Higher-order graph clustering with network motifs

Austin R. Benson
Cornell University
CS 6241
March 26, 2019

Joint research with
David Gleich, Purdue
Jure Leskovec, Stanford
Background. Networks are sets of nodes and edges (graphs) that model real-world systems.

**Social networks**
- nodes are people
- edges are friendships

**Currency**
- nodes are accounts
- edges are transactions

**Brains**
- nodes are neurons
- edges are synapses

**Electrical grid**
- nodes are power plants
- edges are transmission lines

Tim Meko, Washington Post
Networks are defined by nodes and edges, so we design our analysis, models, and algorithms in terms of nodes and edges.
Background. Networks are sets of nodes and edges (graphs) that model real-world systems.

Key insight [ Flake00; Newman04,06; many others…]. Networks for real-world systems have modules, clusters, communities.

• We want algorithms to uncover the clusters automatically.
• Main idea has been to optimize metrics involving the number of nodes and edges in a cluster. Conductance, modularity, density, ratio cut, …
**Background.** Networks are sets of nodes and edges (graphs) that model real-world systems.

**Key insight** [Milo+02]. Networks modelling real-world systems contain certain small subgraphs patterns way more frequently than expected.

Triangles in social relationships.  
Bi-directed length-2 paths in brain networks. 
Signed feed-forward loops in genetic transcription. 

We call these small subgraph patterns **motifs**.
Motifs are the fundamental units of complex networks.

We should design our clustering algorithms around motifs.
Higher-order graph clustering is our technique for finding clusters based on motifs.

Different motifs give different clusters.
Higher-order graph clustering
Main points and overview

- We will generalize spectral clustering, a classical technique to find clusters or communities in a graph, to use motifs as the fundamental unit to partition.
- Based on a higher-order (motif-based) conductance metric that generalizes the traditional conductance.
- Comes with theoretical guarantees.

- We’ll first briefly review how spectral clustering works.
- Then we’ll see how to adapt it to work with network motifs.
- Then we’ll see the impact of this approach on various real-world data.
Background. Spectral clustering is a classic technique to partition graphs by looking at eigenvectors.

[Fiedler 1973, many more…]
Background. The (normalized) graph Laplacian.

Recall from lecture that $A$ is the adjacency matrix. $A_{ij} = 1$ if (i, j) is an edge in the graph, 0 otherwise.

Our fundamental matrices…

$D = \text{diag}(A1)$ \hspace{1cm} Diagonal degree matrix (1 is the vector of all ones).

$L = D - A$ \hspace{1cm} The graph Laplacian

$\mathcal{L} = D^{-1/2}LD^{-1/2}$ \hspace{1cm} The normalized graph Laplacian
Conductance is one of the most important cluster quality scores [Schaeffer07] used in Markov chain theory, spectral clustering, bioinformatics, vision, etc.

**Background.** Spectral clustering works based on conductance

The conductance of a *set of vertices* $S$ is the ratio of edges leaving to total edges

$$\phi(S) = \frac{\text{cut}(S)}{\min(\text{vol}(S), \text{vol}(\bar{S}))}$$

(edges leaving $S$) (edge end points in $S$)

small conductance $\Leftrightarrow$ good cluster

$\text{cut}(S) = 7$
$\text{vol}(S) = 85$
$\text{vol}(\bar{S}) = 151$
$\phi(S) = 7/85$
Background. Conductance and expansion are similar.

**Conductance.**

\[
\phi(S) = \frac{\text{cut}(S)}{\min(\text{vol}(S), \text{vol}(\bar{S}))}
\]

(edges leaving \(S\))

(edge end points in \(S\))

**Expansion.**

\[
\alpha(S) = \frac{\text{cut}(S)}{\min(|S|, |\bar{S}|)}
\]

(edges leaving \(S\))

(nodes in \(S\))

**Normalized graph Laplacian.**

\[
D = \text{diag}(A1)
\]

\[
\mathcal{L} = D^{-1/2}(D - A)D^{-1/2}
\]

**Graph Laplacian.**

\[
D = \text{diag}(A1)
\]

\[
L = D - A
\]
Background. Spectral clustering has theoretical guarantees [Cheeger70, Alon-Milman85]

Finding the smallest conductance set is NP-hard. 😞

- Cheeger realized the eigenvalues of the Laplacian provided surface area to volume bounds in manifolds.
- Alon and Milman independently realized the same thing for a graph (conductance)!

