Information-theoretic approach to network modularity

Etay Ziv

College of Physicians & Surgeons, Department of Biomedical Engineering, Columbia University, New York, New York 10027, USA

Manuel Middendorf

Department of Physics, Columbia University, New York, New York 10027, USA

Chris H. Wiggins

*Department of Applied Physics and Applied Mathematics, Center for Computational Biology and Bioinformatics,
Columbia University, New York, New York 10027, USA*

Exploiting recent developments in information theory, we propose, illustrate, and validate a principled information-theoretic algorithm for module discovery and the resulting measure of network modularity. This measure is an order parameter (a dimensionless number between 0 and 1). Comparison is made with other approaches to module discovery and to quantifying network modularity (using Monte Carlo generated Erdös-like modular networks). Finally, the network information bottleneck (NIB) algorithm is applied to a number of real world networks, including the "social" network of coauthors at the 2004 APS March Meeting.

I. INTRODUCTION

Modeling is the description of system in terms of less complex degrees of freedom while retaining information deemed relevant [1]. Approaches to complexity reduction for networks include characterizing the network in terms of simple statistics (such as degree distributions [2] or clustering coefficients [3]), subgraphs over-represented relative to an assumed null model (see, for example, [4–6]), and communities (see, for example, [7]).

In the case of communities, networks are coarse-grained into clusters of nodes, or modules, where nodes belonging to one cluster are highly interconnected, yet have relatively few connections to nodes in other clusters. This type of network complexity reduction may be particularly promising as an approach to network analysis, since many naturally occurring networks, including biological [8] and sociological [7,9] networks, are thought to be modular. Clearer, quantitative understanding of these ideas would be valuable in finding reduced complexity descriptions of networks, in visualizing networks, and in revealing global design principles. Two current challenges facing the community regarding network modularity include (i) the ability to quantify to what extent a given network is "modular" and (ii) the ability to identify the modules of a given network.

With regard to quantifying modularity, to our knowledge no mathematical definition has yet been proposed for a measure of modularity that could compare networks regardless of size, origin, or choice of partitioning. In their recent book, Schlosser and Wagner [10] write, "a generally accepted definition of a module does not exist and different authors use the concept in quite different ways." They proceed to warn of the "danger that modularity will degenerate into a fashionable but empty phrase unless its precise meaning is specified." Some positive first steps in this direction have been suggested by Newman's "assortativity coefficient" [11], which quantifies the level of assortative mixing in a network,

and its unnormalized form, called "modularity" [7]. However, these measures quantify the quality of a particular partitioning of the network for a given number of modules, but are *not* a property of the network itself that could serve to compare networks of different origins.

As for module discovery, a range of techniques for identifying the modules in a network have been utilized with various success. In his review article, Newman [12] summarizes these efforts under the broad category of hierarchical clustering in which one poses a similarity metric between pairs of vertices. Hierarchical clustering can be agglomerative, where the most similar nodes are iteratively merged (e.g., [13]) or divisive, where edges between the least similar nodes are iteratively removed (e.g., [7]). By modifying traditional divisive approaches to focus on most "between" edges rather than least similar vertices, Newman and Girvan [7] recently proposed a new class of divisive algorithms for finding modules. Various measures of "edge-betweenness" are defined to identify edges that lie between communities rather than within communities. By iteratively removing edges with the highest betweenness, one can break down the network into disconnected components which define the modules.

In this paper, we take a markedly different approach. The problem of finding reduced descriptions of systems while retaining information deemed relevant has been well-studied in the learning theory community. In particular, the information bottleneck [1,14] provides a unified and principled framework for complexity reduction. By applying the information bottleneck on probability distributions defined by graph diffusion, we propose a new, principled, information-theoretic algorithm to identify modules in networks. We demonstrate that the resulting network information bottleneck (NIB) algorithm outperforms the currently used technique of edge-betweenness (i) in correctly assigning nodes to modules and (ii) in determining the optimal number of existing modules. Moreover, the new method naturally defines a

network modularity measure that quantifies the extent to which any undirected network can be summarized by modules over all scales. Information-theoretic bounds constrain this measure to be between 0 and 1. Finally, we apply our method to a collaboration network derived from the 2004 APS March Meeting and the *E. coli* transcriptional regulatory network.

II. THE INFORMATION BOTTLENECK: A REVIEW

Brief [1] and detailed [14] discussions of the information bottleneck can be found elsewhere; here we review only the most salient features. The fundamental quantity in information theory is Shannon entropy $H[p(x)] \equiv -\Sigma_x p(x) \log p(x)$ measuring lack of information (or disorder) in a random variable X, and uniquely (up to a constant) defined by three axioms [15]. Knowledge of a second random variable Y decreases the entropy in X on average by an amount

$$I(X,Y) \equiv H(X) - \langle H(X|Y) \rangle \equiv H[p(x)] - \sum_y p(y) H[p(x|y)]$$

$$(1)$$

called the *mutual information* [16], the average information gained about X by the knowledge of Y. Equation (1) is equivalent to

$$I(X,Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = \left\langle \log \frac{p(x,y)}{p(x)p(y)} \right\rangle$$

$$(2)$$

revealing its symmetry in X and Y. The mutual information thus measures how much information one random variable tells about the other, and is the basis of the information bottleneck.

Clustering can generally be described as the problem of extracting a compressed description of some data that captures information deemed relevant or meaningful. For example, we might want to cluster protein sequences, expecting that the cluster assignments contain information about the fold of the proteins; or we might want to cluster words in documents, expecting that the clusters capture information about the topic in which the words appear. Tishby *et al.*'s [1] key insight into this problem is the inclusion in the clustering algorithm of another random variable, called the *relevance variable*, which describes the information to be preserved. In the case of protein sequences, the relevance variable might be the protein fold; in the case of clustering words over documents, the relevance variable might be the topic.

Let $x \in X$ be the input random variable (e.g., protein sequences in the set of all observed sequences; or words in a given dictionary, in the two examples above), $y \in Y$ the relevance variable, and $z \in Z$ the cluster assignment random variable [31]. The information bottleneck outputs a probabilistic cluster assignment function $p(z|x)$ equal to the probability to be in cluster z for a given input x. The clustering minimizes the mutual information between X and Z ("maximally compressing the data set"), while constraining the possible loss in mutual information between Z and Y ("preserv-

ing relevant information"). In other words, one seeks to pass or squeeze the information that X provides about Y through the "bottleneck" formed by the compressed Z.

