EXPLORING BIOLOGICAL NETWORK DYNAMICS WITH ENSEMBLES OF GRAPH PARTITIONS

SAKET NAVLAKHA AND CARL KINGSFORD*

Center for Bioinformatics and Computational Biology, and
Department of Computer Science,
University of Maryland, College Park, MD 20742.

Unveiling the modular structure of biological networks can reveal important organizational patterns in the cell. Many graph partitioning algorithms have been proposed towards this end. However, most approaches only consider a single, optimal decomposition of the network. In this work, we make use of the multitude of near-optimal clusterings in order to explore the dynamics of network clusterings and how those dynamics relate to the structure of the underlying network. We recast the modularity optimization problem as an integer linear program with diversity constraints. These constraints produce an ensemble of dissimilar but still highly modular clusterings. We apply our approach to four social and biological networks and show how optimal and near-optimal solutions can be used in conjunction to identify deeper community structure in the network, including inter-community dynamics, communities that are especially resilient to change, and core-and-peripheral community members.

1. Introduction

Many types of biological networks, such as protein interaction networks and metabolic networks, are known to be modular in nature[15]. Modules are typically composed of a set of nodes that are all functionally related. Uncovering such functional building blocks is useful because it can provide us with a systems-level understanding of how the cell is organized. Several graph partitioning algorithms have been recently proposed for this purpose (e.g.[40,2,30,32,34]), but these algorithms typically select only a single solution from the vast space of possible clusterings. The chosen solution is meant to characterize the modular structure of the data, but it ignores the horde of near-optimal solutions.

Near-optimal solutions are crucial in many respects. For example, they can help assess confidence in the optimal partitioning. If a near-optimal solution is nearly as good as the optimal, we may be unsure whether it is the near-optimal or the optimal partitioning that represents the true community structure. This is especially true in the presence of noise, when the true community structure might be obscured and as a result only emerge as some near-optimal solution. More locally, pairs of nodes that are co-clustered in many near-optimal partitionings can be confidently determined to be members of the same community. Equivalence classes of these frequently co-clustered nodes can be considered the "core" members of a community. Others ought to be considered tenuous or "peripheral" members. Thus, unlike single solution approaches that treat each individual as an equivalent community member, near-optimal solutions provide a way to measure the strength of membership of members to each community. Further, understanding inter-and intra-community interactions can be used to quantify how robust or resilient a community is to change. By taking such interactions into account, we transition from treating communities as static, independent blobs to dynamic blobs with varying memberships. Finally, there is also theoretical and empirical evidence suggesting that single point solutions in high-dimensional spaces do not represent the data as well as ensembles of solutions[4]. This is particularly true in machine learning, where ensembles of classifiers have been consistently shown to outperform single models[36,37].

In this article, we look at a broad collection of social and biological networks and show how near-optimal clusterings impart information into community dynamics that would otherwise be missed using single solution approaches. We use the popular modularity partitioning criteria proposed by Newman[34]. Modularity has received mixed reviews regarding its relevance to biological networks. It was shown to perform poorly at

*Corresponding author: carlk@cs.umd.edu.

recovering functional modules from large protein interaction networks[31], but it has been effective at finding modules in smaller metabolic networks[11]. We consider it here for smaller networks, but use it simply as a template to investigate near-optimal solutions. It is likely that other approaches will also reap similar benefits by considering ensembles of solutions.

To explore near-optimal community partitions, we cast modularity optimization as an integer linear program (ILP), as has been done before[1,3], but add diversity constraints so that each subsequent clustering is not only adequately different from all previous solutions, but also has high modularity. This way, we directly optimize for both diversity and quality. The collection of solutions returned constitute a partial "energy landscape" that represents overlaid decompositions of the network.

Several techniques have been proposed for finding ensembles of optimal and near-optimal solutions to similar ILP problems. For example, both randomly perturbing objective function weights by a small amount[29,12], or perturbing the input data itself and re-clustering[16,21], can help explore different regions of the clustering space (though selecting the size of that perturbation can be difficult). Alternatively, a randomized rounding procedure to convert a fractional solution to an integral one can be used[1], yielding a slightly different partitioning each time. In addition, heuristic techniques such as simulated annealing[11,27] can be used instead of ILPs to optimize the modularity. Such approaches explicitly explore the state space, and an ensemble of partitionings can be generated by saving any good solutions observed. But these techniques are all based on the idea of randomization: perturbing the inputs or the outputs randomly, or randomly transitioning between solutions. Such randomized procedures suffer from at least two deficiencies. First, they often yield solutions very similar to the optimal because large deviations are improbable to be generated at random. Secondly, there is no guarantee that the perturbed solutions have high modularity. The randomized procedure may generate many diverse solutions of poor quality. Other approaches vary input parameters, such as the number of clusters to return[22], though there can exist multiple reasonable clusterings that have the same number of clusters. Recently, another approach was proposed that systematically perturbs the input data such that the transformed and original data retain similar properties; an alternative clustering is then found by clustering the transformed data[38]. Here, we take the approach of explicitly constraining for diversity within the clustering process itself. This guarantees that each successive solution is both sufficiently different from previously obtained solutions and achieves the maximum possible modularity attainable under the given diversity criteria.

We explore a broad spectrum of social and biological networks in an attempt to show the types of insights that can be extracted from large collections of near-optimal solutions. We begin with Zachary's karate club social network[43], which documents the fission of a group of university students after an internal dispute over the price of karate lessons. Interestingly, we find that the clustering closest to the actual resulting fission of the club (i.e. the true clustering) does not appear until the 31^{th} near-optimal solution. We also show that exploring near-optimal solutions can help identify fringe members of the two factions.

We next look at the ERK1/ERK2 mitogen-activated protein kinase (MAPK[18]) signal-transduction pathway. We identify functional subunits that correspond well to known submodules of the pathway, and we classify their robustness across the modularity landscape. Two portions of the ERK pathway consistently remain tightly bound, whereas all other components are eventually split. We also identify *gatekeeper* nodes that lie between functional modules in the Integrin signalling pathway[26].