Eigenvalues of the Laplacian $\mathcal{L}$

$0 = \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n \leq 2$

$\phi_* = $ set of smallest conductance

$\phi_*^2/2 \leq \lambda_2 \leq 2\phi_*$

Cheeger Inequality

$D = \text{diag}(A1)$

$\mathcal{L} = D^{-1/2}(D - A)D^{-1/2}$
We can find a set $S$ that achieves the Cheeger bound.

1. Compute the eigenvector $z$ associated with $\lambda_2$ and scale to $f = D^{-1/2}z$
2. Sort the vertices by their values in $f$: $\sigma_1, \sigma_2, \ldots, \sigma_n$
3. Let $S_r = \{\sigma_1, \ldots, \sigma_r\}$ and compute the conductance of $\phi(S_r)$ of each $S_r$.
4. Pick the set $S_m$ with minimum conductance.

$$\phi(S_m) \leq 2\sqrt{\phi_*}$$
Background. The sweep cut visualized

[Mihail89, Chung92]

\[ \phi(S) = \frac{\text{cut}(S)}{\min(\text{vol}(S), \text{vol}(\overline{S}))} \]
Spectral clustering is theoretically justified for finding edge-based clusters in undirected, simple graphs.

We want to cluster with richer data
Motifs that may be directed, signed, colored, feature-valued, etc.

Signed feed-forward loops in genetic transcription [Mangan+03]
Gene X activates transcription in gene Y.
Gene X suppresses transcription in gene Z.
Gene Y suppresses transcription in gene Z.
Our contributions


- A generalized conductance metric for motifs.
- A new spectral clustering algorithm to minimize the generalized conductance.
- AND an associated motif Cheeger inequality guarantee.
- Naturally handles directed, signed, colored, weighted, and combinations of motifs.
- Scales to networks with billions of edges.
- Applications in ecology, biology, and transportation.
How do we find clusters based on motifs?
Motif-based conductance

Need new notions of cut and volume

\[
\phi(S) = \frac{\text{cut}(S)}{\min(\text{vol}(S), \text{vol}(S))}
\]

\[
\phi_M(S) = \frac{\text{cut}_M(S)}{\min(\text{vol}_M(S), \text{vol}_M(\bar{S}))}
\]

\[
\text{cut}(S) = \#(\text{edges cut})
\]

\[
\text{cut}_M(S) = \#(\text{motifs cut})
\]

\[
\text{vol}(S) = \#(\text{edge end points in } S)
\]

\[
\text{vol}_M(S) = \#(\text{motif end points in } S)
\]
Motif-based conductance

Figure 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, $\phi_M(S) = \frac{\text{motifs cut}}{\text{motif volume}} = \frac{1}{8}$. C: The higher-order network clustering framework. Given a graph and a motif of interest (in this case, $M_7$), 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 produces nested sets $S_r = \{1, \ldots, r\}$ 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.
Higher-order clustering

**Problem**  Given a motif $M$ and a graph $G$, we want to find a set of nodes $S$ that minimizes motif conductance. This is NP-hard. [Wagner-Wagner93]

**Our solution.** Generalize spectral clustering for motifs

1. Form new weighted, undirected graph $W^{(M)}$ based on $M$ and $G$
2. Compute Fiedler vector of Laplacian matrix of $W^{(M)}$ [Fiedler73, Alon-Milman85]
3. Use “sweep cut” procedure to output clusters [Mihail89, Chung92]

**Theorem (motif Cheeger inequality)**
resulting clusters will obtain near optimal motif conductance
Motif-based spectral clustering

