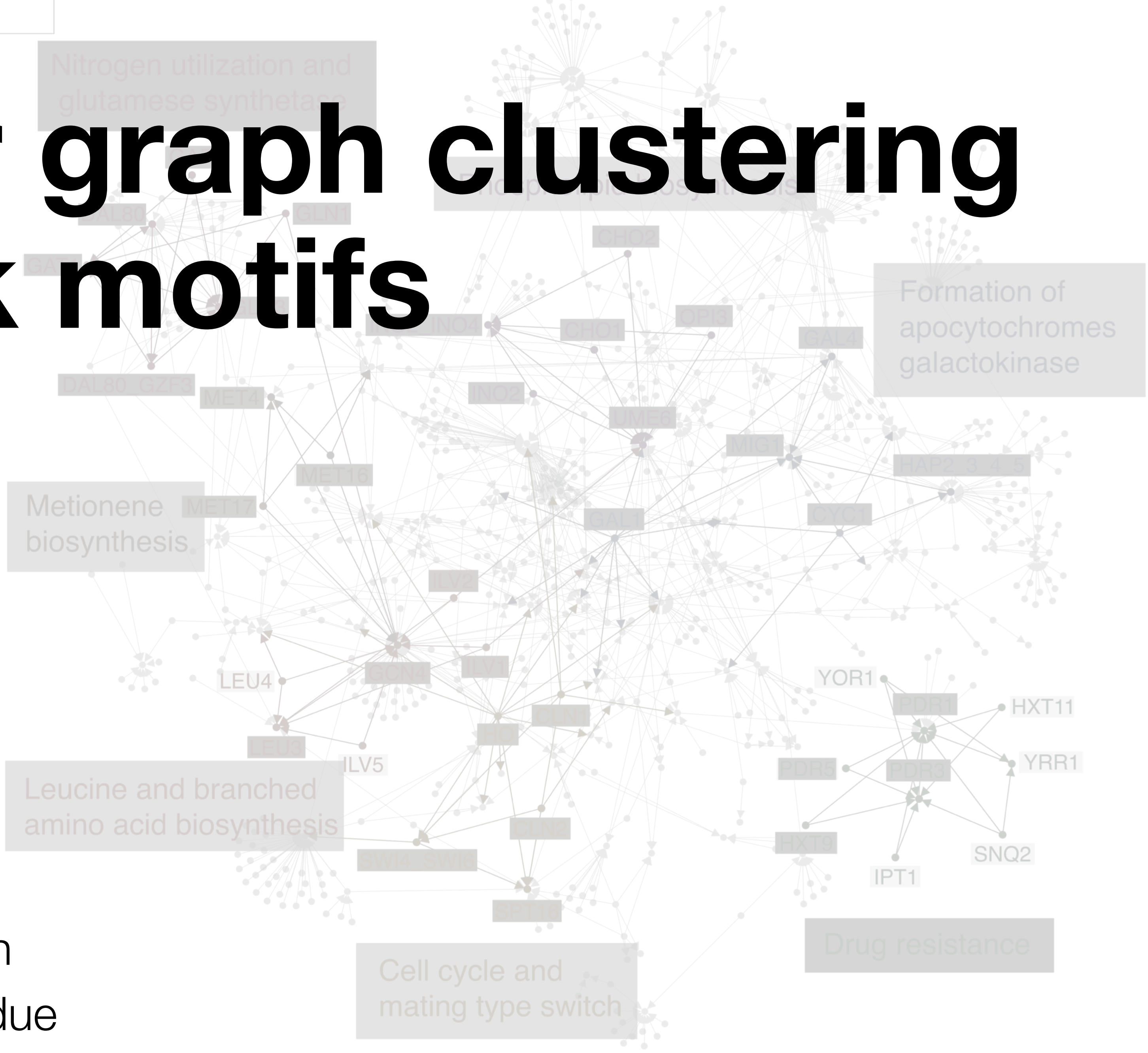
Higher-order graph clustering with network motifs

Austin R. Benson
Cornell University

CS 6241

March 26, 2019

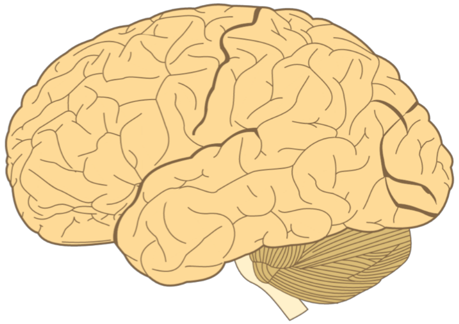


Joint research with
David Gleich, Purdue
Jure Leskovec, Stanford

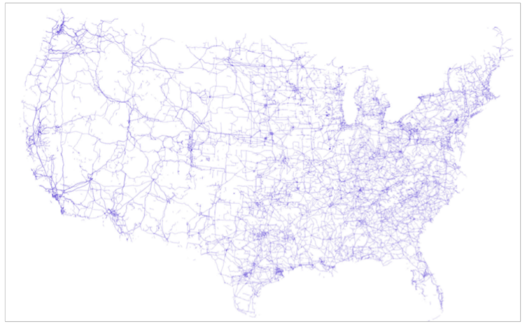
Background. Networks are sets of nodes and edges (graphs) that model real-world systems.

Social networks
nodes are people
edges are friendships

Currency
nodes are accounts
edges are transactions



Brains
nodes are neurons
edges are synapses



Tim Meko, Washington Post

Electrical grid
nodes are power plants
edges are transmission lines

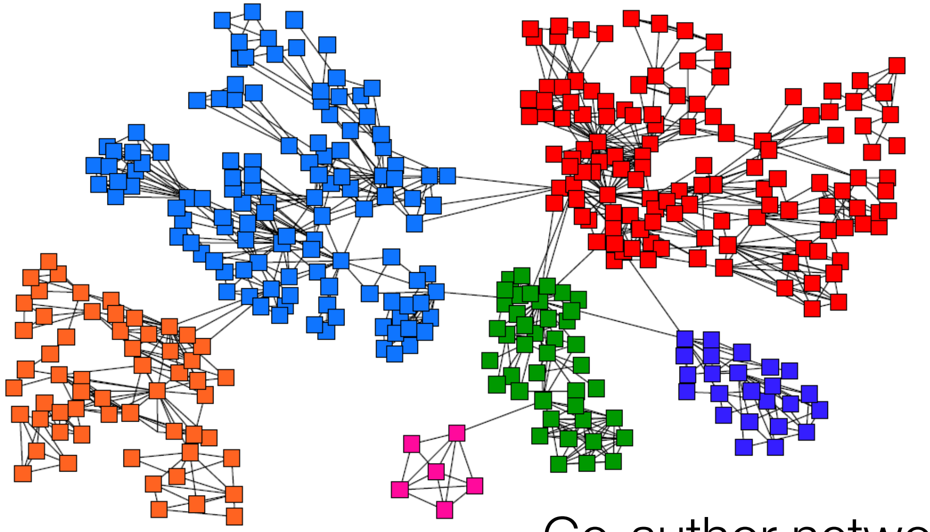
Networks are defined by nodes and edges,
so we design our
analysis, models, and algorithms
in terms of nodes and edges.

Background. Networks are sets of nodes and edges (graphs) that model real-world systems.

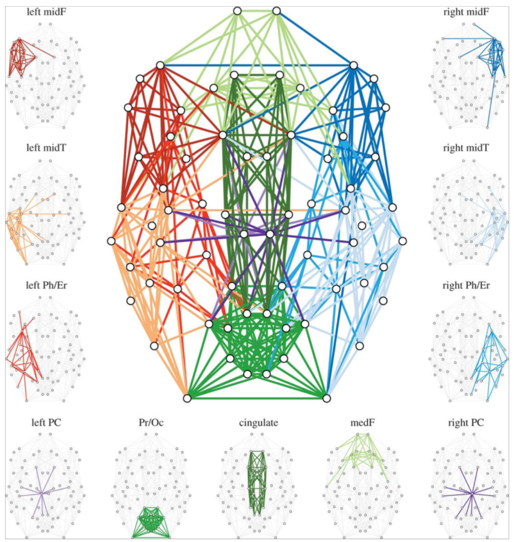
Key insight [Flake00; Newman04,06; many others…].

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, …



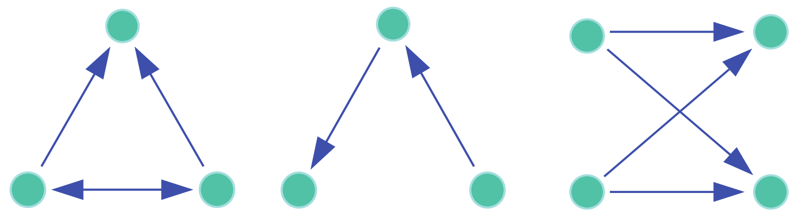
Co-author network



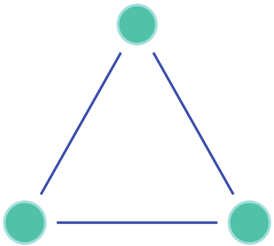
Brain network, de Reus et al., *RSTB*, 2014.

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.

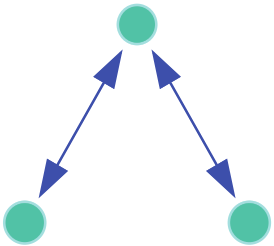


Triangles in social relationships.



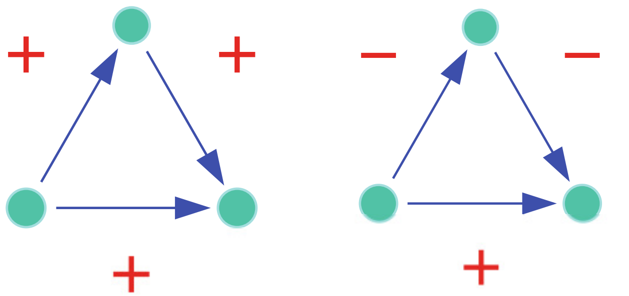
[Simmel 1908;
Rapoport 1953;
Granovetter 1973]

Bi-directed length-2 paths in brain networks.



[Sporns-Kötter 2004;
Sporns+ 2007; Honey+ 2007]

Signed feed-forward loops in genetic transcription.