Finally, we consider a network of cortical-cortical connections in the human brain and find 53 of the first 60 near-optimal solutions are within 1% of the optimal modularity. Of these, 12 have a $> 3\%$ advantage in spatial coherence over the optimal clustering, indicating that they might better represent the true modules of the brain. Differentially classified nodes in this case can be used to identify spatial outliers with respect to the topology. The immense number of similar solutions also suggests tremendous uncertainty in the optimal partitioning.

In all four networks, we find insights conveyed by near-optimal partitionings that helps augment our current understanding of community structure and dynamics.

2. Generating A Diverse Ensemble of Partitionings

Below, we describe our procedure for generating an ensemble of distinct, high-modularity clusterings using integer linear programming (ILP). All superscripts used below indicate indices, not exponentiation.

2.1. *Integer Programming for Modularity*

Intuitively, maximizing modularity corresponds to finding communities where the number of edges lying within a cluster is much greater than we would expect by chance (under an Erdös-Renyi null distribution), and the number of edges connecting two different clusters is much less. Formally, the modularity $q(G, \mathcal{C})$ of an undirected, unweighted network G with community decomposition $\mathcal{C}$ is defined as

$$q(G, \mathcal{C}) := \sum_{u,v \in V} \left(A_{uv} - k_u k_v/(2m)\right)(1 - x_{uv}), \tag{1}$$

where A_{uv} is an entry in the adjacency matrix for G (it is 1 if u and v interact and 0 otherwise), k_u is the degree of node u, m is the total number of edges, and the variables x_{uv} describe $\mathcal{C}$ by indicating which vertices are in the same community. More specifically, we have a variable x_{uv} for every pair of nodes $u < v$, with the interpretation that $x_{uv} = 1$ if u and v belong to different clusters, and $x_{uv} = 0$ otherwise. Letting $m_{uv} = A_{uv} - k_u k_v/(2m)$, a pair of nodes u, v in the same cluster contributes m_{uv} to the total modularity (m_{uv} may be negative). Hence, we seek to maximize $\sum_{u,v} m_{uv}(1 - x_{uv})$ by setting the x_{uv} variables appropriately.

To ensure that the nodes identified as co-clustered are consistent with each other, we must enforce the triangle inequality. This leads to the following integer linear program, MOD-ILP:

$$\text{maximize} \quad \sum_{u \in V} \sum_{v \in V} m_{uv}(1 - x_{uv}) \tag{2}$$

subject to

$$x_{uv} + x_{vw} \geq x_{uw} \qquad \text{for all } u, v, w \in V \tag{3}$$

$$x_{uv} \in \{0, 1\} \tag{4}$$

This ILP is identical to the one proposed by Agarwal et al.[1] for modularity maximization and is similar to the ILP proposed for correlation clustering by Charikar et al.[5] Another similar ILP, where instead $x_{uv} = 1$ indicates that u and v are in the same cluster and with consequently modified constraints, was proposed by Brandes et al.[3] Here, we use MOD-ILP as a tool to generate ensembles of diverse community decompositions, as described in the next section. The ILP can be solved to optimality via branch-and-bound using an ILP solver such as `glpk`[25] or CPLEX[17]. There are $\binom{n}{2}$ variables and $3\binom{n}{3}$ constraints, where n is the number of nodes. For large networks solving the ILP to optimality can be time consuming. Hence, a rounding heuristic has been proposed[1] based on an approximation algorithm for correlation clustering[5]. In this approach, the integrality constraints (4) are replaced by constraints requiring $0 \leq x_{uv} \leq 1$ and the fractional solution is rounded, treating the fractional x_{uv} values as pairwise distances between the nodes. In this article, we focus on smaller networks that can be solved to optimality. However, for larger networks the LP-relaxation of MOD-ILP (with subsequent rounding) can be used, along with the same diversity constraints that are discussed below.

2.2. *Diversity Constraints*

A solution to MOD-ILP reveals only one possible partitioning of the network. Suppose X^0 is a $\binom{n}{2}$-vector $\langle x^0_{uv} \rangle$ representing an optimal solution to MOD-ILP, and let $\vec{1}$ be the $\binom{n}{2}$-vector with every component equal to 1. The following constraints require a vector X to be different from vector X^0:

$$X^0 \cdot (\vec{1} - X) \geq d^0_{\text{merge}} \tag{5}$$

$$(\vec{1} - X^0) \cdot X \geq d^0_{\text{split}} \tag{6}$$

Here, $\cdot$ denotes the dot product between the vectors. Considering X^0, d^0_{merge} and d^0_{split} to be constants, the constraints represented in (5) and (6) are linear. By adding them to MOD-ILP and finding a new optimal, the ILP is forced to return a solution X that is different from X^0. The amount of difference is governed by the parameters d^0_{split} and d^0_{merge}. Equation (5) requires that at least d^0_{merge} variables change from 1 to 0, thereby requiring that d^0_{merge} pairs of nodes formerly in separate clusters become co-clustered. Similarly, equation (6) requires that at least d^0_{split} pairs that were co-clustered in X^0 are placed in separate clusters in X. The parameters d^0_{merge} and d^0_{split} can be set to vary the level and type of diversity desired. Both constraints are required to balance between larger-and smaller-sized clusters, respectively. We can avoid setting separate levels of each diversity type by consolidating these constraints:

$$X^0 \cdot (\vec{1} - X) + (\vec{1} - X^0) \cdot X \geq d_{\text{changes}}\,, \tag{7}$$

where the left-hand side is equivalent to the Hamming distance, $\Delta(X, X^0)$, between vectors X and X^0. Re-solving MOD-ILP with constraint (7) added will find an alternative optimal (if one exists) or will find a second-best partitioning.