**Step 1.** Given directed graph $G$ and motif $M$, form a weighted graph $W^{(M)}$.

**Step 1.**

Given directed graph $G$ and motif $M$, form a weighted graph $W^{(M)}$.

$$W_{ij}^{(M)} = \#\{\text{instances of motif } M \text{ that contain nodes } i \text{ and } j\}$$
Motif-based spectral clustering

**Step 1.** Given directed graph $G$ and motif $M$, form a weighted graph $W^{(M)}$.

**Key insight**

Classical spectral clustering on weighted graph $W^{(M)}$ finds clusters of low motif conductance.

$$\phi_M(S) = \frac{\text{motifs cut}}{\text{motif volume}}$$

$$W_{ij}^{(M)} = \#\{\text{instances of motif } M \text{ that contain nodes } i \text{ and } j\}$$
Motif-based spectral clustering

Step 2. Compute the eigenvector $f^{(M)}$ associated with $\lambda_2$ of the normalized Laplacian matrix of $W^{(M)}$

\[
D = \text{diag}(W^{(M)}1) \\
\mathcal{L}^{(M)} = D^{-1/2}(D - W^{(M)})D^{-1/2} \\
\mathcal{L}^{(M)}Z = \lambda_2 Z \\
f^{(M)} = D^{-1/2}Z
\]

Takes roughly $O(\# \text{ edges})$ time.
Motif-based spectral clustering

Step 3 (motif sweep cut) [Mihail89, Chung92]

- Sort nodes by values in $f^{(M)} \rightarrow \sigma_1, \sigma_2, \ldots \sigma_n$.
- Pick set $S_r = \{\sigma_1, \ldots, \sigma_r\}$ with smallest motif conductance.

$\sigma = (4, 5, 1, 3, 2, 7, 6, 9, 8, 10)$
Motif Cheeger inequality

**Theorem** If the motif has three nodes, then the sweep procedure on the weighted graph finds a set \( S \) of nodes for which

\[
\phi_M(S) \leq 2\sqrt{\phi_M^*}
\]

For 4+ nodes, need slightly different notion of conductance.

Key Proof Step

\[
M(G) = \{\text{instances of } M \text{ in } G\}
\]

\[
cut_M(S, G) = \sum_{\{i,j,k\} \in M(G)} \text{Indicator}[x_i, x_j, x_k \text{ not the same}]
\]

\[
= \frac{1}{4}(x_i^2 + x_j^2 + x_k^2 - x_i x_j - x_j x_k - x_k x_i)
\]

= quadratic in \( x \)
Applications

1. We do not know the motif of interest.
   food webs and new applications

2. We know the motif of interest from domain knowledge.
   yeast transcription regulation networks, connectome, social networks

3. We seek richer information from our data.
   transportation networks and new applications
Application 1

We do not know the motif of interest.
Application 1. Food webs

Florida bay food web

- Nodes are species
- Edges represent carbon exchange $i \rightarrow j$ if $j$ eats $i$
- Motifs represent energy flow patterns

http://marinebio.org/oceans/marine-zones/
Application 1. Food webs

Which motif clusters the food web?

Our approach

- Run motif spectral clustering for all 3-node motifs as well as for just edges.
- Examine the sweep profile to see which motif gives the best clusters.
Application 1. Food webs

**Our finding.** Motif $M_6$ organizes the food web into good clusters.
Application 1. Food webs

Motif $M_6$ reveals aquatic layers

61% accuracy vs. 48% with edge-based methods
Application 2

We know the motif of interest from domain knowledge.
Application 2. Yeast transcription regulation networks

- Nodes are groups of genes
- Edge $i \rightarrow j$ means $i$ regulates transcription to $j$
- Sign + / - denotes activation / suppression
- Coherent feedforward loops encode biological function
  [Mangan+03, Alon07]
Clustering based on coherent feedforward loops identifies functions studied individually by biologists [Mangan+03] 97% accuracy vs. 68–82% with edge-based methods
Figure S8: Higher-order organization of the *S. cerevisiae* transcriptional regulation network.