The simplicity of the model Z relative to that of the world X is quantified by the entropy reduction $\mathcal{S}[p(z|x)] \equiv H(X) - I(X,Z) = H(X,Z) - H(Z) > 0$. The gain in simplicity, however, comes with a loss of fidelity in our description of the world, quantified by the error $\mathcal{E} \equiv I(X,Z) - I(Y,Z) > 0$, the loss in information about the world when described by a model Z instead of the primitive description X. The trade-off between the error and the simplicity can be expressed in terms of the functional

$$\mathcal{F}[p(z|x)] = \mathcal{E} - T\mathcal{S} = \mathcal{F}_0 - I(Y,Z) + TI(X,Z) \qquad (3)$$

in which the temperature T parametrizes the relative importance of simplicity over fidelity. The term $\mathcal{F}_0$ is independent of the cluster assignment $p(z|x)$. Since $p(y|x,z) = p(y|x)$, this is the only degree of freedom over which the free energy $\mathcal{F}$ is to be minimized. In the annealed ground state ($T \to 0$), each possible state of the world $x \in X$ is assigned with unit probability to one and only one state of the model $z \in Z$ [i.e., $p(z|x) \in \{0,1\}$, a limit called "hard clustering"]. If the cardinalities $|Z|$ and $|X|$ are equal, we arrive at the fully detailed, trivial solution where the clusters Z simply copy the original X. A formal solution to the information bottleneck problem is given in [1] and yields the following three self-consistent equations (with $\beta = 1/T$):

$$p(z|x) = \frac{p(z)}{Z(\beta,x)} e^{-\beta D_{KL}[p(y|x) \| p(y|z)]},$$

$$p(z) = \sum_x p(z|x)p(x), \qquad (4)$$

$$p(y|z) = \frac{1}{p(z)} \sum_x p(y|x)p(z|x)p(x),$$

where $Z(\beta,x)$ is a normalization (partition) function and $D_{KL}[p\|q] \equiv \Sigma_x p(x) \log[p(x)/q(x)]$ is the Kullback-Leibler divergence (also called the relative entropy). The first of these equations makes clear that as one anneals to the ground state, where $T \to 0$ and $\beta \to \infty$, the only solution is the hard clustering [$p(x|z) \in \{0,1\}$] limit. These three equations naturally lend themselves to an iterative algorithm proposed in [1] which is provably convergent and finds a locally optimal solution.

While in many applications a "soft" clustering might be of interest, for clarity we only consider the hard case in this paper: each node is associated with one and only one module. We use two different algorithms to find approximate solutions to the information bottleneck problem. Both of them take a fixed $|Z|(|Z| < |X|)$ as input and output hard clustering assignments for every node.

The first algorithm (*self-consistent* NIB) uses β as an annealing parameter that starts at low values and increases step by step. At every given β, the locally optimal solution is computed by iterating over Eq. (4). The solution for given β is then taken as a starting point for the iterations with the next β. The second algorithm (*agglomerative* NIB) uses an

agglomerative approach [17]. At every step, a pair of nodes is merged into a single node, where the pair is chosen so as to maximize the relevant information $I(Y, Z)$. It thus reduces $|Z|$ by one at every step, and stops when the desired $|Z|$ is reached.

To summarize, we wish to find a representation of a network in which a group of nodes has been represented by effective nodes; we argue that a modular description of the network is most successful when relevant information about the network is preserved. Posed in this language, it is clear that the act of finding modules in a network is a type of clustering, and the appropriate cluestering framework is one that preserves the information deemed relevant.

III. DIFFUSIVE DISTRIBUTIONS DEFINED OVER GRAPHS

Formulation of graph clustering in terms of the information bottleneck requires a joint distribution $p(y, x)$ to be defined on the graph, where x designates nodes and y designates a relevance variable. An appropriate distribution that captures structural information about the network without introducing additional parameters is the distribution defined by graph diffusion. The relevance random variable y then ranges over the nodes, as does x, and is defined by the node at which a random walker stands at a given time t if the random walker was standing at node x at time 0. The conditional probability distribution $G_{ij}^t \equiv p^t(y_i | x_j)$ is a Green's function describing propagation from node j to node i. For discrete time diffusion, one can easily derive [18]

$$G^t = [WT^{-1}]^t \ (t \in \mathbb{N}), \tag{5}$$

where W is a symmetric weighted affinity matrix of positive entries and $T_{ij} \equiv \delta_{ij}\Sigma_l W_{il} \equiv \delta_{ij}k_i$. For a graph, with identically weighted edges, k_i is the conventional degree (the number of neighbors of node i), and W is the adjacency matrix ($W_{ij} = 1$ iff i is adjacent to j). Note that here we only consider connected graphs and, as defined, this approach treats directed and undirected graphs identically.

In the continuous time limit,

$$G^t = e^{(WT^{-1}-1)t} = e^{-LT^{-1}t} \ (t > 0), \tag{6}$$

where we defined $L \equiv T - W$, the graph Laplacian [19]. In the machine learning literature, a "graph kernel" [20] has been defined as

$$G^t = e^{-Lt} \tag{7}$$

to learn from structured data. It corresponds to a probability distribution associated with a different diffusion rule assuming a degree-dependent permeability at every node. For comparison, we consider both of these joint distributions as possible input to the information bottleneck algorithm.

The characteristic time scale τ of the system is given by the inverse of the smallest nonzero eigenvalue of the diffusion operator exponent (LT^{-1} or L). This time reflects the finite system size and characterizes large-scale behaviors. For example, in one dimension on a bounded domain of size ℓ, the smallest nonzero eigenvalue of the Laplacian with diffusion constant D is $\pi^2 D / \ell^2$. For our algorithm we will thus choose $t = \tau$.

To calculate the joint probability distribution $p(y, x) = p(y|x)p(x)$ from the conditional probability distribution $G^\tau = p(y|x)$, we must specify a prior $p(x)$: the distribution of random walkers at time 0. Natural definitions include (i) a flat prior $p(x) = 1/N$, N being the total number of nodes, and (ii) a prior corresponding to the steady-state distribution associated with the diffusion operator: $p(x) = 1/N$ or $p(x) = k_x/\Sigma_x k_x$ for $G^\tau = e^{-L\tau}$ or $G^\tau = e^{-LT^{-1}\tau}$, respectively, where k_x is the degree of node x.