To speed up the solution of the ILP, we can use a heuristic algorithm to find a reasonable partitioning and then supply that partitioning to the integer programming solver as an initial basis. Here, this was necessary only for the Integrin pathway and the human brain network, where we used the partitioning found by Newman's spectral method[34] as a starting basis. This provided the solver a starting point for the branch-and-bound process and resulted in convergence in minutes (as opposed to hours). Such an initial basis does not alter the optimality of the solution found.

2.3. *Modularity Landscape*

A partial "modularity landscape" of a network can be generated by iteratively solving MOD-ILP including constraint (7) while increasing d_{changes}. If X^i is the solution of the ith iteration, in the $i+1$ interaction, we set

$$d^{i+1}_{\text{changes}} = \Delta(X^0, X^i) + 1. \tag{8}$$

In contrast to repeated sampling using, e.g., simulated annealing[11,27], this approach guarantees that successively obtained partitionings maximize modularity while still being sufficiently different from the optimal, X^0. We call this the *distance-based* method of generating diverse solutions.

An alternative method for generating an ensemble of diverse, high-modularity partitionings is to repeatedly resolve MOD-ILP with the addition of several constraints of the form of Equation 7, one for each previously uncovered solution. In other words, on the ith iteration, for each previous solution X^j $(0 \leq j < i)$, we add a constraint $X^j \cdot (\vec{1} - X) + (\vec{1} - X^j) \cdot X \geq 1$ to MOD-ILP. A new solution will have at least one difference from each previously uncovered solution. We call this the *point-based* method because it is akin to avoiding specific markers on the clusterings space. The point-based method produces clusterings that are finer-grained than the distance-based approach because there can exist many solutions having distance between d^i_{changes} and d^{i+1}_{changes} that the distance-based method would miss. Using the point-based method, the i^{th} solution returned is a provably i^{th} optimal network decomposition in terms of modularity (clusterings with identical modularity will be ordered arbitrarily). The distance-based method more quickly samples a more diverse collection of solutions. By setting $d_{\text{changes}} > 1$, the point-based approach could also be adopted to more rapidly sample the solution space. In the results described below, we experiment with the distance-based approach and the point-based approach with $d_{\text{changes}} = 1$.

2.4. *Determining Core and Peripheral Community Members*

Nodes that travel together across the modularity landscape can be thought of as *core* members of a community. Such nodes remain together despite the additional diversity constraints added, which implies that their cohesion is stronger than that of other pairs of nodes. Nodes whose co-clustered neighbors fluctuate across

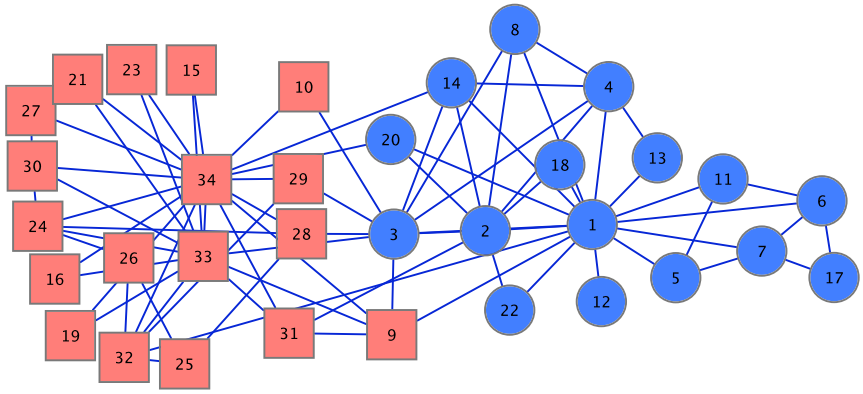

Figure 1. **Zachary's karate club social network[43].** The network consists of 34 nodes and 78 edges. Blue circles correspond to Mr. Hi's faction. Red squares correspond to the officers' faction.

solutions can be considered *peripheral* members that lie on the outskirts of the community. We find core and peripheral members of communities by creating a co-clustering matrix whose entries equals the number of clusterings in the landscape in which nodes u and v are co-clustered. Dense blocks in the matrix correspond to core members; cavities within dense blocks indicate peripheral activity or overlapping modules. Such matrices have been previously investigated in a different context — consensus clustering[28,9] — where the goal is typically to return a centroid clustering that lies centrally amongst a given set of input clusterings. Finding core and peripheral proteins within dense subgraphs in protein interaction networks has also recently been shown to be useful for protein complex identification[8,23,24]. We use the co-clustering matrix as a means to identify inter-and intra-module clustering dynamics.

3. Results

We used MOD-ILP with diversity constraints to produce modularity landscapes for the karate club social network[43], the ERK1/ERK2 MAPK[18] and Integrin[26] metabolic pathways, and a coarse-level human brain network[13]. For each network, we show how exploring ensembles of near-optimal solutions reveals clustering dynamics that would otherwise be missed by single solution approaches.

3.1. *Karate Club Network*

We begin by studying the modularity landscape of Zachary's karate club network[43], shown in Figure 1. This network consists of 34 nodes and 78 social-interaction edges. Due to an internal dispute over the price of karate lessons, the group split into two factions, one corresponding to the club's karate instructor, Mr. Hi, and the other to the club's officers. Although not a network derived from molecular biology, it has the advantage of being small enough to examine by hand and to have hand-curated evidence regarding social interactions and community membership.

The distance-based approach found 82 different clusterings, after which no more feasible clusterings existed. These clusterings had between 1 and 5 communities. Figure 2 shows the modularity landscape produced by MOD-ILP with diversity constraints using the distance-based approach. In each panel, the x-axis gives the solution number. The y-axis in the top panel shows the distance from each solution to the optimal solution; the y-axes of the middle and bottom panels show the solution's modularity and number of communities, respectively. The number of communities does not monotonically decrease with lower modularity. Instead, different components join or break-off as dictated by the resulting modularity and diversity constraints. This implies that our ensembles do not simply correspond to iteratively choosing different levels in the modularity hierarchical tree decomposition[6].