A: The four higher-order structures used by our higher-order clustering method, which can model signed motifs. These are coherent feedforward loop motifs, which act as sign-sensitive delay elements in transcriptional regulation networks (46). The edge signs refer to activation (positive) or repression (negative).

B: Six higher-order clusters revealed by the motifs in (A). Clusters show functional modules consisting of several motifs (coherent feedforward loops), which were previously studied individually (45). The higher-order clustering framework identifies the functional modules with higher accuracy (97%) than existing methods (68–82%).

C–D: Two higher-order clusters from (B). In these clusters, all edges have positive sign. The functionality of the motifs in the modules correspond to drug resistance (C) or cell cycle and mating type match (D). The clustering suggests that coherent feedforward loops function together as a single processing unit rather than as independent elements.
Application 3

We seek richer information from our data.
Application 3. Transportation networks

- North American air transport network.
- Nodes are cities.
- \( i \to j \) if you can travel from \( i \) to \( j \) in < 8 hours.

[Frey-Dueck07]
Application 3. Transportation networks

Weighted adjacency matrix already reveals hub-like structure

$$W_{i,j}^{(M)} = \#\{\text{bi-directional length-2 paths from } i \text{ to } j\}$$

Important motifs from literature [Rosvall+14]
Application 3. Transportation networks

West coast non-hubs
- Monterey, CA
- Redding, CA
- Kodiak, AK
- Salina, KS
- Morgantown, WV
- Atlantic City, NJ

East coast non-hubs
- New York, NY
- Denver, CO
- Chicago, IL
- Dallas, TX
- San Francisco, CA
- Charlotte, NC

Top 8 U.S. hubs
- Atlanta, GA
- Charlotte, NC
- Atlanta, GA
- New York, NY
- San Francisco, CA
- Chicago, IL
- Phoenix, AZ
- Denver, CO

Atlanta, the top hub, is next to Salina, a non-hub.

Primary spectral coordinate

Secondary spectral coordinate

MOTIF SPECTRAL EMBEDDING

EDGE SPECTRAL EMBEDDING
Application 3. Transportation networks
Applications 4, 5 & 6

Just some extra fun things we found.
Application 4. Anomaly detection in social networks

The up-linked triangle finds an anomalous cluster in Twitter.

Anomalous cluster in the 1.4B edge Twitter graph. All nodes are holding accounts for a company, and the orange nodes have incomplete profiles.
Application 5. Hierarchical structure in web graphs

The “uplinked triangle” has been observed to occur much more frequently than in random graph models. [Milo+02]

Periphery groups link to each other.

Core group with large in-degree.
Application 6. Nictation control in a neural network

We find the control mechanism that explains nictation based on the bi-fan motif (Milo et al. found it over-expressed).
Recap. Higher-order graph clustering

- Generalization of graph clustering to higher-order structures (motifs) through a new objective (motif conductance).
- Generalizing old ideas from spectral graph theory admits a new algorithm and a motif Cheeger inequality.
- Applications in ecology, biology, transportation, social networks, the Web, and neuroscience.
Key takeaways

- Organizing graphs according to motifs reveals new insights into data
- Simple & scalable framework with theoretical guarantees
- Impact in the community
  - Motif-Based Analysis of Effective Connectivity in Brain Networks, Meier et al., 2016
  - Motif correlation clustering Li et al., 2016
  - Network analytics in the age of big data, Pržulj & Malod-Dognin, 2016

Higher-order clustering


Code + data [http://snap.stanford.edu/higher-order](http://snap.stanford.edu/higher-order)
Intermission…
Timestamped connections are everywhere

**Private communication**
- e-mail, phone calls, text messages, instant messages

**Biology**
- cell signaling

**Public communication**
- Q&A forums, Facebook walls, Wikipedia edits

**Technical infrastructure**
- packets over the Internet, messages over supercomputer

**Payments**
- credit card transactions, Bitcoin, Venmo
Current methods for analyzing temporal networks

1. **Models for network growth**
   Growth of academic collaborations, Internet infrastructure, etc. [Leskovec+07]