IV. QUANTIFYING MODULARITY

A. Partition modularity—Quality of a partitioning

Newman and Girvan [7] propose a modularity, a "measure of a particular division of a network," as $Q = \Sigma_i[e_{ii} - (\Sigma_j e_{ij})(\Sigma_k e_{ik})]$, where e_{ij} is the fraction of all edges connecting module i and module j. It can be interpreted as the difference between the fraction of within-module edges and the expected fraction of within-module edges in an ensemble of networks created by randomizing all connections while holding constant the number of edges emanating from each module. Q should therefore go to 0 for randomly connected networks, and tend to $1 - 1/|Z|$ for a perfectly modular network with $|Z|$ equally sized modules. Herein we refer to the measure Q as *partition modularity* to distinguish it from *network modularity*, which we define below based on information-theoretic quantities. Newman *et al.* also study the number of modules $|Z|_{max}$ which maximizes Q given a particular module discovery algorithm.

B. Network modularity—Summarizability of network structure

Here we propose a new modularity measure M, a property of a given network rather than of a given partitioning, which quantifies the extent to which a network can be summarized in terms of modules.

Every clustering solution $p(z|x)$ determines a normalized "input information" $0 < I(Z, X)/H(X) < 1$ between input variable X and cluster assignment Z, and an "output information" $0 < I(Z, Y)/I(X, Y) < 1$ between cluster assignment and relevance variable Y. The *information curve* is then plotted as $I(Z, Y)/I(X, Y)$ versus $I(Z, X)/H(X)$ for every solution of Eq. (4) for every possible number of clusters [1]. An example is shown in Fig. 3. The curve traced by minimizers of the functional Eq. (3), which will not necessarily be computed by such approximating schemes as the AIB, is provably concave. For perfectly random data, which cannot be summarized, this curve lies along the diagonal $y = x$. Consistent with these observations, we find that synthetic graphs with high connectivity within defined modules and low connectivity between different modules exhibit a larger area under the information curve (data not shown). We thus define a new measure of *network modularity*: the area under the information curve (data not shown). That is, we measure the area enclosed by expanding the volume from $|Z| = 1$ to $|Z| = X$ at
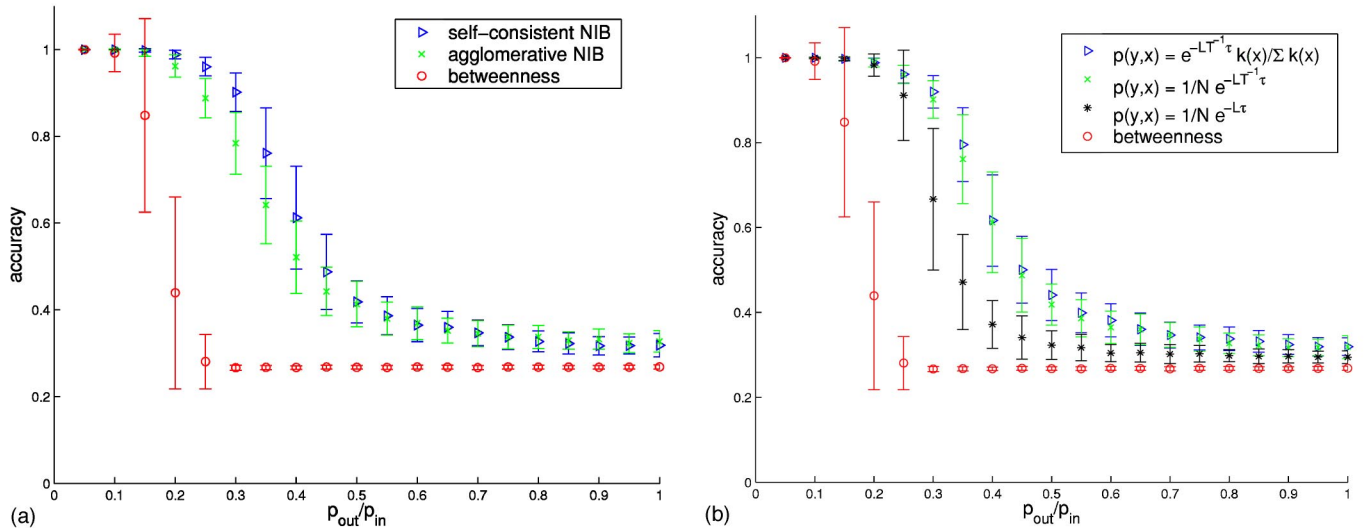
FIG. 1. (a) (Color online) Accuracy for different algorithms. Measured is the accuracy of recovering modular structure in synthetic networks under varying noise levels. Every point represents an average over 100 networks, each with 128 nodes and an average degree of 8. Both NIB algorithms outperform the betweenness algorithm. (b) (Color online) Accuracy for different diffusion operators. Accuracy is measured in the same way as in (a), now using the self-consistent NIB algorithm with various diffusion operators to define probability distributions $p(y,x)$ over nodes. For comparison, the betweenness results are also shown. The operator for physical diffusion $e^{-LT^{-1}t}$ outperforms the "graph kernel" diffusion operator proposed in the machine learning literature [20].

fixed $T=0$, followed by heading from $T=0$ to $T=1$ [the highest temperature allowed by the data-processing inequality $I(Z,Y) < \min(I(X,Y),I(Y,Z))$] and fixed volume, compressing from $|Z|=|X|$ to $|Z|=1$ at fixed temperature, and finally cooling from $T=1$ to $T=0$ at fixed volume. In the soft clustering case, the information curve is continuous since solutions vary with every choice of $\beta \in [0,\infty)$. In the hard clustering case, which we study here, the information curve is only defined at discrete points corresponding to solutions for every possible number of clusters $|Z|$. The area can then be calculated by linear interpolation. Information-theoretic bounds constrain the range of m allowing comparison of networks of varying number of nodes and edges, and are a property of the network itself, rather than a given partitioning.