Although the true structure consisted of 2 communities, the optimal modularity solution (with modularity

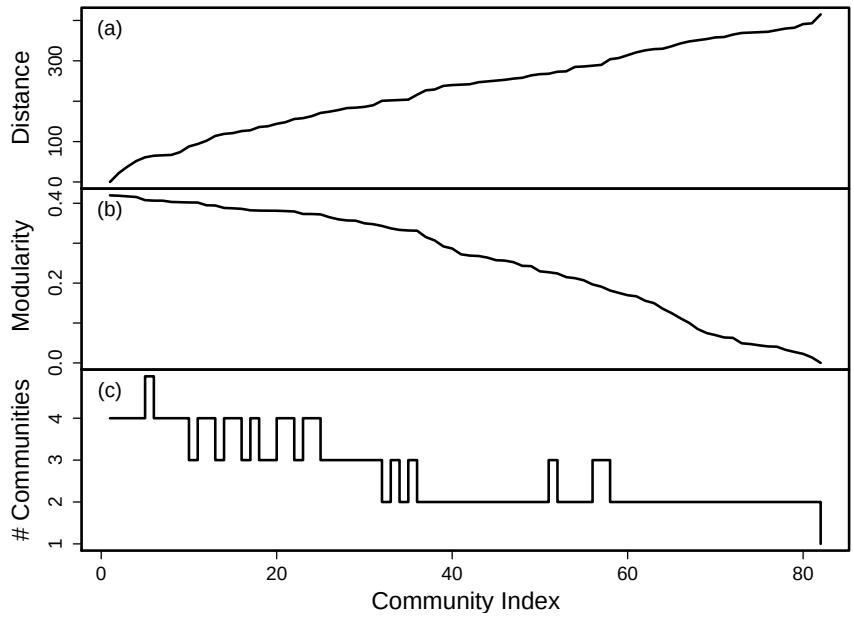

Figure 2. **Modularity landscape of the karate club network.** The x-axis in all panels shows the community index (ordered list of clusterings returned by iterative runs of MOD-ILP with distance-based diversity constraints). The 0^{th} community index corresponds to the optimal modularity clustering. (a) The Hamming distance from each clustering to the optimal modularity clustering. (b) The modularity of each clustering. There are 10 clusterings with modularity > 0.4, and 37 with modularity > 0.3. (c) The number of communities in each clustering.

0.419790) had four clusters (with each faction broken into two communities). The network is not split into the two communities until the 31^{st} solution. This solution has modularity 0.343195 and corresponds closely to the actual groups formed (with the exception of nodes $9, 10, 20$, and 31 — all topologically fringe, three of which were weak supporters of their faction leaders[43]). Such a solution would never be found unless near-optimal solutions were considered. Further, randomized rounding procedures would be unable to generate diverse solutions for this network, because even when the integrality constraints were relaxed, allowing $x_{uv} \in [0, 1]$, an integral solution was returned. This argues for the necessity of a constraint-based approach.

The point-based method, which only constrains each solution to be minimally different from all previous solutions, produced many more finer-grained solutions corresponding to incremental merging and splitting of communities. In fact, the 100^{th} solution of the point-based approach still had a modularity above 0.4. Although this level of detail could be useful for some applications, here we seek to more coarsely characterize the clustering dynamics, and therefore only further consider the distance-based solutions.

Dynamics for individual nodes can be better understood by looking at near-optimal solutions. For example, the solution with the provably second-best modularity, which is also the clustering that is output by Newman's spectral method[34], consists of 4 clusters but with slightly smaller modularity (0.418803) than the optimum. The difference lies in the classification of node 10, which, in the second-best clustering is placed with Mr. Hi and in the optimal clustering is placed with the officer's faction. Zachary measures the strength of friendship between pairs of individuals based on their interactions in other social contexts (for example, academic classes, student pubs, and other karate studios[43]) and finds that node 10 had nearly equal interaction with members from both factions. Node 10 was also not a strong believer in either faction's ideology (although he ultimately chose the officer's club after the fission). Hence, it makes sense that node 10 was the first to jump from one clustering to the other.

Another interesting case occurs for node 20. He lies in Mr. Hi's faction in the optimal clustering, but in subsequent clusterings is co-clustered with members from the officer's faction. According to Zachary, node 20 ultimately chooses Mr. Hi's club, but only weakly supported Mr. Hi's position in the dispute[43]. Looking