[Mangan+ 2003; Alon 2007]

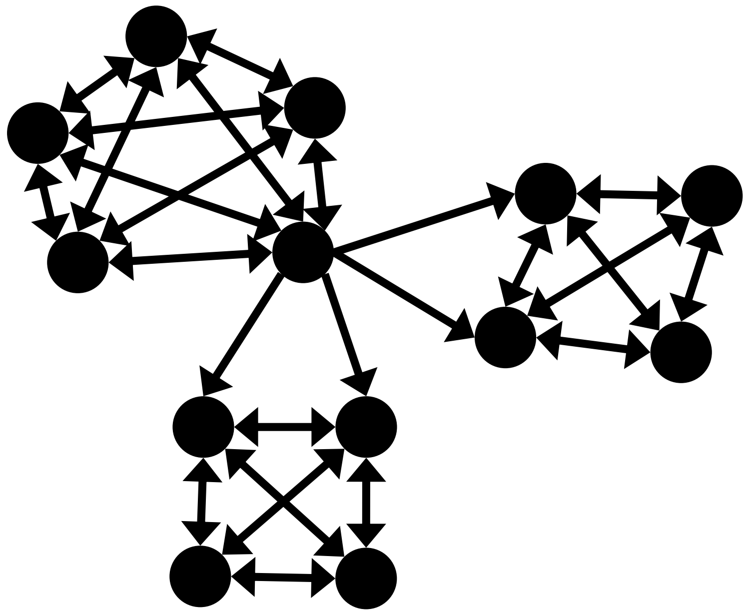
We call these small subgraph patterns *motifs*.

Motifs are the fundamental units of complex networks.

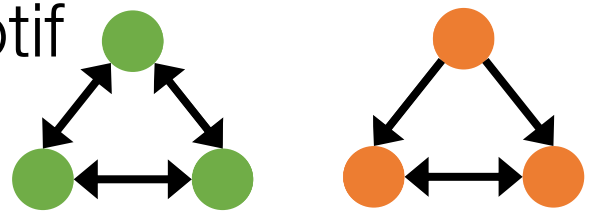
We should design our clustering algorithms around motifs.

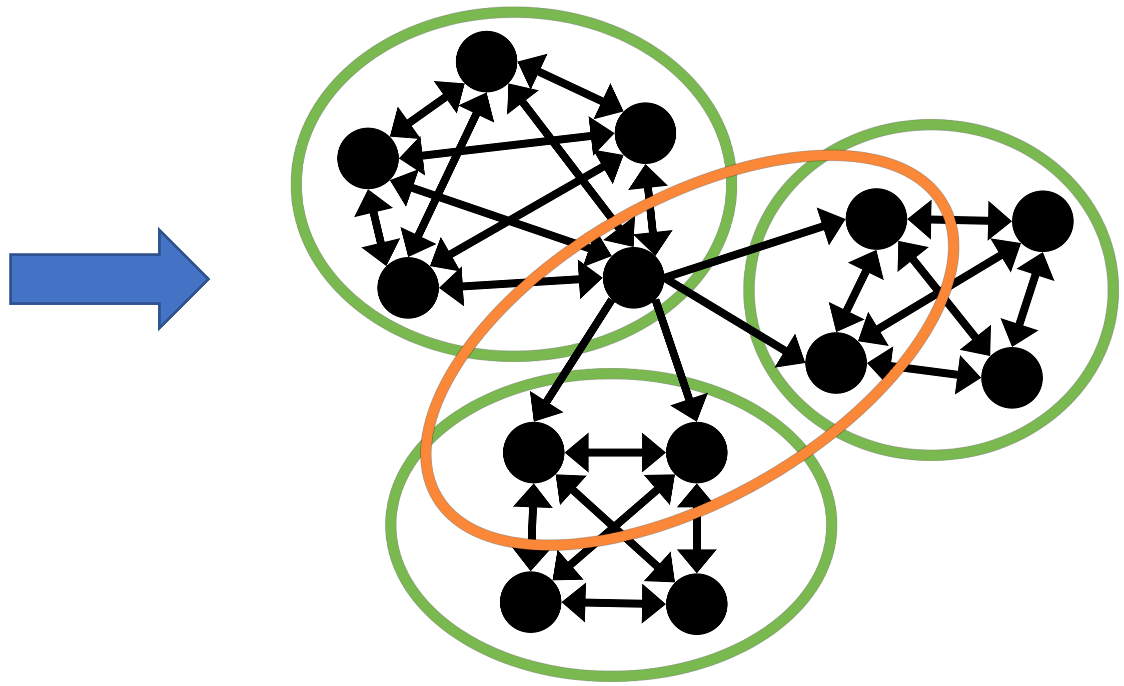
Higher-order graph clustering is our technique for finding clusters based on motifs

Network



Motif





Different motifs give different clusters.

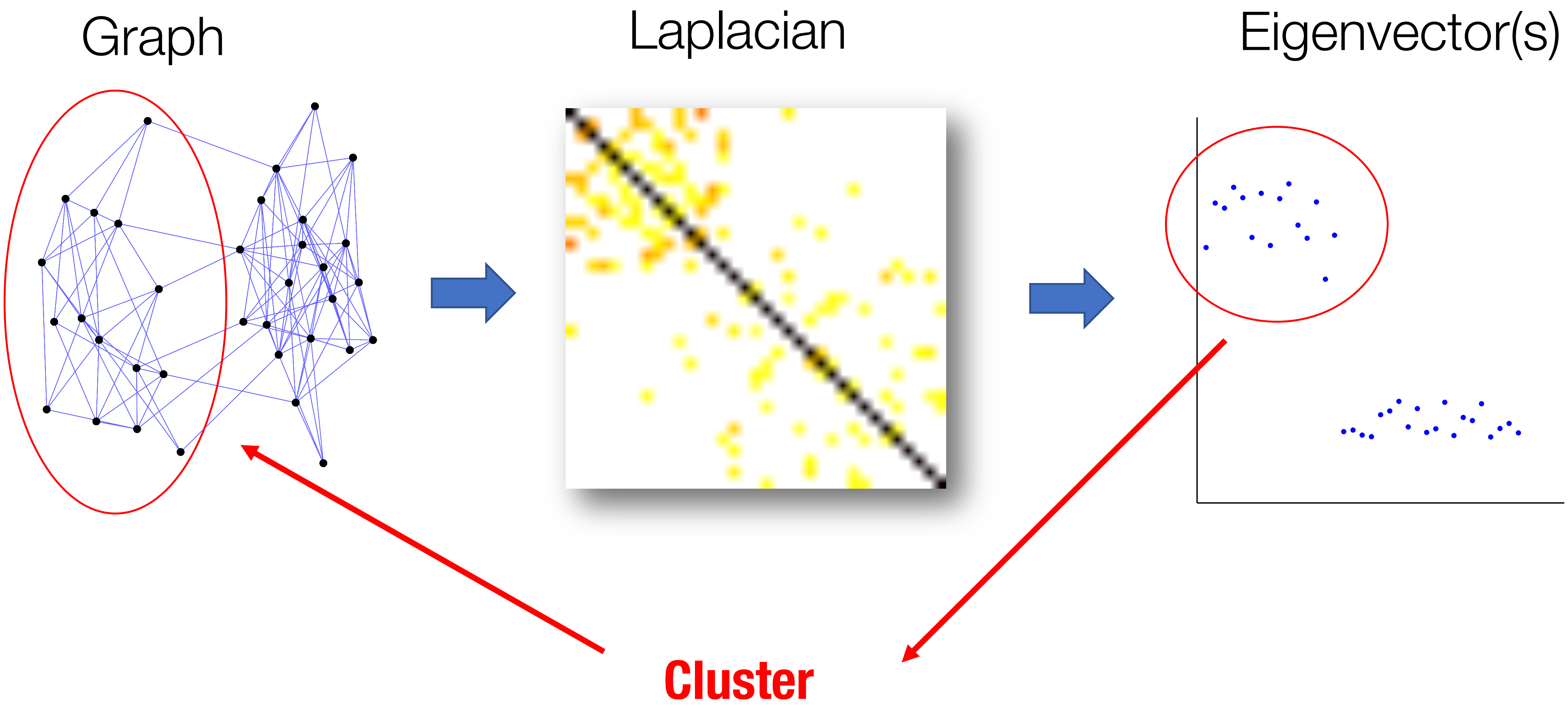
Higher-order graph clustering
Main points and overview

- 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…

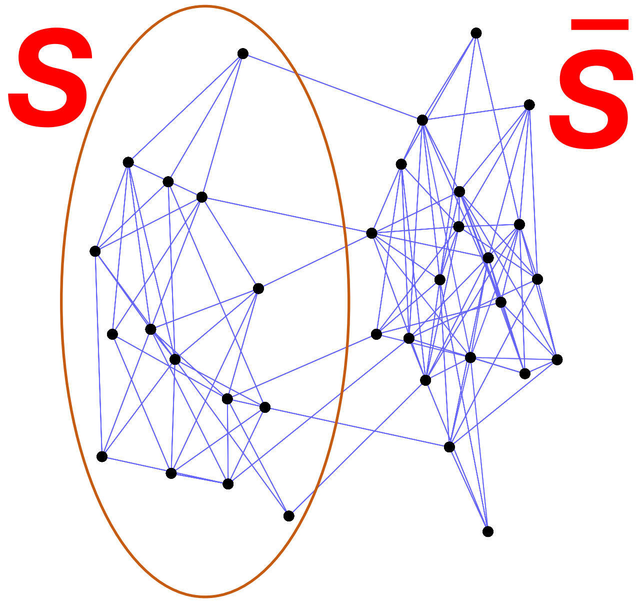
$D = \text{diag}(A1)$ Diagonal degree matrix (1 is the vector of all ones).

$L = D - A$ The graph Laplacian

$\mathcal{L} = D^{-1/2} L D^{-1/2}$ The normalized graph Laplacian

Background. Spectral clustering works based on conductance

Conductance is one of the most important cluster quality scores [Schaeffer07]
used in Markov chain theory, spectral clustering, bioinformatics, vision, etc.



$\mathrm{cut}(S) = 7$

$\mathrm{vol}(S) = 85$

$\mathrm{vol}(\bar{S}) = 151$

$\phi(S) = 7/85$

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

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

(edges leaving S)

(edge end points in S)

small conductance ⇔ good cluster

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.

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

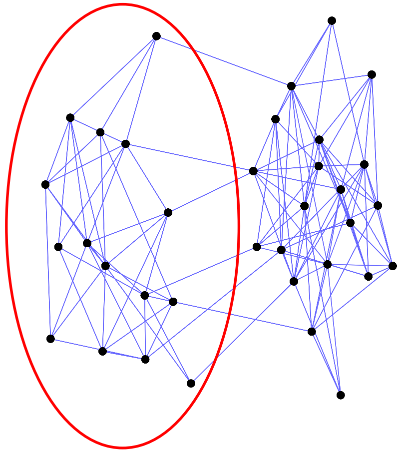
$$L = D - A$$

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)!