V. TESTS ON SYNTHETIC NETWORKS

A. Accuracy of the partitioning

Here we test how well various NIB implementations with different diffusion operators can reconstruct modules in a network generated with a known modular structure. We also compare our method to the "edge-betweenness" algorithm recently proposed in [7] for the same purpose of finding modules or "communities." In [7], the network is broken down into isolated components by iteratively removing edges with highest "betweenness" (several definitions of edge-betweenness are tested in [7]; here we use the "shortest path" betweenness, which was shown in [7] to perform optimally).

As in [7], we generate synthetic Erdös-like graphs via Monte Carlo with 128 nodes each and average degree 8 (average total of 512 edges). We also demand that the graphs be

connected by rejecting generated graphs that have disconnected components. We impose a structure of four modules with 32 nodes each by introducing two different probabilities: p_{in} for edges inside modules and p_{out} for edges between different modules. The level of noise in the graph is thus controlled by p_{out}. The higher p_{out}, the harder it will be to recover the different modules. We first generate networks with $p_{out}=0$ and then increase p_{out} while adjusting p_{in} such that the average degree remains fixed. When $p_{in}=p_{out}$, all modular structure is lost and we obtain a usual Erdös graph. We measure the accuracy of a proposed partitioning using the following computation. In principle, any module proposed by the algorithm could match any "true" module with an associated error. We therefore try every possible permutation of the four proposed modules matching the four "true" modules, and consider the one permutation with the smallest total number of incorrectly assigned nodes. We define *accuracy* as the total fraction of correctly assigned nodes.

Figure 1(a) shows the accuracy of the recovered modules as a function of p_{out}/p_{in} for three different algorithms: self-consistent NIB, agglomerative NIB, and betweenness. Both NIB algorithms use the physical diffusion operator $e^{-LT^{-1}t}$ and a flat prior $1/N$ to define a joint probability distribution. We observe that both NIB algorithms are much more successful in recovering the modular structure than the betweenness algorithm. A threshold noise level is achieved at around $p_{out}/p_{in} \approx 1/3$ for the NIB algorithms, and around $p_{out}/p_{in} \approx 1/6$ for the betweenness algorithm. The figure also shows that the self-consistent NIB in general finds a better partitioning than the agglomerative NIB.

Figure 1(b) shows the same measurements for self-consistent NIB algorithms using different diffusion operators as explained in Sec. III. For comparison, the betweenness results are also plotted. Physical diffusion, defined by the
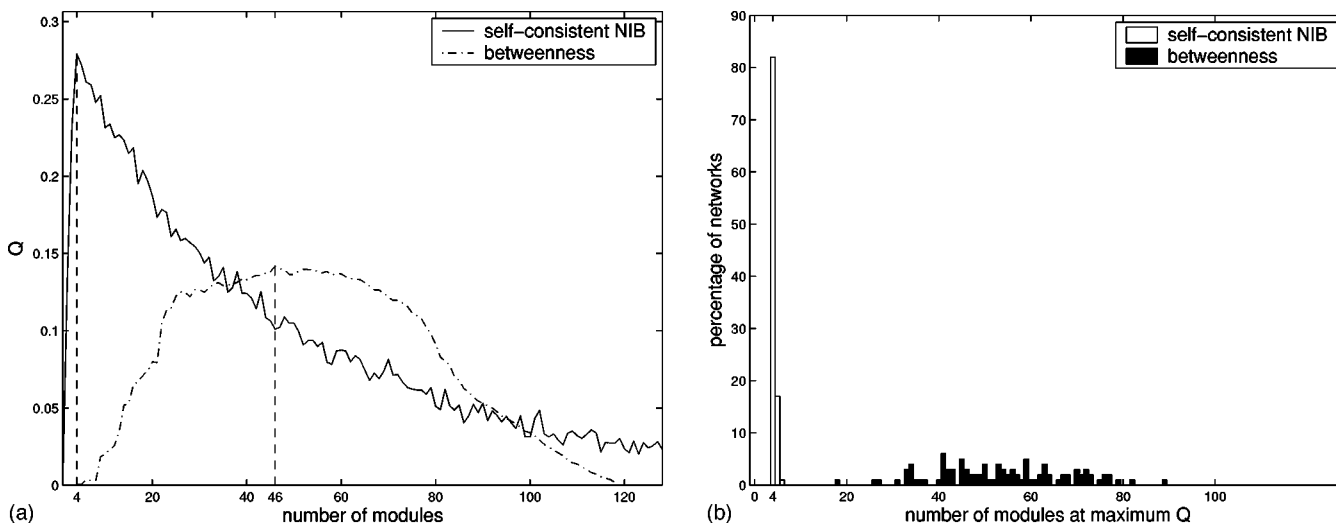
FIG. 2. (a) Q vs $|Z|$. The quality of the partitioning Q is computed for the partitionings output by the self-consistent NIB algorithm (with physical diffusion operator), and the betweenness algorithm, for every given number of modules $|Z|$. While the NIB algorithm correctly determines $|Z|=4$ as the best number of modules (at maximum Q), Q calculated by the betweenness algorithm peaks at $|Z|=46$. (b) Histogram of $|Z|_{\max}$ for 100 different networks. For 82% of the networks, the NIB algorithm is able to find the correct number of modules $|Z|_{\max}=4$, and comes close to it ($|Z|_{\max}=5$ or 6) in all other cases. The betweenness algorithm gives $|Z|_{\max}$ between 18 and 89, far from the correct value for all networks.

continuous time limit, with an initial state $p(x)$ given by the equilibrium distribution $p(x) \propto k_x$, gives the best performance.

B. Finding the optimal number of modules

In most real world problems, the correct number of modules $|Z|$ present in the network is unknown *a priori*. It is therefore important to have an algorithm which not only computes a good partitioning for a given $|Z|$ but also gives a good estimate for $|Z|$ itself. To this end, here we make use of the partition modularity Q as described in Sec. IV A.