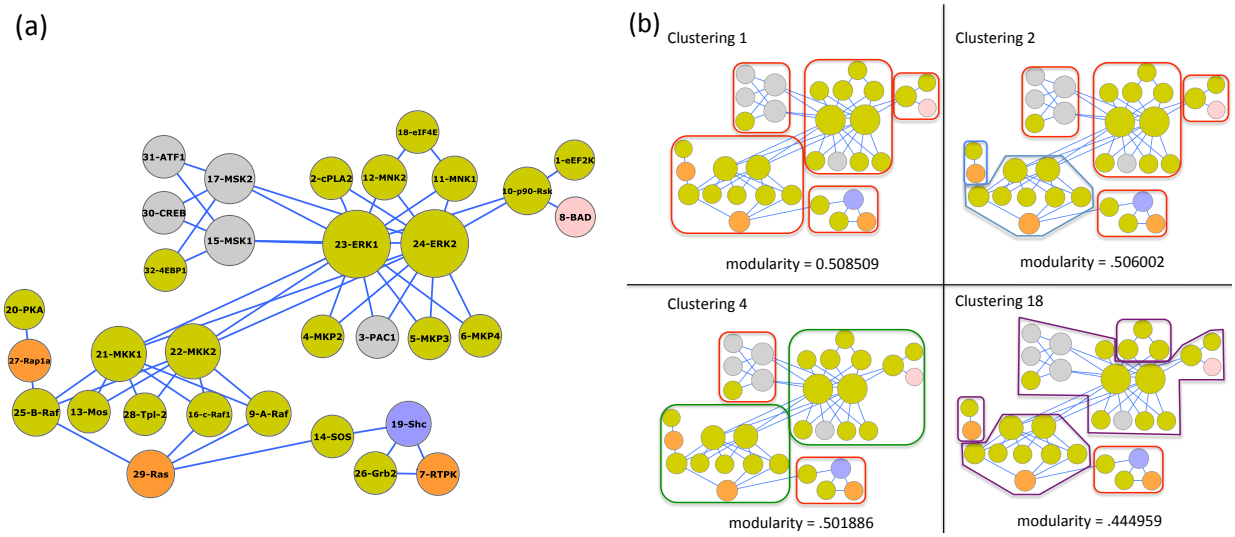

Figure 3. (a) **The ERK1/ERK2 MAPK signalling pathway[18].** The network consists of 32 nodes and 54 edges. The color of the node indicates the subcellular localization of the signalling component (green = cytosol, orange = plasma membrane, gray = nucleus, blue = plasma membrane translocation, and pink = mitochondrion). The network was drawn using Cytoscape[39]. (b) **"Flip-book" showing the clustering dynamics of the ERK1/ERK2 MAPK pathway**. Each of the four blocks corresponds to a clustering produced by MOD-ILP with distance-based diversity constraints. The number of the clustering is shown at the top, and its modularity on the bottom. Each cluster is blocked within a polygonal shape. A variety of near-optimal clusterings provide alternative, legitimate decompositions of the network.

at the network, 20 is connected to both faction leaders, plus an additional supporter of Mr. Hi. Topologically and anecdotally, it seems to make sense then that node 20 is a peripheral member of Mr. Hi's karate club.

Trying to identify core and peripheral nodes by only looking at the neighbors of a node, however, can be misleading. Node 3, for example, is a topologically fringe node with 10 total edges, 5 to members in both factions. But, according to Zachary[43], node 3 was a strong supporter of Mr. Hi, whose club he joined after fission. In our ensemble, we only see node 3 switch from a Mr. Hi-dominant clustering to a clustering dominated by officer members three times. These all occur near the end of the landscape, at clusterings 72, 78, and 80, which have a very low modularity (average = 0.030188). Using only network neighbors to classify a node as core or peripheral is therefore not always sufficient. Further, the landscape also provides a way to confidently say what groups of nodes do not belong together. A static analysis of the optimal clustering will clearly be unable to understand these type of community dynamics.

3.2. *Signalling Networks*

We considered the ERK1/ERK2 mitogen-activated protein kinase (MAPK) pathway[18] shown in Figure 3a. MAPK is a signal-transduction pathway that is highly-conserved across eukaryotes. MAPKs phosphorylate serines and threonines of target proteins and regulate a vast array of cellular functions, including gene expression, mitosis, and metabolism[19]. The extra-cellular signal-regulated kinases (ERKs) play a functional role in cell division, in particular meiosis and mitosis[19]. Identifying functional modules in such pathways is important because modules are often conserved across organisms, and thus can be used to generate new pathways from reference pathways[20,41]. The pathway consists of 32 nodes and 54 edges.

Figure 3b shows four snapshots of the modularity landscape. The optimal modularity (clustering 1) consists of five clusters roughly corresponding to nodes surrounding the Ras activation module, the Raf and MEK kinase modules, and the larger ERK module (split into three) — all known submodules of the pathway. In subsequent clusterings, nearby cores are either split or merged together, corresponding to finer- and coarser-grained functional subunits of the pathway. As in the karate network, we also find that the number of clusters does not simply monotonically decrease (or increase) as the diversity constraint, d_{changes},

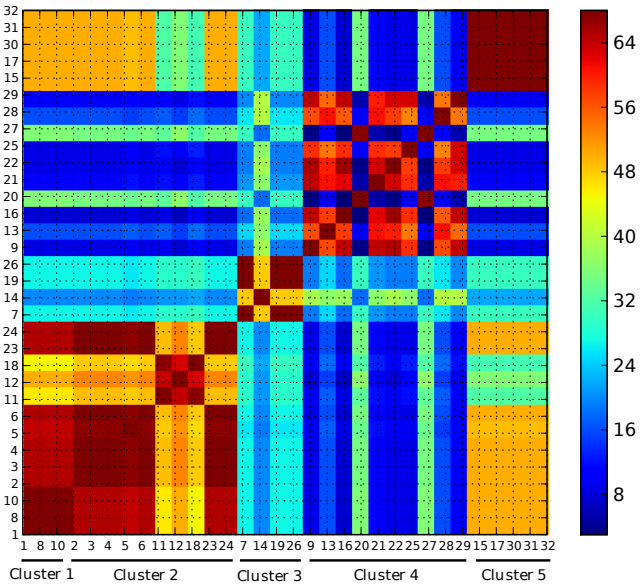

Figure 4. **Co-clustering heatmap for the ERK1/ERK2 MAPK pathway[18].** A broad view of how pairs of nodes traverse the modularity landscape. Each cell (u, v) in the heatmap corresponds to the number of clusterings in which nodes u and v were placed together. The nodes are ordered according to the optimal modularity found by MOD-ILP. Though outlines of the five optimal modules are present, the fluctuation of activity within and between the five blocks reveal interesting inter- and intra-community interactions.

is increased.

Figure 4 shows a global view of how the affiliation between each pair of nodes changes across clusterings. The intensity of cell (u, v) in the heatmap corresponds to the number of clusterings in the landscape in which nodes u and v are co-clustered. A similar picture was obtained by setting the intensity of a cell (u, v) to be the total modularity sum of all clusterings in which u and v were co-clustered. The nodes are ordered based on the clusters from the optimal modularity clustering.

