Higher-order organization of complex networks

Austin R. Benson,1 David F. Gleich,2 Jure Leskovec3

Networks are a fundamental tool for understanding and modeling complex systems in physics, biology, neuroscience, engineering, and social science. Many networks are known to exhibit rich, lower-order connectivity patterns that can be captured at the level of individual nodes and edges. However, higher-order organization of complex networks—at the level of small network subgraphs—remains largely unknown. Here, we develop a generalized framework for clustering networks on the basis of higher-order connectivity patterns. This framework provides mathematical guarantees on the optimality of obtained clusters and scales to networks with billions of edges. The framework reveals higher-order organization in a number of networks, including information propagation units in neuronal networks and hub structure in transportation networks. Results show that networks exhibit rich higher-order organizational structures that are exposed by clustering based on higher-order connectivity patterns.

Networks are a standard representation of data throughout the sciences, and higher-order connectivity patterns are essential to understanding the fundamental structures that control and mediate the behavior of many complex systems (1–7). The most common higher-order structures are small network subgraphs, which we refer to as network motifs (Fig. 1A). Network motifs are considered building blocks for complex networks (1, 8). For example, feed-forward loops (Fig. 1A, $M_2$) have proven fundamental to understanding transcriptional regulation networks (9); triangular motifs (Fig. 1A, $M_1$–$M_2$) are crucial for social networks (4); open bidirectional wedges (Fig. 1A, $M_3$) are key to structural hubs in the brain (10); and two-hop paths (Fig. 1A, $M_3$–$M_5$) are essential to understanding air traffic patterns (5). Although network motifs have been recognized as fundamental units of networks, the higher-order organization of networks at the level of network motifs largely remains an open question.

Here, we use higher-order network structures to gain new insights into the organization of complex systems. We develop a framework that identifies clusters of network motifs. For each network motif (Fig. 1A), a different higher-order clustering may be revealed (Fig. 1B), which means that different organizational patterns are exposed, depending on the chosen motif.

Conceptually, given a network motif $M$, our framework searches for a cluster of nodes $S$ with two goals. First, the nodes in $S$ should participate in many instances of $M$. Second, the set $S$ should avoid cutting instances of $M$, which occurs when only a subset of the nodes from a motif are in the set $S$ (Fig. 1B). More precisely, given a motif $M$, the higher-order clustering framework aims to find a cluster (defined by a set of nodes $S$) that minimizes the following ratio:

$$\phi_M(S) = \text{cut}_M(S,\tilde{S})/\min[\text{vol}_M(S), \text{vol}_M(\tilde{S})]$$

where $\tilde{S}$ denotes the remainder of the nodes (the complement of $S$), $\text{cut}_M(S,\tilde{S})$ is the number of instances of motif $M$ with at least one node in $S$ and one in $\tilde{S}$, and $\text{vol}_M(S)$ is the number of nodes in instances of $M$ that reside in $S$. Equation 1 is a generalization of the conductance metric in spectral graph theory, one of the most useful graph partitioning scores (11). We refer to $\phi_M(S)$ as the motif conductance of $S$ with respect to $M$.

Finding the exact set of nodes $S$ that minimizes the motif conductance is computationally infeasible (12). To approximately minimize Eq. 1 and, hence, to identify higher-order clusters, we developed an optimization framework that provably finds near-optimal clusters [supplementary materials (13)].

We extend the spectral graph clustering methodology, which is based on the eigenvalues and eigenvectors of matrices associated with the graph (11), to account for higher-order structures in networks. The resulting method maintains the properties of traditional spectral graph clustering: computational efficiency, ease of implementation, and mathematical guarantees on the near-optimality of obtained clusters. Specifically, the clusters identified by our higher-order clustering framework satisfy the motif Cheeger inequality (14), which means that our optimization framework finds clusters that are at most a quadratic factor away from optimal.

The algorithm (illustrated in Fig. 1C) efficiently identifies a cluster of nodes $S$ as follows:

- Step 1: Given a network and a motif $M$ of interest, form the motif adjacency matrix $W_M$ whose entries $(i, j)$ are the co-occurrence counts of nodes $i$ and $j$ in the motif $M$: $(W_M)_{ij} = \text{number of instances of } M \text{ that contain nodes } i \text{ and } j$.