2. **Sequence of snapshot aggregates**
   Daily phone call graph [Araujo+14], Per-year co-authorship [Dunlavy+2010]

**Opportunity** these methods do not capture the pulse of temporal networks that are constantly in motion.
How can we generalize motifs for temporal networks to provide a new type of analysis?
Temporal networks are lists of directed edges with timestamps

<table>
<thead>
<tr>
<th>source</th>
<th>destination</th>
<th>timestamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>d</td>
<td>14s</td>
</tr>
<tr>
<td>c</td>
<td>a</td>
<td>15s</td>
</tr>
<tr>
<td>a</td>
<td>c</td>
<td>17s</td>
</tr>
<tr>
<td>a</td>
<td>b</td>
<td>25s</td>
</tr>
<tr>
<td>a</td>
<td>c</td>
<td>28s</td>
</tr>
<tr>
<td>a</td>
<td>c</td>
<td>30s</td>
</tr>
<tr>
<td>c</td>
<td>d</td>
<td>31s</td>
</tr>
<tr>
<td>c</td>
<td>a</td>
<td>32s</td>
</tr>
<tr>
<td>a</td>
<td>c</td>
<td>35s</td>
</tr>
</tbody>
</table>

many timestamps between the same pair of nodes!

Timestamps are fine-grained
1 second resolution and O(years) span
Temporal network motifs

<table>
<thead>
<tr>
<th>source</th>
<th>destination</th>
<th>timestamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>d</td>
<td>14s</td>
</tr>
<tr>
<td>c</td>
<td>a</td>
<td>15s</td>
</tr>
<tr>
<td>a</td>
<td>c</td>
<td>17s</td>
</tr>
<tr>
<td>a</td>
<td>b</td>
<td>25s</td>
</tr>
<tr>
<td>a</td>
<td>c</td>
<td>28s</td>
</tr>
<tr>
<td>a</td>
<td>c</td>
<td>30s</td>
</tr>
<tr>
<td>c</td>
<td>d</td>
<td>31s</td>
</tr>
<tr>
<td>c</td>
<td>a</td>
<td>32s</td>
</tr>
<tr>
<td>a</td>
<td>c</td>
<td>35s</td>
</tr>
</tbody>
</table>

**Temporal network motif**
1. Directed multigraph with k edges
2. Edge ordering
3. Maximum time span \( \delta \)

**Motif instance** k temporal edges that match the pattern that all occur within \( \delta \) time

Wrong order! (c, a) before (a, c)

Paranjape, Benson, & Leskovec, WSDM, 2017
Algorithmic challenge of temporal motifs

**Given** a temporal network and a temporal network motif, count the number of motif instances in the network.

3 instances
Summary of new algorithms

In a network with \( m \) temporal edges and \( T \) static triangles and a motif with \( k \) temporal edges.

1. General algorithm for any motif.  
   faster than \( O(m^k) \) brute force approach
   
   2-nodes, \( k \) temporal edges. \( O(k^2 m) \),  
   linear time in size of data for const. \( k \)

Optimized algorithms for special cases

2. 3 nodes, 3 temporal edges, stars. \( O(m) \)  
   linear time in size of data

3. 3 nodes, 3 temporal edges, triangles. \( O(T^{1/2}m) \)  
   faster than previous state-of-the-art \( O(Tm) \)
New algorithms let us analyze large datasets

Processing a phone call network with 2 billion temporal edges takes just a few hours (single threaded).
Temporal motifs expose one-to-one and one-to-many behavior in communication systems

\[ \delta = 1 \text{ hour} \]
Temporal network motifs
Code + data  http://snap.stanford.edu/temporal-motifs

Key takeaways

- Temporal network motifs are a simple and effective way to analyze temporal networks, a data type for which we have few tools.

- Requires algorithmic insights to scale to large networks.