The outlines of the five optimal blocks in Figure 4 provide a basic hint about the modular structure of the pathway, but it does not tell the whole story. For example, nodes 20 (PKA) and 27 (Rap1a) travel together much more than 27 and 13 (Mos), even though all three were placed together in the same optimal module. From the layout shown in Figure 3a, this makes sense — PKA and Rap1a are connected to the core Raf module by only one edge, and are also connected to each other. This suggests that they play a peripheral role in the module in which they were placed, or perhaps that they should be placed together in their own module.

The heatmap also provides a way to measure the confidence in a community by looking at how a group of nodes change their membership with respect to each other. For example, nodes 15, 17, 30, 31, and 32, corresponding to a portion of the ERK module, were co-clustered across all clusterings, as indicated by the solid red block in the upper-right corner of Figure 4. This implies that we are very confident in this module, more so than any other. Other clusters vary greatly with respect to how often their members travel together. An optimal clustering alone would yield a heatmap with solid red blocks for all clusters, which is much less informative of community membership strength.

We also looked at the Integrin signalling pathway[26], known to be vital for cell migration and growth. This pathway is longer and less dense than the ERK/MAPK pathway. The optimal modularity clustering found a reasonable decomposition consisting of modules with long chains of nodes. These long chains are often prefaced by *gatekeeper* nodes that branch off multiple non-overlapping paths. Network centrality measures

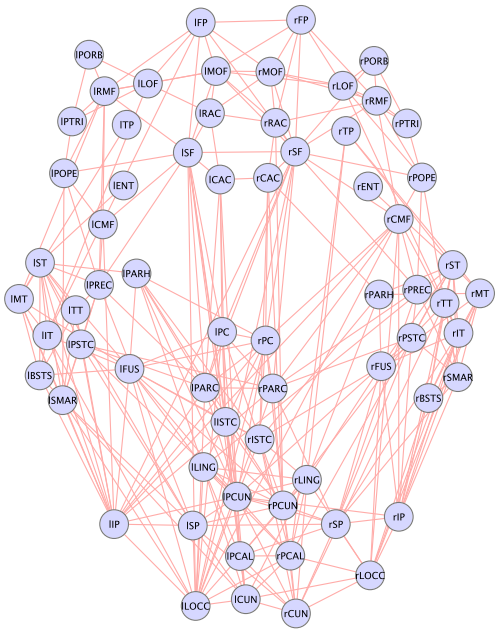

Figure 5. **Anatomical network of the human cerebral cortex[13].** The network consists of 66 nodes (brain regions) and 2,149 multiedges (dense axonal-pathways). The 'r' and 'l' prefixes correspond to the right and left hemispheres of the brain. The remaining portion of the names correspond to cortical regions (e.g. ENT = entorhinal cortex, TP = temporal pole, PC = posterior cingulate cortex, CUN = cuneus, and PARH = parahippocampal cortex). The layout is set to spatially agree with the actual positions of the regions in the brain. Coordinates of the regions were estimated from Figure 6 in Hagmann et al.[13], as well as from the Brede database[35]. The network was drawn using Cytoscape[39].

have also been used to globally identify highly "between" nodes,[42,14] though they do not typically take modules into account. Interestingly, many of the near-optimal clusterings identified these gatekeepers by placing them into different modules corresponding to the various branches. For example, the Cdc42 protein acts as a between-module node that ultimately leads to activation of actin and the c-Jun N-terminal kinases (JNK). It was first placed amongst nodes in the JNK module, but later switches into the actin module. A similar dynamic was seen for branches leading out of the focal adhesion kinase (FAK), which is involved in cellular adhesion and migration.

3.3. *Brain Network*

Lastly, we investigated a network representing the axonal-pathways within the cortex of the human brain. Brain maps have typically been constructed using functional magnetic resonance imaging (fMRI), which measures neural activity via blood flow (e.g.[10,7]). Recently, Hagmann et al.[13] used a technique called diffusion spectrum imaging (DSI) which identifies neuronal fiber trajectories by looking at the diffusion of water molecules in brain tissues. DSI produces a 3D water-flow gradient at specified positions in the brain to which tractography can be applied to recover the underlying neural tracts. Tractography identifies, for each position, the diffusion of water to that position from all other directions. Thus, we can determine the axonal trajectories across white matter, i.e. the connectivity across different regions in the brain. Regions are typically defined manually after white-gray matter segmentation (white = nerve connections, gray = congregations of neurons). The resulting network is composed of nodes (brain regions) and weighted edges, corresponding to the density of the connection between brain regions.

Hagmann et al.[13] applied their technique to generate a brain "connectome" for five human participants. Each connectome consists of 998 regions of interest. They also created a coarser network by condensing the 998 regions into 66 anatomical regions. An edge (u, v) in the anatomical network was weighted by computing the average of all edges that map to (u, v). To handle weighted networks, we used an extended version of