- Step 2: Form the Laplacian of the motif $M$:

$$L_M = D_M - W_M$$

where $D_M$ is the degree matrix of $M$.

- Step 3: Find the first $K$ eigenvectors of $L_M$: $\{v_1, \ldots, v_K\}$.

- Step 4: Set $S$ as the indices $\{i : v_1[i] > \alpha \}$ of the first eigenvector $v_1$ above some threshold $\alpha$.

The motif Cheeger inequality provides a bound on the motif conductance of the set $S$ identified by the algorithm.

Here, we refer to $S$ as a near-optimal higher-order cluster (13).

Fig. 1. Higher-order network structures and the higher-order network clustering framework. (A) Higher-order structures are captured by network motifs. For example, all 13 connected three-node directed motifs are shown here. (B) Clustering of a network based on motif $M_7$. For a given motif $M$, our framework aims to find a set of nodes $S$ that minimizes motif conductance, $\phi_M(S)$, which we define as the ratio of the number of motifs cut (filled triangles cut) to the minimum number of nodes in instances of the motif in either $S$ or $\tilde{S}$ (12). In this case, there is one motif cut. (C) The higher-order network clustering framework. Given a graph and a motif of interest (in this case, $M_3$), the framework forms a motif adjacency matrix ($W_M$) by counting the number of times two nodes co-occur in an instance of the motif. An eigenvector of a Laplacian transformation of the motif adjacency matrix is then computed. The ordering of the nodes provided by the components of the eigenvector (15) produces nested sets $S_r = \{s_1, \ldots, s_r\}$ of increasing size $r$. We prove that the set $S_r$ with the smallest motif-based conductance, $\phi_M(S_r)$, is a near-optimal higher-order cluster (13).
Step 2: Compute the spectral ordering $s$ of the nodes from the normalized motif Laplacian matrix constructed via $WM_{(15)}$.

Step 3: Find the prefix set of $s$ with the smallest motif conductance (the argument of the minimum), formally, $S := \arg \min_r \phi_M(S_r)$, where $S_r = \{s_1, ..., s_r\}$.

For triangular motifs, the algorithm scales to networks with billions of edges and, typically, only takes several hours to process graphs of such size. On smaller networks with hundreds of thousands of edges, the algorithm can process motifs up to size 9 (13). Although the worst-case computational complexity of the algorithm for triangular motifs is $\Theta(m^{1.5})$, where $m$ is the number of edges in the network, in practice, the algorithm is much faster. By analyzing 16 real-world networks where the number of edges $m$ ranges from 159,000 to 2 billion, we found the computational complexity to scale as $\Theta(m^{1.2})$. Moreover, the algorithm can easily be parallelized, and sampling techniques can be used to further improve performance (16).

The framework can be applied to directed, undirected, and weighted networks, as well as motifs (13). Moreover, it can also be applied to networks with positive and negative signs on the edges, which are common in social networks (friend versus foe or trust versus distrust edges) and metabolic networks (edges signifying activation versus inhibition) (13). The framework can be used to identify higher-order structure in networks where domain knowledge suggests the motif of interest. In the supplementary materials, we also show that when a domain-specific higher-order pattern is not known in advance, the framework can also serve to identify which motifs are important for the modular organization of a given network (13). Such a general framework allows complex higher-order organizational structures in a number of different networks by using individual motifs and sets of motifs. The framework and mathematical theory immediately extend to other spectral methods, such as localized algorithms that find clusters around a seed node (17) and