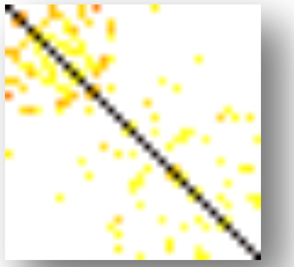
Eigenvalues of the Laplacian $\mathcal{L}$

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

ϕ_* = set of smallest conductance

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

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



Laplacian

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

Cheeger Inequality

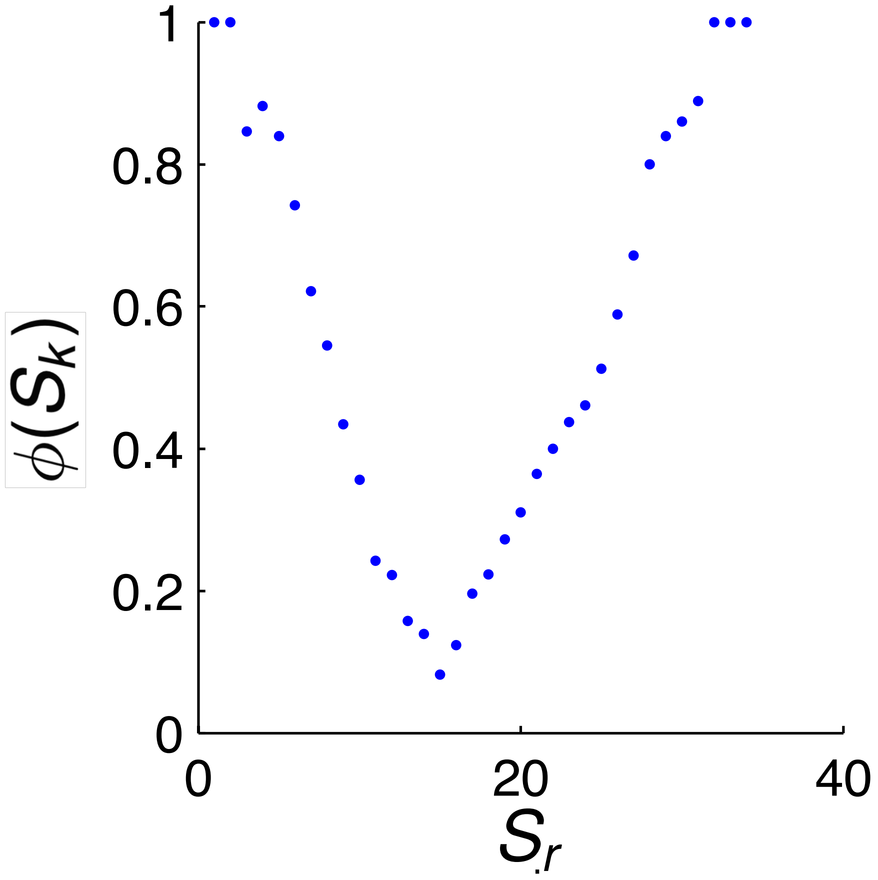
Background. The sweep cut algorithm realizes the guarantee

[Mihail89, Chung92]

We can find a set S that achieves the Cheeger bound.

1. Compute the eigenvector $\boldsymbol{z}$ associated with λ_2 and scale to $\boldsymbol{f} = D^{-1/2}\boldsymbol{z}$
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.

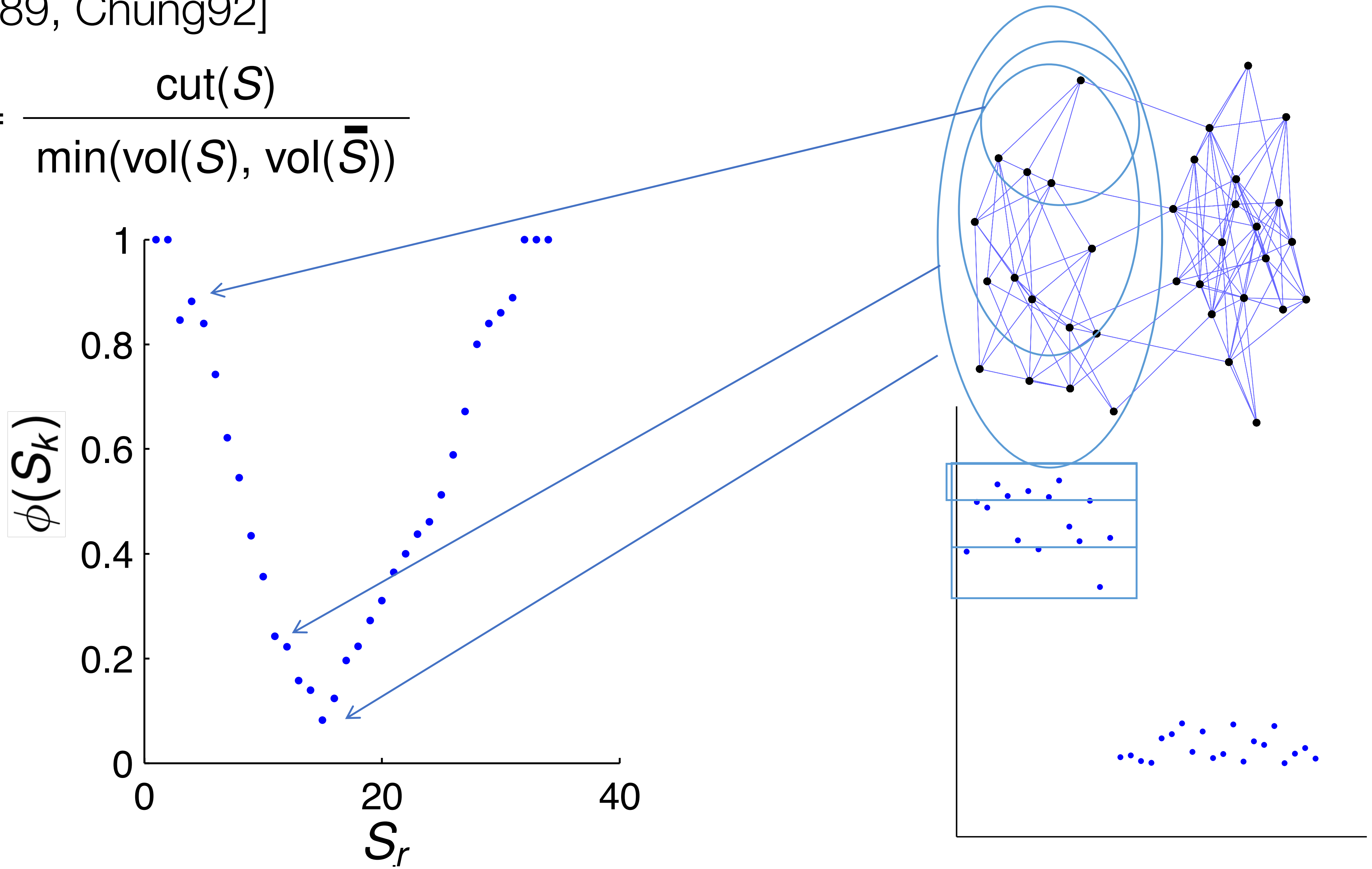
$$\phi(S_m) \leq 2\sqrt{\phi_*}$$



Background. The sweep cut visualized

[Mihail89, Chung92]

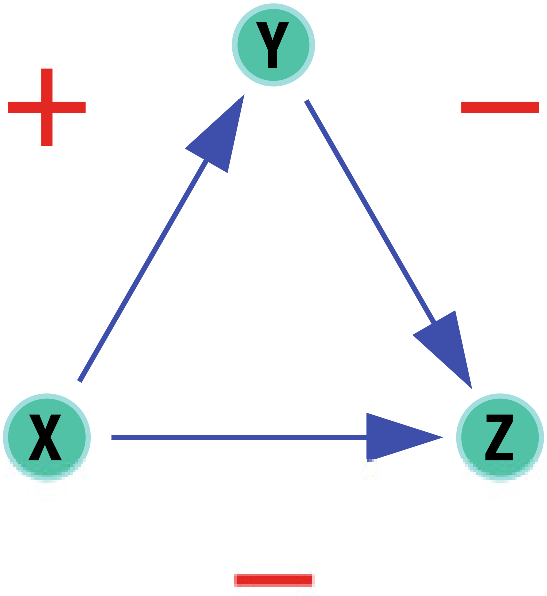
$$\phi(S) = \frac{\text{cut}(S)}{\min(\text{vol}(S), \text{vol}(\bar{S}))}$$



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.

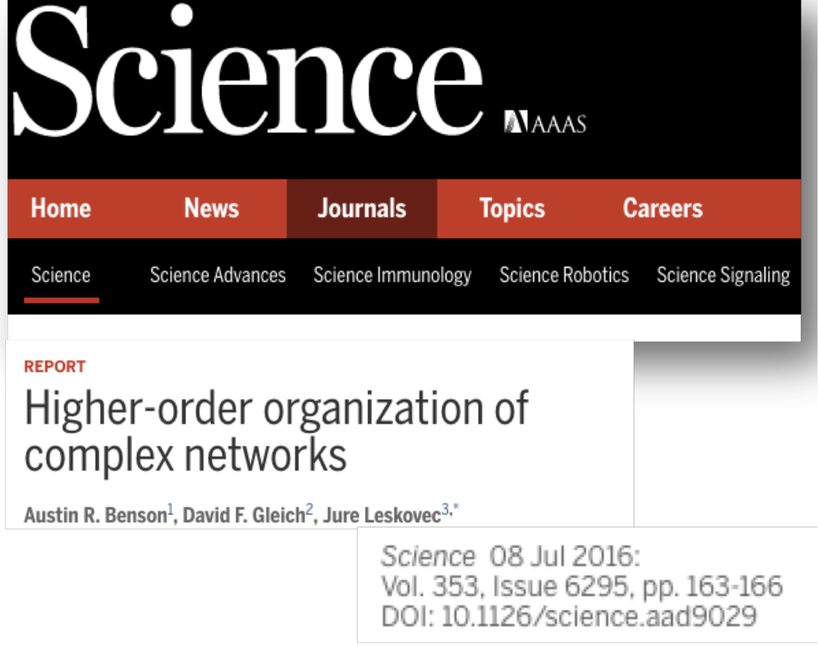


Signed feed-forward loops in genetic transcription [Mangan+03]

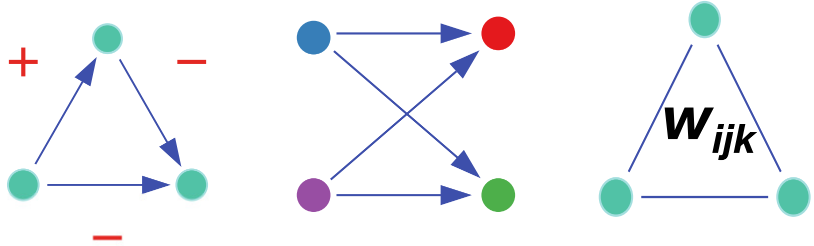
Gene X activates transcription in gene Y.
Gene X suppresses transcription in gene Z.
Gene Y suppresses transcription in gene Z.