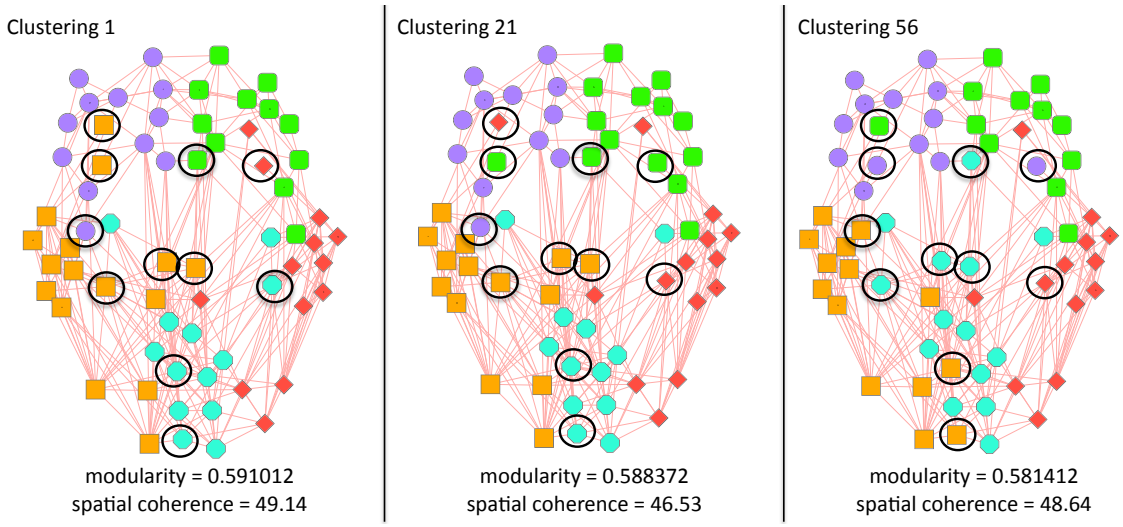

Figure 6. **"Flip-book" showing the clustering dynamics of the anatomical brain network[13].** Each of the three blocks corresponds to a clustering produced by MOD-ILP with distance-based diversity constraints. The number of the clustering is shown at the top, with modularity and spatial coherence on the bottom. Co-clustered nodes share the same color and shape. Black circles highlight nodes whose communities change across clusterings. The optimal modularity clustering does not have the highest coherence with the spatial coordinates of the regions.

modularity that converts weighted edges to unweighted, multi-edges[33]. In particular, in the multi-edged anatomical network we created $\lfloor 1000 \cdot w(u,v) \rfloor$ edges between nodes u and v, where $w(u,v)$ is the weight of edge (u,v) in the weighted anatomical network. The only change required in the definition of modularity is with A_{uv}, which is now the number of edges that go between u and v, instead of just 0 or 1. The final anatomical network contained 66 nodes and $2,149$ multiedges. Hagmann et al.[13] applied modularity to the anatomical network to identify regional hubs.

We ran MOD-ILP with diversity constraints on the first subject's human connectome (Figure 5). The similarity between the modularity values of the near-optimal solutions suggest extreme uncertainty in whether the optimal solution represents the true partitioning. Figure 6 shows the optimal clustering plus two near-optimal clusterings returned by the distance-based approach. The near-optimal clusterings are only slightly less topologically modular. In fact, amongst the first 60 solutions, we find that 53 are within 1% of the optimal modularity.

The brain network is unique among those that we consider because the nodes have a fixed spatial position. Hagmann et al.[13] assigned spatial coordinates to each region corresponding to its center of mass, but because not all spatial coordinates were available, the layout in Figure 5 is a taken from the layout drawn in Hagmann et al.[13] Three-dimensional spatial coordinates were directly available for 23 of the 66 regions. An additional 15 regions were assigned spatial coordinates based on averaging the coordinates from several studies for the relevant region using the Brede neuroimaging database[35].

The spatial coordinates themselves define a rough clustering, which can be used as an additional measure (along with modularity) to evaluate the likelihood of a particular brain network partitioning. We defined the *spatial coherence* of a clustering as the average Euclidean distance between anatomical regions placed in the same cluster. The near-optimal clusterings shown in Figure 6 have a better spatial coherence than the optimal solution at only a tiny decrease in modularity, despite having the same number of clusters. In fact, out of the 60 near-optimal solutions 30 of these solutions have a $> 1\%$ advantage in spatial coherence, 24 have a $> 2\%$ advantage, and 12 have a $> 3\%$ advantage. Naturally, clusters and nodes that do not match what is expected spatially may be the most interesting to investigate further. The nodes that are differentially clustered within the ensemble of solutions (circled in black in Figure 6) are typically such spatial outliers.

4. Conclusions

We investigated the clustering dynamics of four social and biological networks to reveal how these networks are organized. In all four settings, we showed how traversing the modularity landscape by explicitly constraining for diversity can be used to uncover deeper community structure that would otherwise be absent from single-solution or randomization-based procedures. In particular, we used ensembles of near-optimal network decompositions to identify resilient communities, core-peripheral community members, and finer- and coarser-grained community structure. We also found cases where near-optimal solutions corresponded better with known community structure than the optimal solution. We presented mostly anecdotal evidence regarding inter-and intra-module dynamics. Testing these notions on a large scale, such as for the automated identification of core-peripheral proteins in protein complexes[8,23,24], is a potential avenue for future work. It would also be interesting to characterize the relationship between clustering dynamics across the energy landscape and clustering dynamics across time. Nonetheless, we believe the insights provided by near-optimal solutions augment our current understanding of community structure and dynamics, and should not be ignored.

Acknowledgments

This work was partially supported by grants 0849899 and 0812111 from the National Science Foundation.

References

1. G. Agarwal and D. Kempe. Modularity-maximizing graph communities via mathematical programming. *Eur. Phys. J. B*, 66(3):409–418, 2008.
2. G. D. Bader and C. W. V. Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4:2, 2003.
3. U. Brandes, D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski, and D. Wagner. On modularity clustering. *IEEE T. Knowl. Data En.*, 20(2):172–188, 2008.
4. L. E. Carvalho and C. E. Lawrence. Centroid estimation in discrete high-dimensional spaces with applications in biology. *Proc. Natl. Acad. Sci. USA*, 105(9):3209–3214, 2008.
5. M. Charikar, V. Guruswami, and A. Wirth. Clustering with qualitative information. In *FOCS '03: Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science*, pages 524–533, 2003.
6. A. Clauset, M. E. Newman, and C. Moore. Finding community structure in very large networks. *Phys. Rev. E*, 70(6 Pt 2):066111, 2004.
7. M. D. Fox, A. Z. Snyder, J. L. Vincent, M. Corbetta, D. C. Van Essen, and M. E. Raichle. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proc. Natl. Acad. Sci. USA*, 102(27):9673–9678, 2005.
8. A. C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dumpelfeld, A. Edelmann, M. A. Heurtier, V. Hoffman, C. Hoefert, K. Klein, M. Hudak, A. M. Michon, M. Schelder, M. Schirle, M. Remor, T. Rudi, S. Hooper, A. Bauer, T. Bouwmeester, G. Casari, G. Drewes, G. Neubauer, J. M. Rick, B. Kuster, P. Bork, R. B. Russell, and G. Superti-Furga. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084):631–636, 2006.
9. A. Goder and V. Filkov. Consensus clustering algorithms: Comparison and refinement. In *ALENEX '08: Proceedings of the Workshop on Algorithm Engineering and Experiments*, pages 109–117. SIAM, 2008.
10. M. D. Greicius, B. Krasnow, A. L. Reiss, and V. Menon. Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proc. Natl. Acad. Sci. USA*, 100(1):253–258, 2003.
11. R. Guimerà and L. A. Nunes Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900, 2005.
12. S. T. Hadjitodorov, L. I. Kuncheva, and L. P. Todorova. Moderate diversity for better cluster ensembles. *Inform. Fusion*, 7(3):264–275, 2006.
13. P. Hagmann, L. Cammoun, X. Gigandet, R. Meuli, C. J. Honey, V. J. Wedeen, and O. Sporns. Mapping the structural core of human cerebral cortex. *PLoS Biol.*, 6(7):e159, 2008.
14. M. W. Hahn and A. D. Kern. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol. Biol. Evol.*, 22(4):803–806, 2005.
15. L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray. From molecular to modular cell biology. *Nature*, 402(6761 Suppl):C47–52, 1999.