Fig. 2. Higher-order cluster in the *C. elegans* neuronal network. [See (29).] (A) The four-node bi-fan motif, which is overexpressed in neuronal networks (1). Intuitively, this motif describes a cooperative propagation of information from the nodes on the left to the nodes on the right. (B) The best higher-order cluster in the *C. elegans* frontal neuronal network based on the motif in (A). The cluster contains three ring motor neurons (RMEL, -V, and -R; cyan) with many outgoing connections, which serve as the source of information; six inner labial sensory neurons (IL2DL, -VR, -R, -DR, -VL, and -L; orange) with many incoming connections, serving as the destination of information; and four URA motor neurons (purple) acting as intermediaries. These RME neurons have been proposed as pioneers for the nerve ring (21), whereas the IL2 neurons are known regulators of nictation (22), and the higher-order cluster exposes their organization. The cluster also reveals that RIH serves as a critical intermediary of information processing. This neuron has incoming links from three RME neurons, outgoing connections to five of the six IL2 neurons, and the largest total number of connections of any neuron in the cluster. (C) Illustration of the higher-order cluster in the context of the entire network. Node locations are the true two-dimensional spatial embedding of the neurons. Most information flows from left to right, and we see that RMEV, -R, and -L and RIH serve as sources of information to the neurons on the right.
The framework can serve to identify a higher-order modular organization of networks. We apply the higher-order clustering framework to the *Caenorhabditis elegans* neuronal network, where the four-node “bi-fan” motif (Fig. 2A) is overexpressed (I). The higher-order clustering framework then reveals the organization of the motif within the *C. elegans* neuronal network. We find a cluster of 20 neurons in the frontal section with low bi-fan motif conductance (Fig. 2B). The cluster shows a way that nivation is controlled. Within the cluster, ring motor neurons (RMEL, -V, or -R), proposed pioneers of the nerve ring (21), propagate information to inner labial sensory neurons, regulators of nivation (22), through the neuron RII (Fig. 2C). Our framework contextualizes the importance of the bi-fan motif in this control mechanism.

The framework also provides new insights into network organization beyond the clustering of nodes based only on edges. Results on a transportation reachability network (23) demonstrate how it finds the essential hub interconnection airports (Fig. 3). These appear as extrema on the primary spectral direction (Fig. 3C) when two-hop motifs (Fig. 3A) are used to capture highly connected nodes and nonhubs. [The first spectral coordinate of the normalized motif Laplacian embedding was positively correlated with the airport city's metropolitan population with Pearson correlation 99% confidence interval (0.33, 0.53).] The secondary spectral direction identified the west-east geography in the North American flight network. [It was negatively correlated with the airport city's longitude with Pearson correlation 99% confidence interval (-0.66, -0.50).] On the other hand, edge-based methods confute geography and hub structure. For example, Atlanta, a large hub, is embedded next to Salina, a nonhub, with an edge-based method (Fig. 3D).

Our higher-order network clustering framework unifies motif analysis and network partitioning—two fundamental tools in network science—and reveals new organizational patterns and modules in complex systems. Prior efforts along these lines do not provide worst-case performance guarantees on the obtained clustering (24) and do not reveal which motifs organize the network (25) but rely on expanding the size of the network (26, 27). Theoretical results in the supplementary materials (13) also explain why classes of hypergraph partitioning methods are more general than previously assumed and how motif-based clustering provides a rigorous framework for the special case of partitioning directed graphs. Finally, the higher-order network clustering framework is generally applicable to a wide range of network types, including directed, undirected, weighted, and signed networks.

REFERENCES AND NOTES
S-Acylation of the cellulose synthase complex is essential for its plasma membrane localization

Manoj Kumar,1 Raymond Wightman,2 Ivan Atanassov,3,4 Anjali Gupta,1 Charlotte H. Hurst,4,5 Piers A. Hemsley,3,4 Simon Turner†

Plant cellulose microfibrils are synthesized by a process that propels the cellulose synthase complex (CSC) through the plane of the plasma membrane. How interactions between membranes and the CSC is regulated is currently unknown. Here, we demonstrate that all catalytic subunits of the CSC, known as cellulose synthase A (CESA) proteins, are S-acylated. Analysis of Arabidopsis CESA7 reveals four cysteines in variable region 2 (VR2) and two cysteines at the carboxy terminus (CT) as S-acylation sites. Mutating both the VR2 and CT cysteines permits CSC assembly and trafficking to the Golgi but prevents localization to the plasma membrane. Estimates suggest that a single CSC contains more than 100 S-acyl groups, which greatly increase the hydrophobic nature of the CSC and likely influence its immediate membrane environment.

Plant cellulose microfibrils are synthesized by a process that propels the cellulose synthase complex (CSC) through the plane of the plasma membrane. How interactions between membranes and the CSC is regulated is currently unknown. Here, we demonstrate that all catalytic subunits of the CSC, known as cellulose synthase A (CESA) proteins, are S-acylated. Analysis of Arabidopsis CESA7 reveals four cysteines in variable region 2 (VR2) and two cysteines at the carboxy terminus (CT) as S-acylation sites. Mutating both the VR2 and CT cysteines permits CSC assembly and trafficking to the Golgi but prevents localization to the plasma membrane. Estimates suggest that a single CSC contains more than 100 S-acyl groups, which greatly increase the hydrophobic nature of the CSC and likely influence its immediate membrane environment.