We again consider synthetic connected networks of 128 nodes and average degree 8 as in the previous section. However, we fix the noise level to a value of $p_{\mathrm{out}}/p_{\mathrm{in}}=0.3$, which was shown to be a critical level for these networks. We run the self-consistent NIB and the betweenness algorithms for every possible number of modules $|Z|=1,2,\ldots,128$ and compute Q for the proposed partitionings. Figure 2(a) shows Q as a function of $|Z|$ for a typical run. While for the NIB algorithm Q sharply peaks at the correct value of $|Z|_{\max}=4$, Q calculated by the betweenness algorithm attains its maximum at $|Z|_{\max}=46$ and does not show a particular signal at $|Z|=4$. Figure 2(b) shows a histogram of $|Z|_{\max}$ for 100 generated networks. The NIB algorithm successfully identifies $|Z|_{\max}=4$ for 82% of the networks, while the betweenness algorithm calculates $|Z|_{\max}$ lying between 18 and 89, notably far from the correct value for any network. These experiments suggest that the NIB algorithm performs well both in accurately assigning nodes to modules and in revealing the optimal scale for partitioning.

VI. APPLICATIONS

A. Collaboration network

Having validated NIB on a toy model of modular networks, we next apply our algorithm to two examples of natu-

rally occurring networks. In the first example, we construct a collaboration network from the 2004 APS March Meeting, where this algorithm was first presented, and in the second example we construct a symmetric version of the *E. coli* genetic regulatory network.

Vertices of the collaboration network represent authors from all talks at the March Meeting; edges represent coauthorship. The largest component of the resulting graph consists of 5603 vertices with 19 761 edges. Network information bottleneck using the agglomerative algorithm and the physical diffusion operator (as defined in Sec. III with its corresponding equilibrium distribution) reveals that this large network is highly modular [$m=0.9775$, see Fig. 3(a)]. For comparison, we also show the information curve for a typical Erdös network, which is clearly less modular. Such a high value of modularity implies that the authors of this component of the network are "easily" compressed or combined into larger clusters of authors. In light of this fact, we study what the clusters of authors reveal about the collaboration network. For example, authors may group themselves according to topics or subject matters of the talks; alternatively, author modules may be more indicative of the authors' affiliations or even geographical location.

To begin to approach these types of questions, we may choose to look at the author groupings given by NIB at a particular number of clusters. While we emphasize that network modularity is a measure over all scales or all numbers of clusters, it is illustrative in this case also to examine the clustering at a particular scale. For the APS network, such an analysis yields the optimal number of modules, $|Z|_{\max}=115$, for this network [see Fig. 3(a)].

In Fig. 4, we plot the 115 modules and their connections where each ellipse represents one module and edges between ellipses represent intermodular connections. The sizes of the ellipses and the thickness of the edges are proportional to the log of the number of authors in a module and the log of the
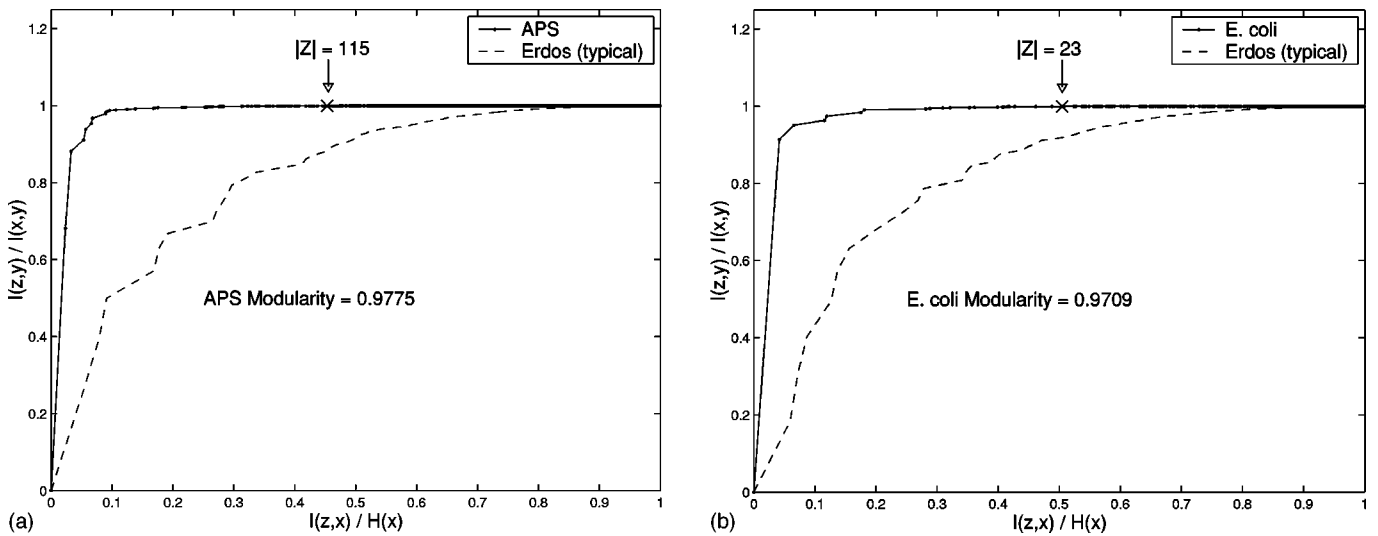
FIG. 3. (a) APS network modularity. Information plane for the collaboration network obtained from the 2004 APS March Meeting (the largest component consists of 5603 authors and 19 761 edges). We use the agglomerative algorithm with the diffusion operator $e^{-LT^{-1}t}$. Network modularity for this graph, defined as the area under the curve, is 0.9775. Comparison is made with a typical information curve obtained from an Erdös graph. The optimal number of modules as defined by the the Newman and Girvan measure is at $|Z|=115$. (b) *E. coli* network modularity. Information plane for the *E. coli* genetic regulatory network (the largest component has 328 nodes and 456 edges). We use the agglomerative algorithm with the diffusion operator $e^{-LT^{-1}t}$. Network modularity for this graph, defined as the area under the curve, is 0.9709. Comparison is made with a typical information curve obtained from an Erdös graph. The optimal number of modules as defined by the Newman and Girvan measure is at $|Z|=23$.