16. J. Hopcroft, O. Khan, B. Kulis, and B. Selman. Tracking evolving communities in large linked networks. *Proc. Natl. Acad. Sci. USA*, 101 Suppl 1:5249–5253, 2004.
17. IBM Ilog, Inc. Solver CPLEX, 2009. http://www.ilog.com/products/cplex/ (accessed 7 July 2009).
18. G. L. Johnson. ERK1/ERK2 MAPK pathway. *Sci. Signal. Connections Map in the Database of Cell Signaling*, 2009.
19. G. L. Johnson and R. Lapadat. Mitogen-activated protein kinase pathways mediated by ERK, JNK, and p38 protein kinases. *Science*, 298(5600):1911–1912, 2002.
20. M. Kanehisa and S. Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nuc. Acids Res.*, 28(1):27–30, 2000.
21. B. Karrer, E. Levina, and M. E. J. Newman. Robustness of community structure in networks. *Phys. Rev. E*, 77(4):046119, 2008.
22. L. I. Kuncheva and S. T. Hadjitodorov. Using diversity in cluster ensembles. In *IEEE International Conference on Systems, Man, and Cybernetics*, volume 2, pages 1214–1219 vol.2, 2004.
23. H. C. Leung, Q. Xiang, S. M. Yiu, and F. Y. Chin. Predicting protein complexes from PPI data: a core-attachment approach. *J. Comp. Biol.*, 16(2):133–144, 2009.
24. F. Luo, B. Li, X. F. Wan, and R. H. Scheuermann. Core and periphery structures in protein interaction networks. *BMC Bioinformatics*, 10 Suppl 4:S8, 2009.
25. A. Makhorin. *GNU Linear Programming Kit, Version 4.26*. GNU Software Foundation, http://www.gnu.org/software/glpk/glpk.html.
26. K. H. Martin, J. K. Slack, S. A. Boerner, C. C. Martin, and J. T. Parsons. Integrin signaling pathway. *Sci. Signal. Connections Map in the Database of Cell Signaling*, 2009.
27. C. P. Massen and J. P. Doye. Identifying communities within energy landscapes. *Phys. Rev. E*, 71(4 Pt 2):046101, 2005.
28. S. Monti, P. Tamayo, J. P. Mesirov, and T. R. Golub. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, 52(1-2):91–118, 2003.
29. E. Nabieva, K. Jim, A. Agarwal, B. Chazelle, and M. Singh. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, 21 Suppl 1:i302–i310, 2005.
30. S. Navlakha, R. Rastogi, and N. Shrivastava. Graph summarization with bounded error. In *SIGMOD '08: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pages 419–432, New York, NY, USA, 2008. ACM.
31. S. Navlakha, M. C. Schatz, and C. Kingsford. Revealing biological modules via graph summarization. *J. Comp. Biol.*, 16(2):253–264, 2009.
32. S. Navlakha, J. White, N. Nagarajan, M. Pop, and C. Kingsford. Finding biologically accurate clusterings in hierarchical tree decompositions using the variation of information. In *RECOMB '09: Proceedings of the 13th Annual International Conference on Research in Computational Molecular Biology*, volume 5541, pages 400–417, 2009.
33. M. E. Newman. Analysis of weighted networks. *Phys. Rev. E*, 70(5 Pt 2):056131, 2004.
34. M. E. J. Newman. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA*, 103(23):8577–8582, June 2006.
35. F. A. Nielsen. The Brede database: a small database for functional neuroimaging. In *The 9th International Conference on Functional Mapping of the Human Brain*, 2003.
36. D. Opitz and R. Maclin. Popular ensemble methods: an empirical study. *J. Artif. Intell. Res.*, 11:169–198, 1999.
37. R. Polikar. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6(3):21–45, 2006.
38. Z. Qi and I. Davidson. A principled and flexible framework for finding alternative clusterings. In *KDD '09: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 717–726, New York, NY, USA, 2009. ACM.
39. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13(11):2498–2504, 2003.
40. S. Van Dongen. Graph clustering via a discrete uncoupling process. *SIAM J. Matrix Anal. A.*, 30(1):121–141, 2008.
41. T. Yamada, S. Goto, and M. Kanehisa. Extraction of phylogenetic network modules from prokayrote metabolic pathways. *Genome Inform.*, 15(1):249–258, 2004.
42. H. Yu, P. M. Kim, E. Sprecher, V. Trifonov, and M. Gerstein. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput. Biol.*, 3(4):e59, 2007.
43. W. W. Zachary. An information flow model for conflict and fission in small groups. *J. Anthropol. Res.*, 33:452–473, 1977.