Our contributions

Benson-Gleich-Leskovec, *Science*, 2016.

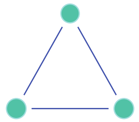

Science AAAS
Home News Journals Topics Careers
Science Science Advances Science Immunology Science Robotics Science Signaling
REPORT
Higher-order organization of complex networks
Austin R. Benson[1], David F. Gleich[2], Jure Leskovec[3,*]
Science 08 Jul 2016:
Vol. 353, Issue 6295, pp. 163-166
DOI: 10.1126/science.aad9029

- 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.



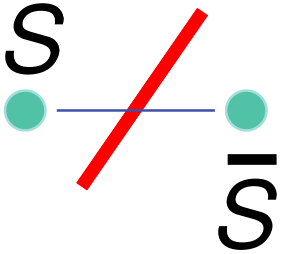
How do we find clusters based on motifs?

Motif-based conductance

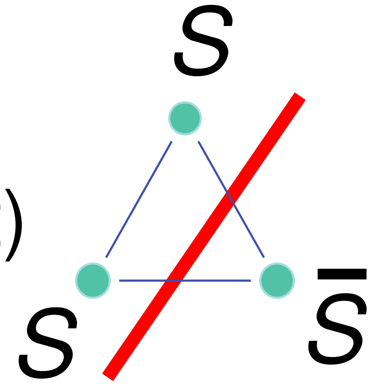
M = triangle motif 

Need new notions of *cut* and *volume*

$\text{cut}(S)$ = #(edges cut)



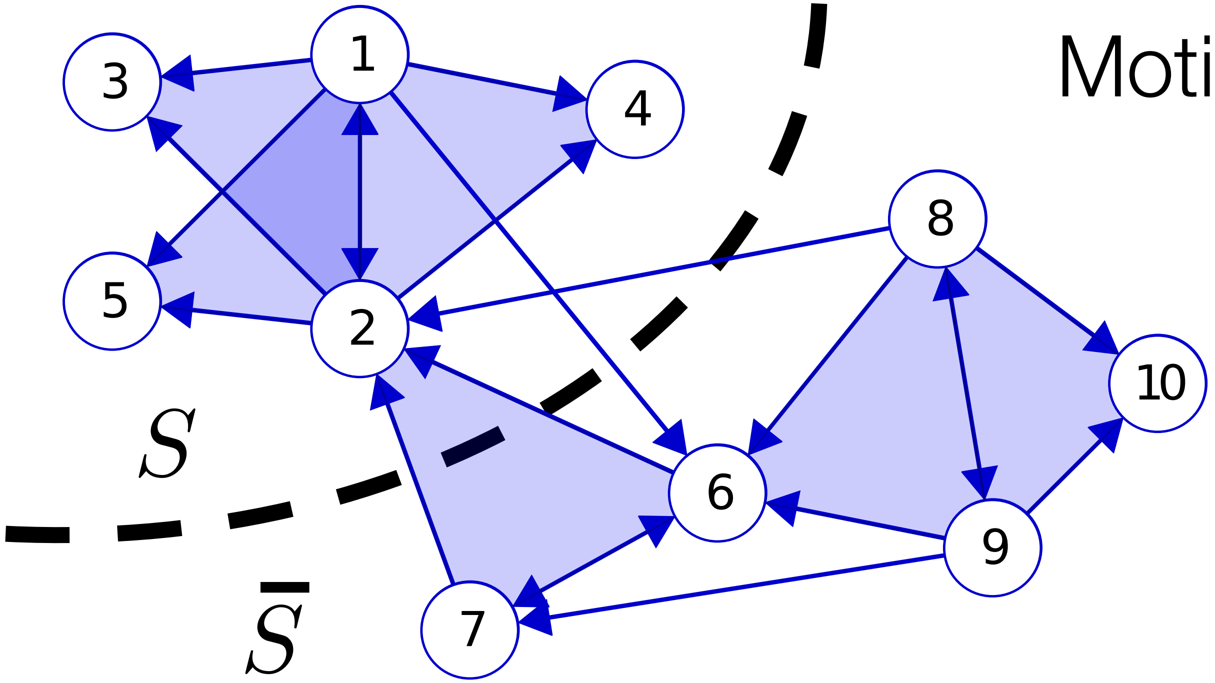
$\Rightarrow$ $\text{cut}_M(S)$ = #(motifs cut)

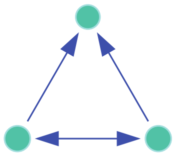


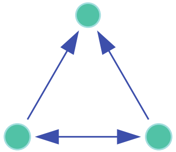
$\text{vol}(S)$ = #(edge end points in S) $\Rightarrow$ $\text{vol}_M(S)$ = #(motif end points in S)

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

Motif-based conductance



Motif M 



$$\phi_M(S) = \frac{\text{motifs cut}}{\text{motif volume}} = \frac{1}{8}$$

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

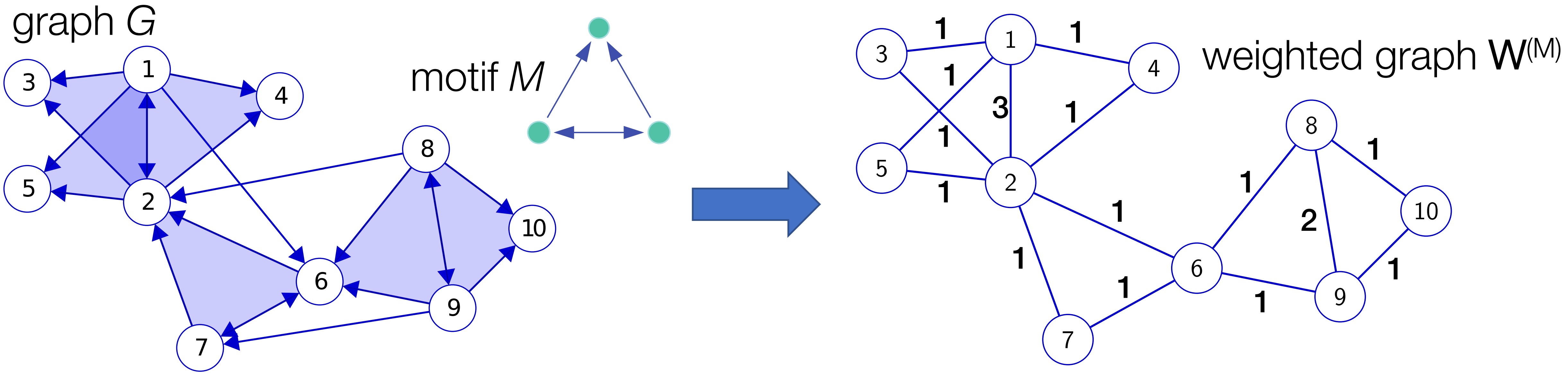
1. Form new weighted, undirected graph $W^{(M)}$ based on *M* and *G*
2. Compute Fiedler vector of Laplacian matrix of $W^{(M)}$ [Fiedler73, Alon-Milman85]
3. Use "sweep cut" procedure to output clusters [Mihail89, Chung92]

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 $W^{(M)}$.



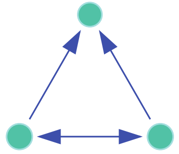
$$W_{ij}^{(M)} = \#\{\text{instances of motif } M \text{ that contain nodes } i \text{ and } j\}$$

Motif-based spectral clustering

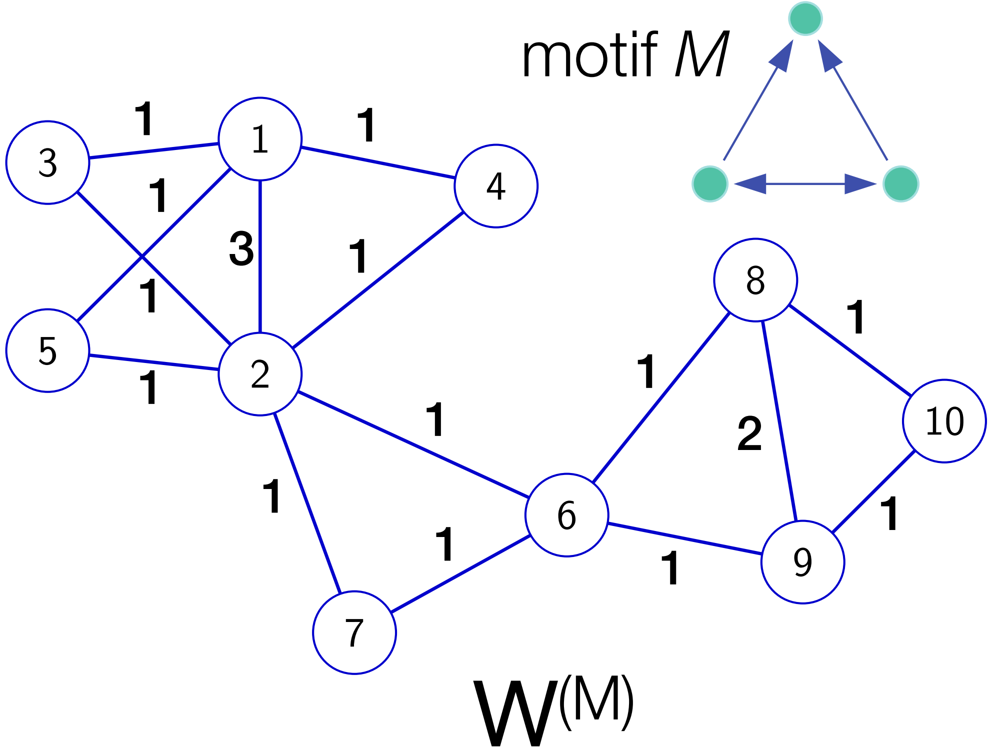
Step 1. Given directed graph G and motif M, form a weighted graph $W^{(M)}$.