number of intermodular connections between modules, respectively. We note the provocative structure revealed in the figure with a large center of highly connected modules (including two of the largest modules), three more or less branching, linear chains of modules, and one large 18-node cycle of modules.
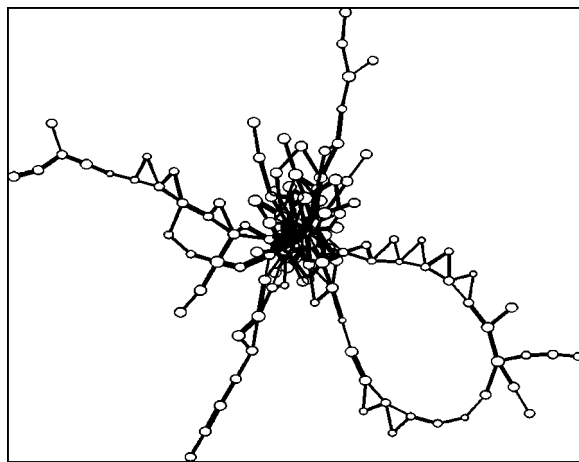

FIG. 4. Adjacency network of the 115 modules of the APS network. Nodes represent modules where the size of the drawn ellipse is proportional to the number of authors in the module. Edges between modules represent collaborations between authors in different modules, where the thickness of the drawn lines is proportional to the number of these intermodule collaborations. The module network reveals a structure with a highly dense center of modules, three branching linear chains of modules, and one cycle of modules.

Closer inspection of a single module demonstrates that for many of the modules, institutional affiliation, and even geography, play a large role in determining collaborations. In Fig. 5, a single 17-node module is plotted where each node now represents an author and edges represent author collaborations. We see that 15 of the 17 authors are affiliated with Columbia University; the remaining two authors are affiliated with Stony Brook, and notably, are adjacent (indicating coauthorship) to each other. The finding that the modules in this collaboration network are somewhat related to institu-
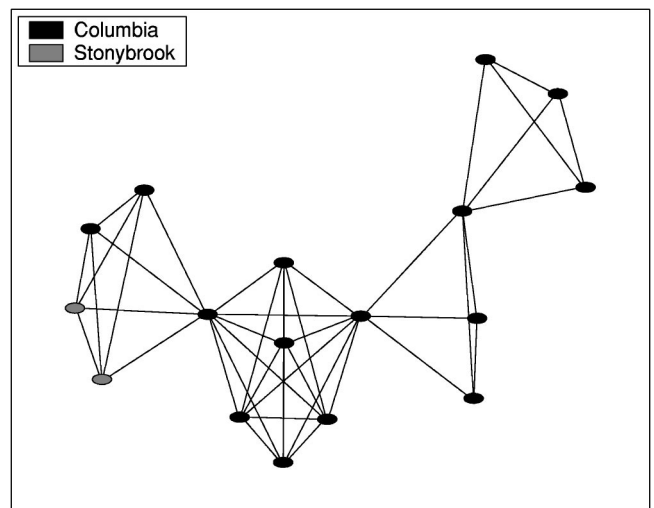

FIG. 5. One of the 115 modules of the APS network. Nodes represent authors and edges represent collaborations. Of the 17 authors in the module, 15 are Columbia University affiliated and two are affiliated with Stony Brook University.
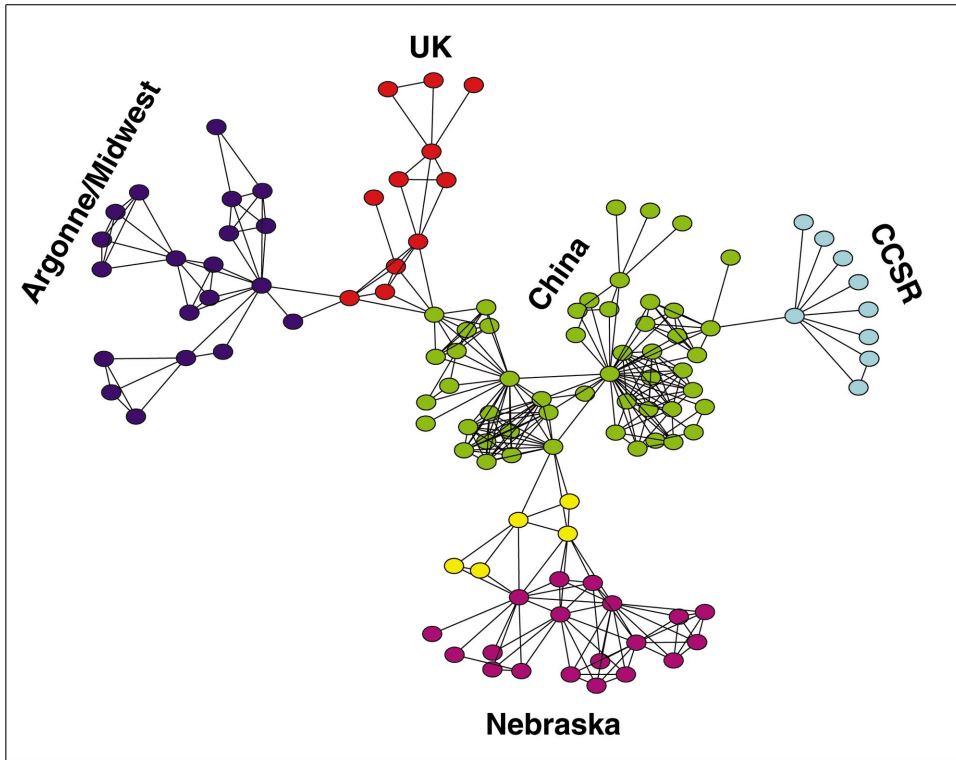
FIG. 6. (Color) Six modules corresponding to the uppermost branched linear chain of modules depicted in Fig. 4. Colors denote modules as defined by the network information bottleneck algorithm. Again the modules roughly correspond to institutional affiliations. Over 50% of the blue nodes have one or more affiliations with the institutions based in and around Chicago (Argonne National Laboratory, University of Illinois at Chicago, and University of Notre Dame). 70% of the red nodes are in England, and 75% of the green nodes are in China, mostly at the Institute of Chemistry, Chinese Academy of Sciences, and all of the cyan nodes are at the Center of Complex Systems Research in Illinois. Both the yellow and magenta modules are mostly affiliated with the University of Nebraska.

tional affiliations and geography is supported by similar results found in other physics collaboration networks previously studied using different techniques [7].