S-Acylation involves reversible addition of an acyl group, often palmitate or stearate, to a cysteine residue, which can affect protein structure or localization (4). A recent study identified many S-acylated proteins in plants (5), including CESA1 and CESA3, which are essential for cellulose synthesis in the primary cell wall (6). We used acyl–resin-assisted capture (acyl-RAC) assays (7) to confirm that CESA1 is S-acylated (Fig. S1) and showed that CESA6 is also S-acylated (Fig. 1A). Furthermore, all three CESA proteins required for cellulose synthesis in the secondary cell wall, CESA4, CESA7, and CESA8, are S-acylated (Fig. 1A), which demonstrates that S-acylation is a common feature of CESA proteins involved in cellulose synthesis in both primary and secondary cell walls.

CESA7 has 26 cysteines (fig. S2A). In order to identify S-acylated cysteines, we mutated individual CESA7 cysteines to serines and tested their ability to complement the cesaconsh1 mutant. None of the eight cysteines in the zinc finger domain (ZF) showed any significant complementation (Fig. 2A and figs. S3 and S4). The structure of the ZF-type zinc-finger domain from CESA7 (Protein Data Bank (PDB) ID: 1WEO) shows that all eight cysteines are involved in coordinating two zinc atoms, which makes them unlikely to be S-acylated. Consequently, we focused our subsequent analysis on other regions of CESA7. Here, we describe the modifications of CESA proteins and demonstrate their importance to the functioning of the CSC.

A cysteine-rich region lies within VR2 (8). The number of VR2 cysteines is conserved among orthologous CESAs from different species but varies between paralogous CESAs (table S1). There are four VR2 cysteines in CESA7 (fig. S2), and mutating them individually has no effect on cellulose biosynthesis (Fig. 2A and C). We hypothesized that if VR2 is a site of CESA S-acylation, the remaining VR2 cysteines may support sufficient S-acylation for CESA function. Consequently, we mutated all four VR2 cysteines in CESA7 (VR2C8). The VR2C8 mutant exhibited no complementation of the cesaconsh1 mutant (Fig. 2C), Thus, the cysteines in this region appear to be functionally redundant.

Having identified the VR2 and CT cysteines as potential S-acylation sites, we proceeded to determine if these sites were S-acylated. We generated a mutant in which both CT cysteines were mutated (CTC8). The CTC8 mutant did not complement the cesaconsh1 mutant (Fig. 2B). Using ACP-RAC assays we consistently found that S-acylation was dramatically reduced in the VR2C8 mutant, although some signal remained. The CTC8 mutants exhibited a smaller decrease in S-acylation (Fig. 1, B and C). We then constructed a mutant in which both the VR2 and CT cysteines were mutated

---

1Faculty of Life Sciences, The University of Manchester, Michael Smith Building, Oxford Road, Manchester M13 9PT, UK.
2Microscopy Core Facility, Sainsbury Laboratory, University of Cambridge, Bateman Street, Cambridge CB2 1LR, UK.
3Division of Plant Sciences, School of Life Sciences, University of Dundee, Dowell Street, Dundee, DD1 5EH, Scotland, UK.
4Cell and Molecular Sciences, The James Hutton Institute, Invergowrie, DD2 5DA, Scotland, UK.
5Present address: AgriBioInstitute, 8 Dragan Tzankov Boulevard, 1645 Sofia Bulgaria. (Corresponding author. Email: simonturner@manchester.ac.uk (S.T.); p.a.hemsley@dundee.ac.uk (P.A.H.).
**Editor's Summary**

**Resolving a network of hubs**

Graphs are a pervasive tool for modeling and analyzing network data throughout the sciences. Benson et al., developed an algorithmic framework for studying how complex networks are organized by higher-order connectivity patterns (see the Perspective by Pr Ćulj and Malod-Dognin). Motifs in transportation networks reveal hubs and geographical elements not readily achievable by other methods. A motif previously suggested as important for neuronal networks is part of a “rich club” of subnetworks. *Science*, this issue p. 163; see also p. 123.