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

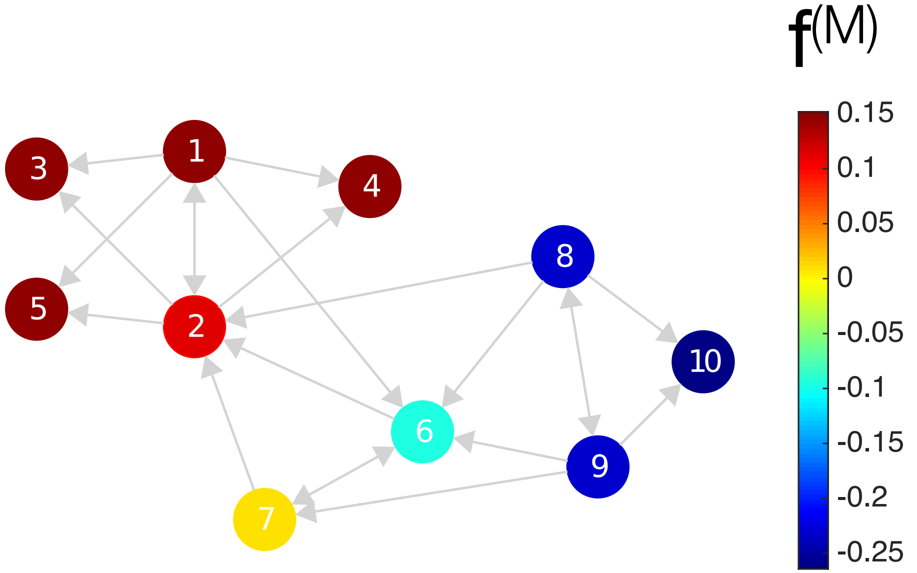
Step 2. Compute the eigenvector $\mathbf{f}^{(M)}$ associated with λ_2 of the normalized Laplacian matrix of $\mathbf{W}^{(M)}$

$$D = \text{diag}(W^{(M)}1)$$

$$\mathcal{L}^{(M)} = D^{-1/2}(D - W^{(M)})D^{-1/2}$$

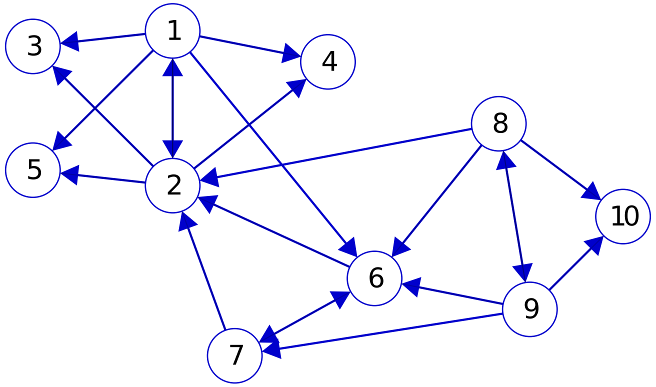
$$\mathcal{L}^{(M)}z = \lambda_2 z$$

$$f^{(M)} = D^{-1/2}z$$



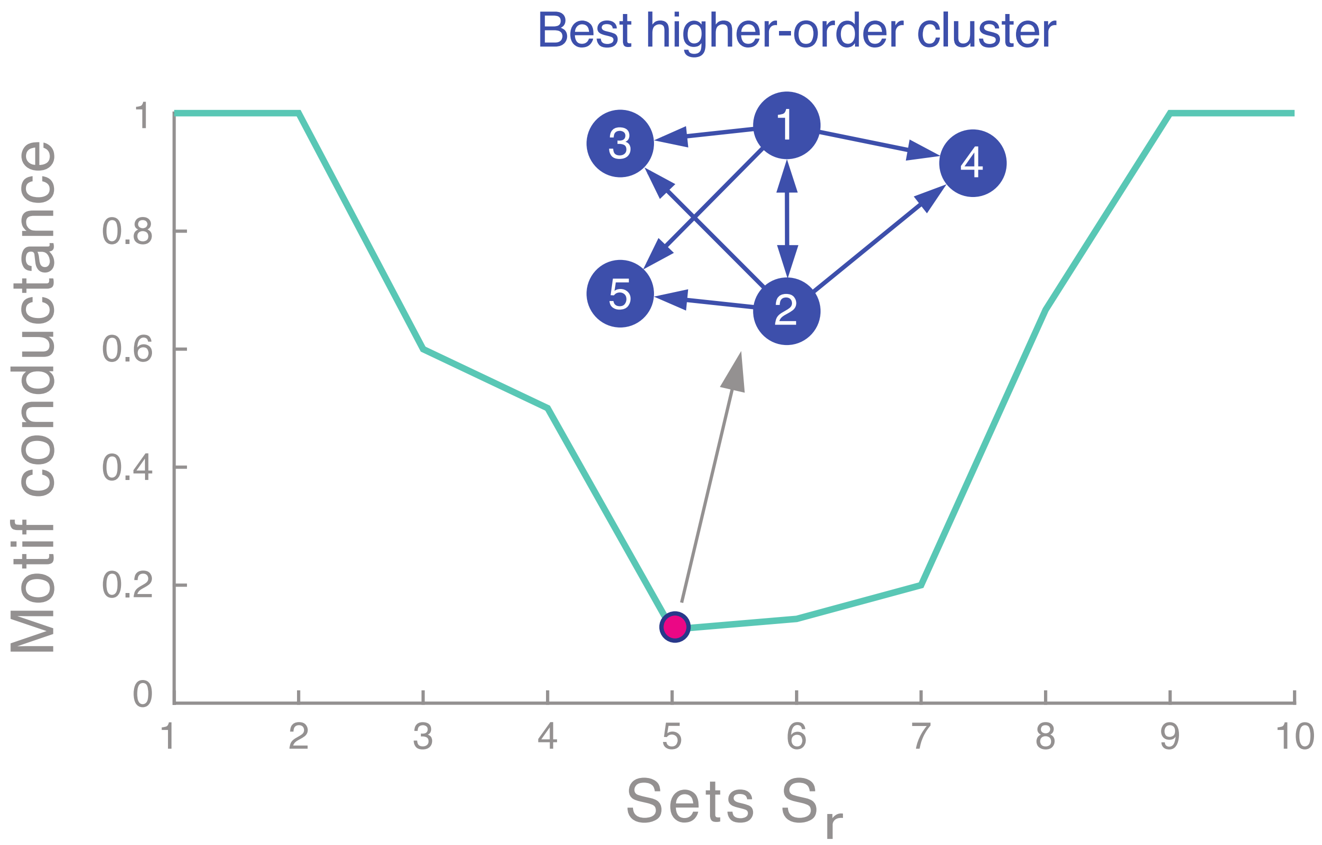
Takes roughly O(# edges) time.

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

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

$$\begin{aligned}
\text{cut}_M(S, G) &= \sum_{\{i,j,k\} \in M(G)} \text{Indicator}[x_i, x_j, x_k \text{ not the same}] \\
&= \frac{1}{4}(x_i^2 + x_j^2 + x_k^2 - x_i x_j - x_j x_k - x_i x_k) \\
&= \text{quadratic in } \mathbf{x}
\end{aligned}$$

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

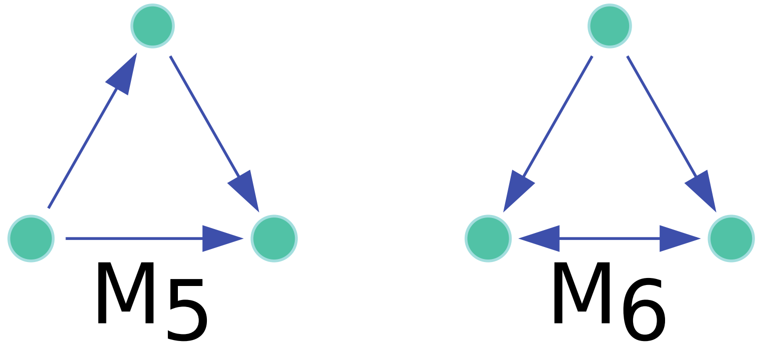
Application 1

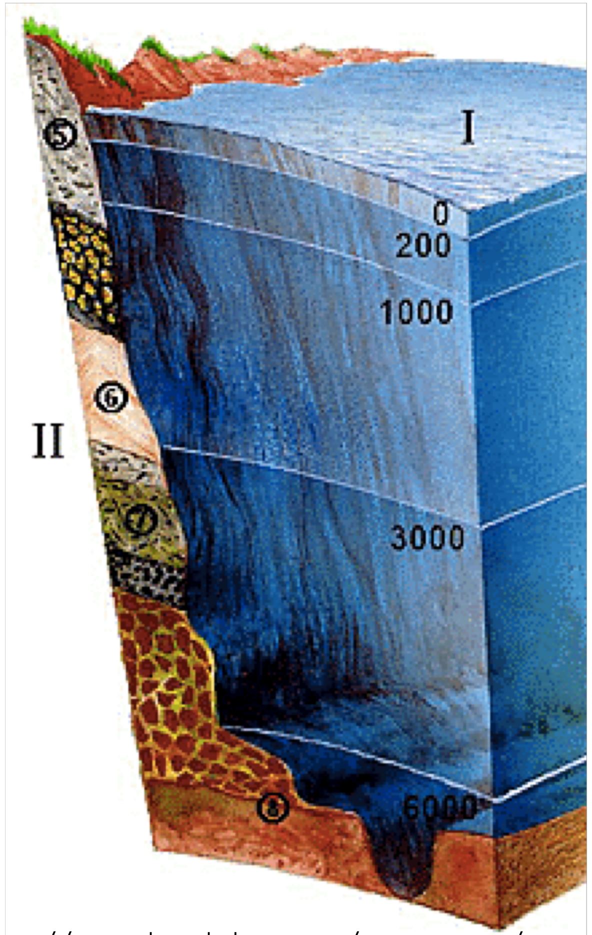
We do not know the motif of interest.

Application 1. Food webs

Florida bay food web

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





http://marinebio.org/oceans/marine-zones/

Application 1. Food webs

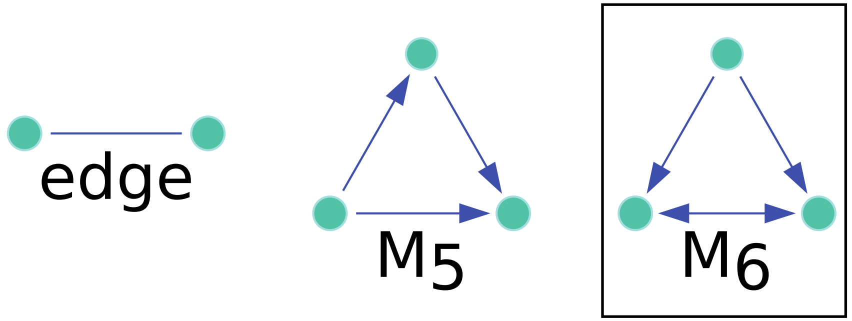
Which motif clusters the food web?