Another possible annotation for this module to consider is that of the APS divisions and topical groups, since each author is associated with at least one talk and each talk is listed under one or more of these APS categories. However, the 14 APS divisions and 10 topical groups appear to be too broad and have too much overlap to clearly define a module. For example, the Columbia University module includes talks under the categories of Polymer, Condensed Matter, Material, and Chemical Physics. On the other hand, the module is
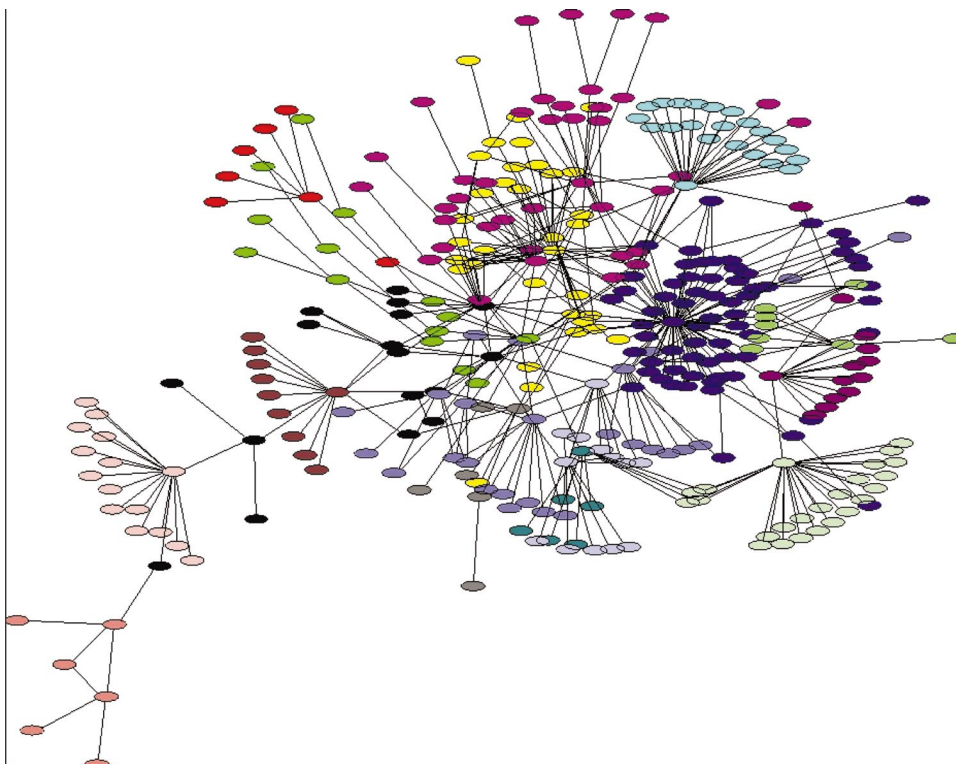

FIG. 7. (Color) *E. coli* gene regulatory network. Largest component of the symmetric version of the *E. coli* genetic regulatory network. Colors denote modules identified by NIB.

essentially representative of researchers at the Columbia University Materials Research Science and Engineering Center (MRSEC) and in particular those interested in the synthesis of complex metal oxide nanocrystals. There is thus both topical and institutional information retained in the modules.

It is also revealing to examine the affiliations of multiple connected modules. For example, Fig. 6 plots the uppermost branching linear chain of Fig. 4. Here, color denotes module assignment as given by NIB. Most of these modules also have clear institutional affiliations. For example, everyone in the cyan module is at the Center of Complex Systems Research (CCSR) in Illinois; close to 80% of the large green module is in China, mostly at the Institute of Chemistry, Chinese Academy of Sciences (ICCAS); and 70% of the red module is in England. The blue module is slightly more diffuse, though an institutional affiliation is also apparent here; over 50% of the authors are affiliated with one of three institutions near Chicago (Argonne National Labs, University of Illinois at Chicago, and University of Notre Dame). The yellow and magenta modules are also overwhelmingly associated with the University of Nebraska, though interestingly our algorithm separates these two modules at this partitioning. In general, one does not anticipate that the optimal number of clusters in a given network will give the most natural partitioning at all scales and over all resulting modules.

B. Biological network

The notion of modularity has been central in the study of a variety of biological networks including metabolic [13], protein [21,22], and genetic [4,8] networks. Certainly most biologists agree that the various networks operating within and between cells have a modular structure, though what they mean by "modular" can vary greatly [10].

NIB allows us to investigate quantitatively and in detail to what extent naturally occurring biological networks are modular. For example, Fig. 7 depicts the undirected form of the largest component of the *E. coli* genetic regulatory network described previously in [4,23]. The network consists of 328 vertices and 456 edges and its modularity is depicted by the curve one traces in the information plane as the network is clustered using the network information bottleneck [see Fig. 3(b)].

To establish whether the modularity of the network should be considered low, high, or moderate, we employ an ansatz popular in several reserach communities in which a distribution of networks is created by holding the in-, out-, and self-degree of each node constant but randomizing the connectivity of the graph, changing which nodes are connected to which neighbors [4,5,21,24–26]. The randomization, a variant of the configuration model [12], produces a distribution of networks from which we sample and then measure the network modularity. The histogram in Fig. 8 shows that *E. coli*'s modularity is higher relative to this ensemble.

VII. CONCLUSIONS AND EXTENSIONS

We have presented a principled, quantitative, parameter-free, information-theoretic definition of network modularity,
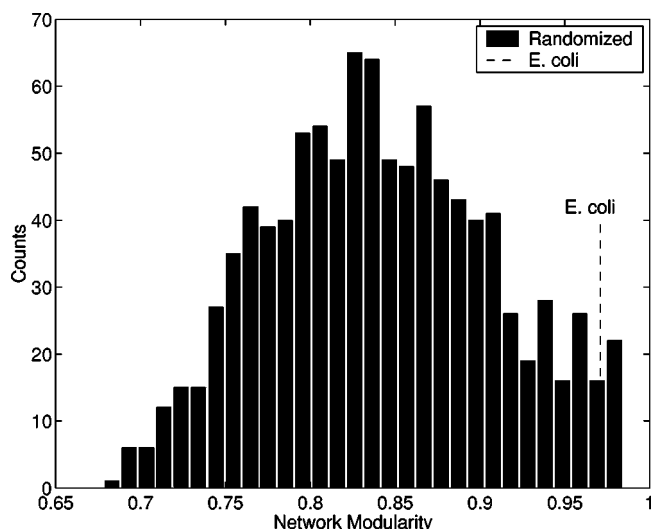

FIG. 8. A histogram of network modularity, defined by the area underneath the curve in the information plane resulting from network compression, for 1000 realizations of the variation of the configuration model. Note that the true *E. coli* network is more modular than the typical network resulting from the variation of the configuration model.

as well as an algorithm for discovering modules of a network. Network modularity is a dimensionless number between 0 and 1 and is a property of a given network over all scales, rather than of a given partitioning with a given number of modules. The measure is applicable to any network, including those with weighted edges. We validate the effectiveness of our algorithm in identifying the correct modules and in finding the true number of modules on synthetic, Monte Carlo generated, Erdös-like, modular networks. Finally, application to two real-world networks, a "social" network of physics collaborations and a biological network of gene interactions, is demonstrated.