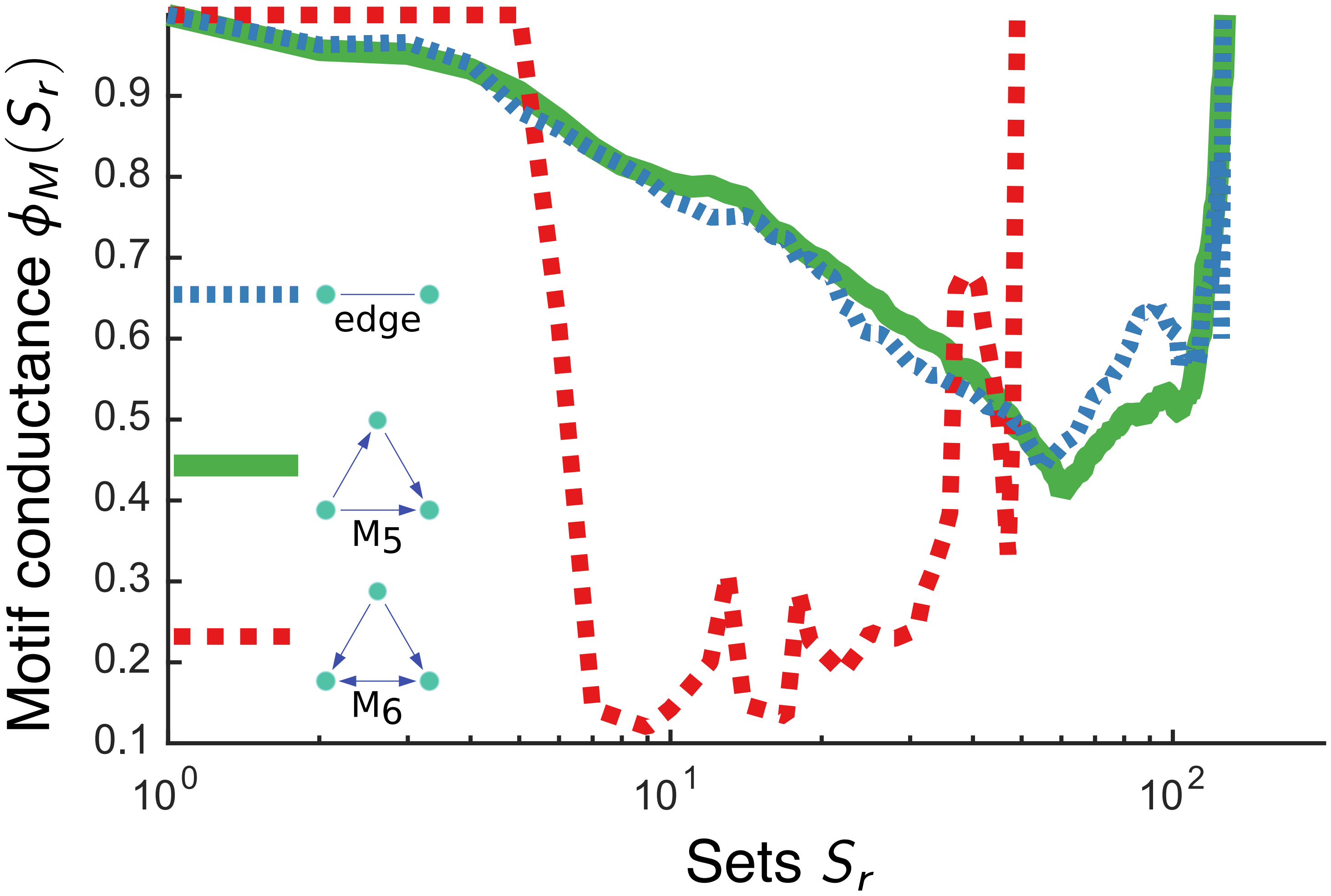
Our approach

- 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.

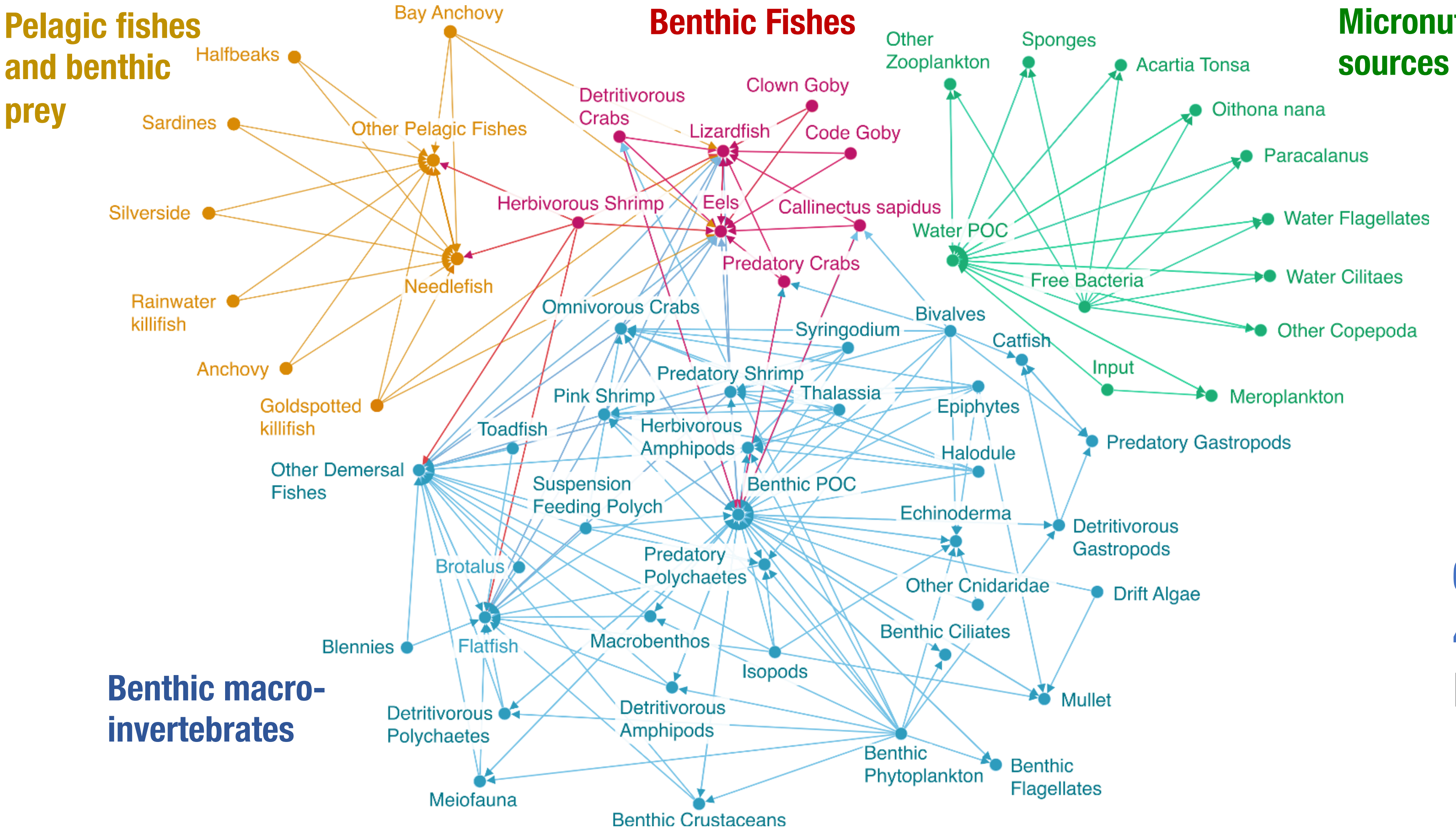
Application 1. Food webs

Our finding. Motif M_6 organizes the food web into good clusters.

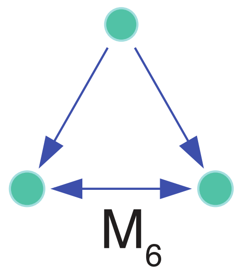




Application 1. Food webs



Motif M_6 reveals aquatic layers



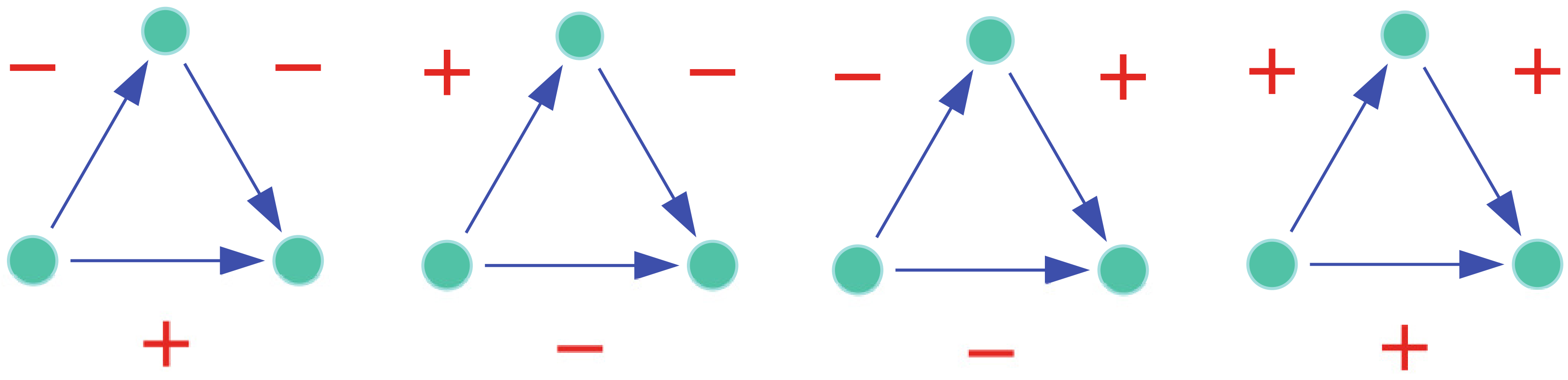
61% accuracy vs. **48%** with edge-based methods

Application 2

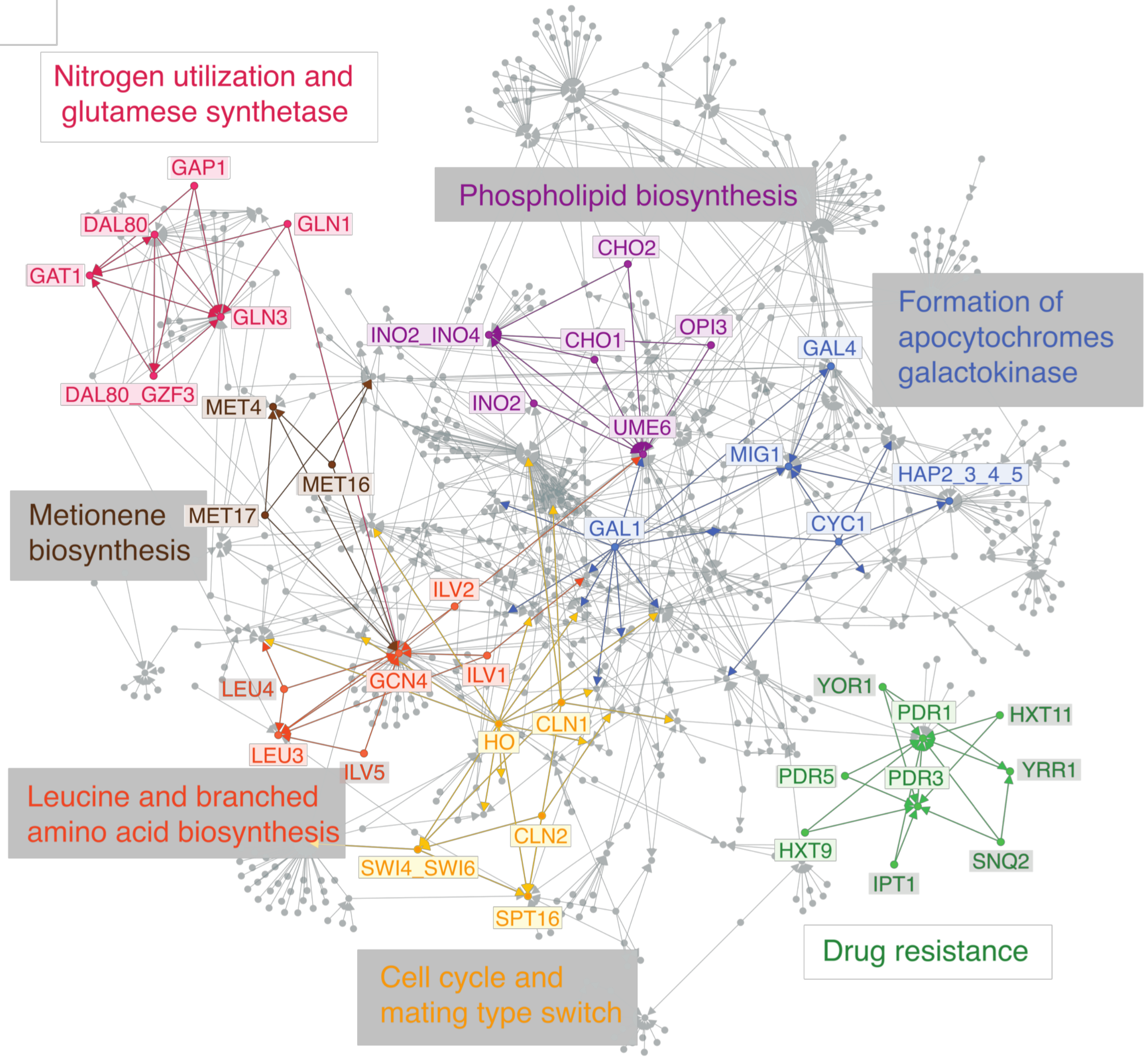
We know the motif of interest from domain knowledge.

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]



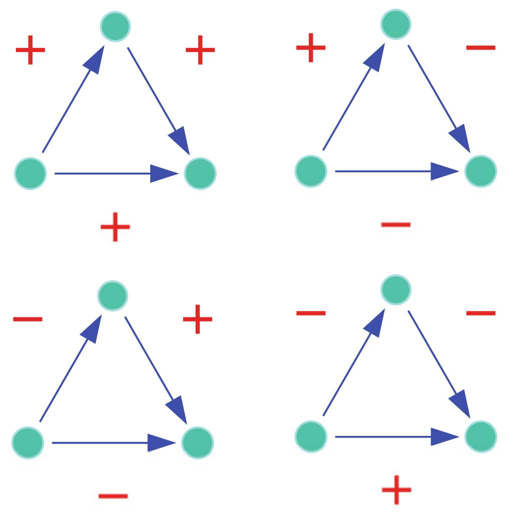
Application 2. Yeast transcription regulation networks



Clustering based on coherent feedforward loops identifies functions studied individually by biologists [Mangan+03]

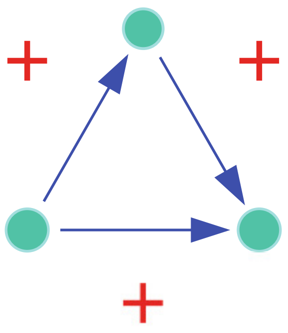
97% accuracy vs.

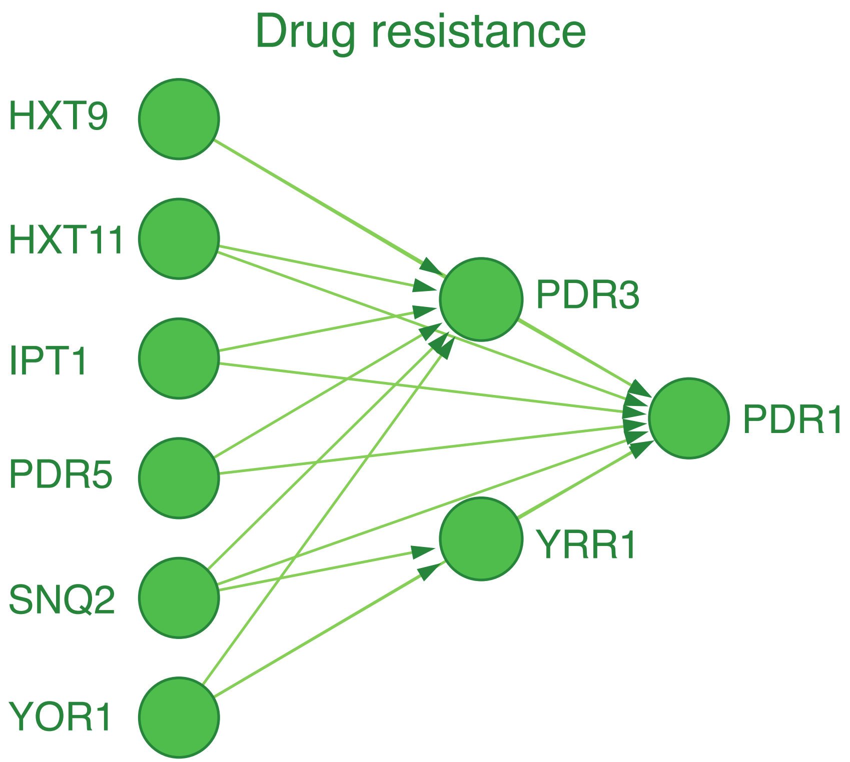
68–82% with edge-based methods

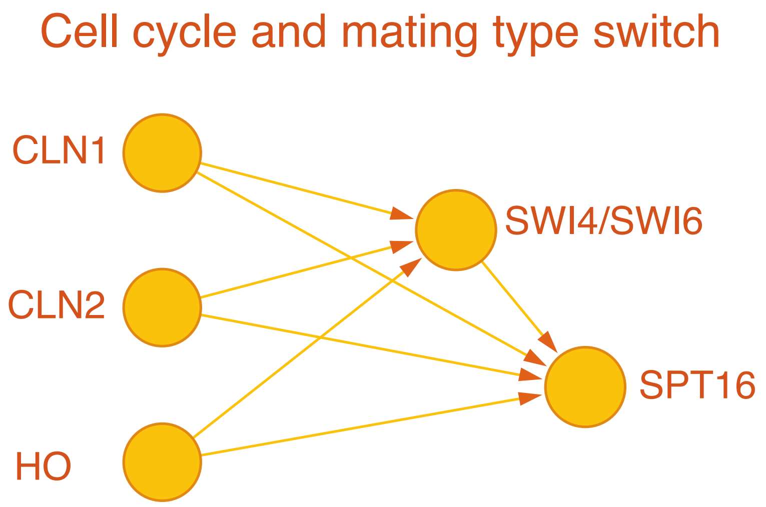


Application 2. Yeast transcription regulation networks

Structure of the found modules
(all edge signs are positive)



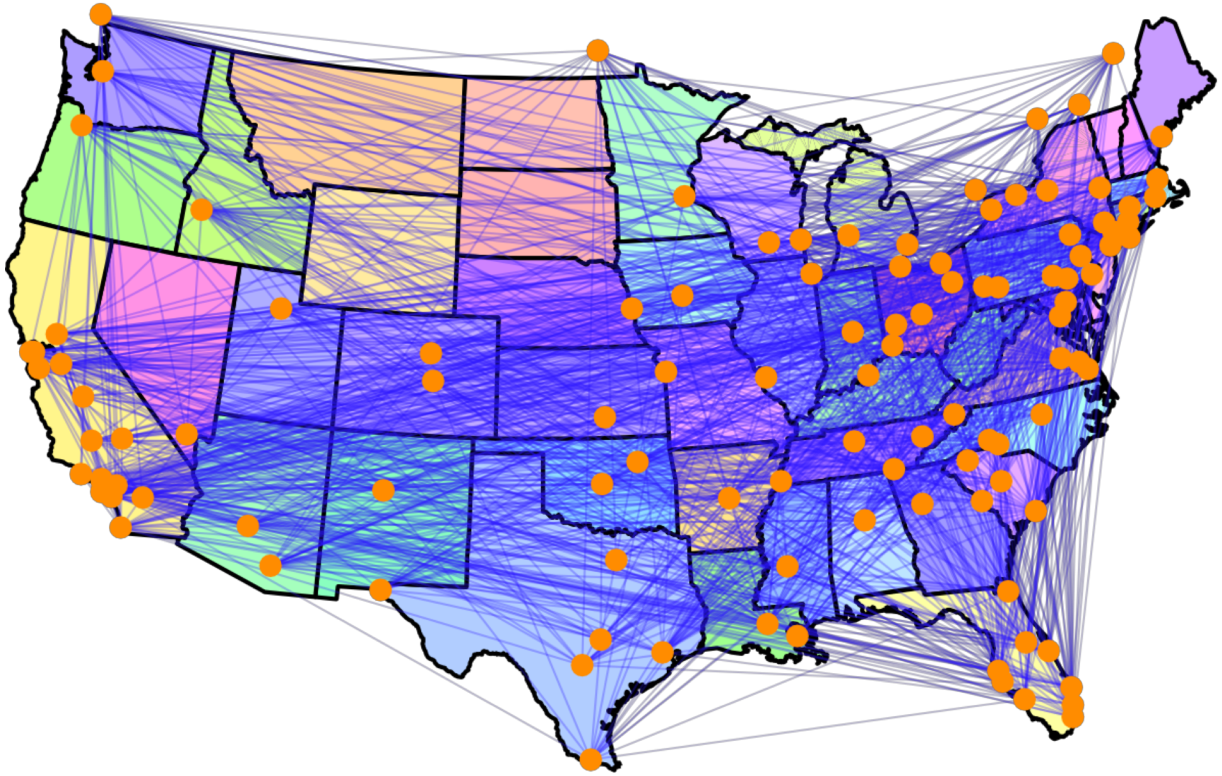




Application 3

We seek richer information from our data.