Network modularity, the area under the curve in the information plane, is but one relevant statistic that we may retrieve from the information curve. Certainly other useful statistics may be culled. For example, the optimal information curve will always be concave [14] and its slope will decrease monotonically. The point at which the slope equals 1 is uniquely determined for each network and can be described as the point after which clustering further results in a greater loss in relative relevant information than gain in relative compression [that is, $\delta(I(Z,Y)/I(X,Y)) = \delta(I(Z,X)/H(X))$]. This *break-even point* is the point at which one can gain further (normalized) simplicity only by losing an equivalent (normalized) fidelity. Numerical experiments and investigating the utility of this measure are currently in progress.

Diffusive distributions are but one general class of distributions on a network. A natural generalization of these ideas is to describe other distributions on a network for which a particular function, energy, or origin is known, and on which some particular degree of freedom (such as chemical concentration or genetic expression as a function of time) may be defined.

Finally, we note that while the information bottleneck is a prescription for finding the highest-fidelity summary of a

system at a given simplicity, algorithms for determining network community structure are usually motivated by various definitions of normalized min-cuts [27–30]. Our results, particularly for the synthetic graphs with prescribed modular structure, demonstrate that *information* modularity implies *edge* modularity, an unexpected finding which motivates further numerical and analytic investigations in progress regarding this relationship.

[1] N. Tishby, F. Pereira, and W. Bialek, in *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing* (University of Illinois Press, Urbana-Champaign, IL, 1999), pp. 368–377, URL arxiv.org/abs/physics/0004057.

[2] H. Jeong, B. Tombor, R. Albert, Z. Oltavi, and A.-L. Barabasi, Nature (London) **407**, 651 (2000).

[3] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).

[4] S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, Nat. Genet. **31**, 64 (2002).

[5] P. Holland and S. Leinhardt, Sociol. Methodol. **7**, 1 (1976).

[6] Y. Artzy-Randrup, S. J. Fleishman, N. Ben-Tal, and L. Stone, Science **305**, 1107c (2004).

[7] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).

[8] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray, Nature (London) **402**, C47 (1999).

[9] S. Wasserman, K. Faust, and M. Granovetter, *Social Network Analysis* (Cambridge University Press, Cambridge, UK, 1994).

[10] G. Schlosser and G. P. Wagner, *Modularity in Development and Evolution* (University of Chicago Press, Chicago, 2004).

[11] M. E. J. Newman Phys. Rev. E **67**, 026126 (2003).

[12] M. Newman, SIAM Rev. **45**, 167 (2003).

[13] E. Ravasz, A. Somera, D. Mongru, Z. Oltavi, and A.-L. Barabasi, Science **297**, 1551 (2002).

[14] N. Slonim, Ph.D. thesis, The Hebrew University of Jerusalem, 2002).

[15] C. Shannon and W. Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Urbana-Champaign, IL, 1949).

[16] T. Cover and J. Thomas, *Elements of Information Theory* (Wiley, New York, 1990).

[17] N. Slonim, N. Friedman, and N. Tishby, in *Proceedings of NIPS-12, 1999* (MIT Press, Cambridge, MA, 2000), pp. 617–623.

[18] F. R. K. Chung, *Spectral Graph Theory*, no. 92 in Regional Conference Series in Mathematics (American Mathematical Society, Providence, RI, 1997).

[19] R. Merris, Linear Algebr. Appl. **197/198**, 143 (1994).

[20] R. I. Kondor and J. Lafferty, in *Proceedings of the 19th International Conference on Machine Learning (ICML)*, edited by C. Sammut and A. Hoffman (Morgan Kauffman, Sydney, 2002), p. 315.

[21] S. Maslov and K. Sneppen, Science **296**, 910 (2002).

[22] J.-D. J. Han *et al.*, Nature (London) **430**, 88 (2004).

[23] H. Salgado, S. G.-C. A. Santos-Zavaleta, D. Millan-Zarate, E. Diaz-Peredo, F. Sanchez-Solano, E. Perez-Rueda, C. Bonavides-Martinez, and J. Collado-Vides, Nucleic Acids Res. **29**, 72 (2001).

[24] E. Ziv, R. Koytcheff, M. Middendorf, and C. H. Wiggins, Phys. Rev. E **71**, 016110 (2005).

[25] E. F. Connor and D. Simberloff, Ecology **60**, 1132 (1979).

[26] T. Snijders, Psychometrika **56**, 397 (1991).

[27] A. Ng, M. Jordan, and Y. Weiss, in *Advances in Neural Information Processing Systems (NIPS) 14*, edited by T. G. Dietterich, S. Becker, and Z. Ghahramani (MIT Press, Cambridge, MA, 2002), pp. 839–856.

[28] Y. Weiss, Tech. Rep., CS Dept., University of California at Berkeley (1999).

[29] J. Shi and J. Malik, Proceedings of IEEE Conference of Comp. Vision and Pattern Recognition (IEEE Computer Science, Puerto Rico, 1997), pp. 731–737.

[30] F. R. Bach and M. I. Jordan, in *Advances in Neural Information Processing Systems 16*, edited by S. Thrun, L. Saul, and B. Schölkopf (MIT Press, Cambridge, MA, 2004).

[31] We use Z here as the cluster variable rather than T as in many papers [14] in order to avoid confusion with time t and temperature T, which will appear later.