Application 3. Transportation networks

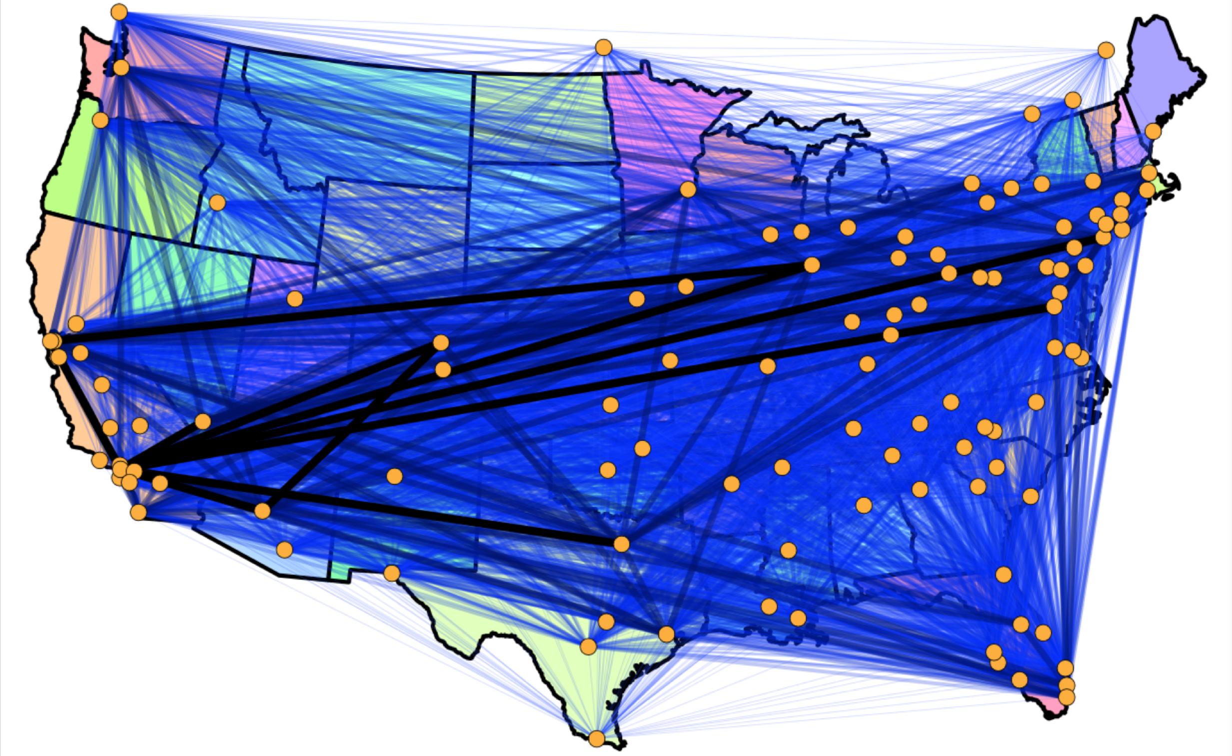


- North American air transport network.
- Nodes are cites.
- $i \to j$ if you can travel from i to j in < 8 hours.

[Frey-Dueck07]

Application 3. **Transportation networks**

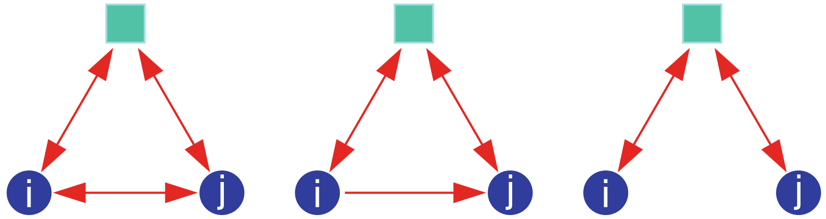
Weighted adjacency matrix already reveals hub-like structure



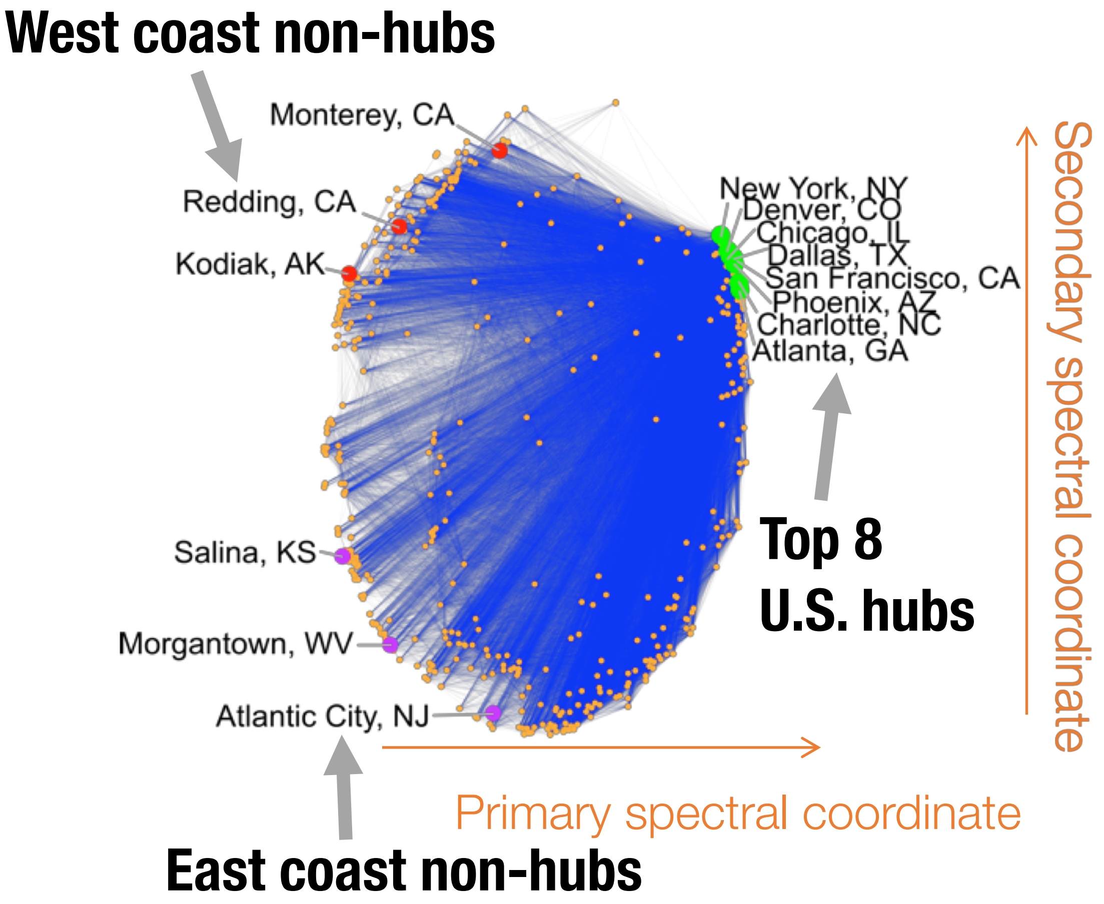
$W_{ij}^{(M)}$ = #{bi-directional length-2 paths from i to j}

Important motifs from literature

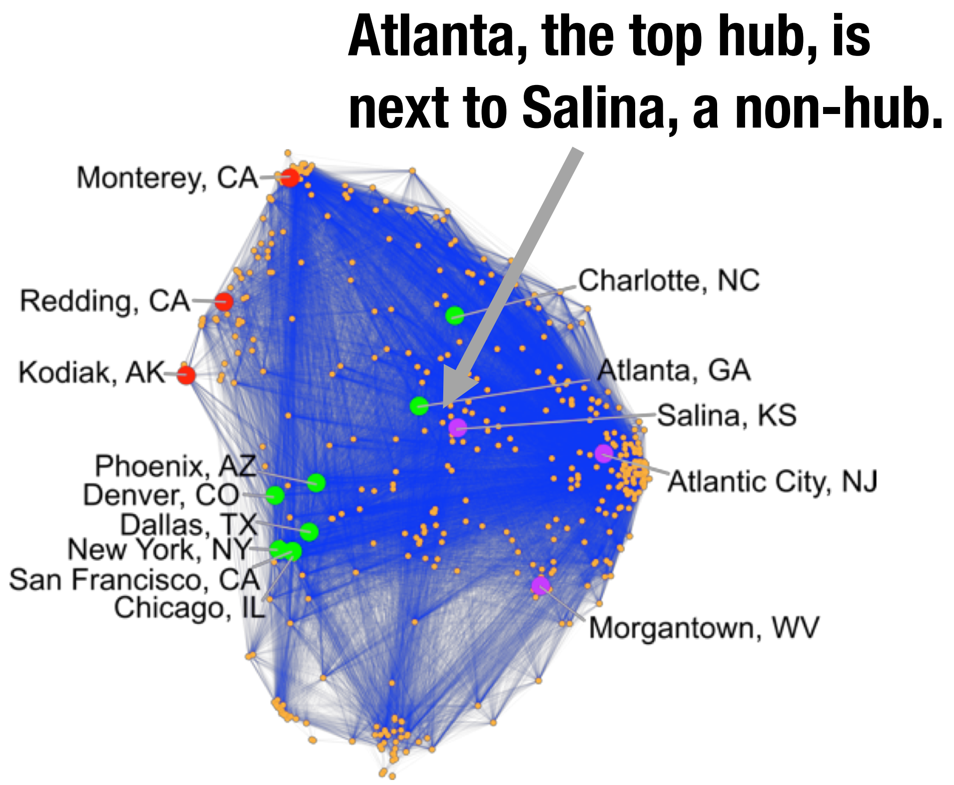
[Rosvall+14]



Application 3. Transportation networks

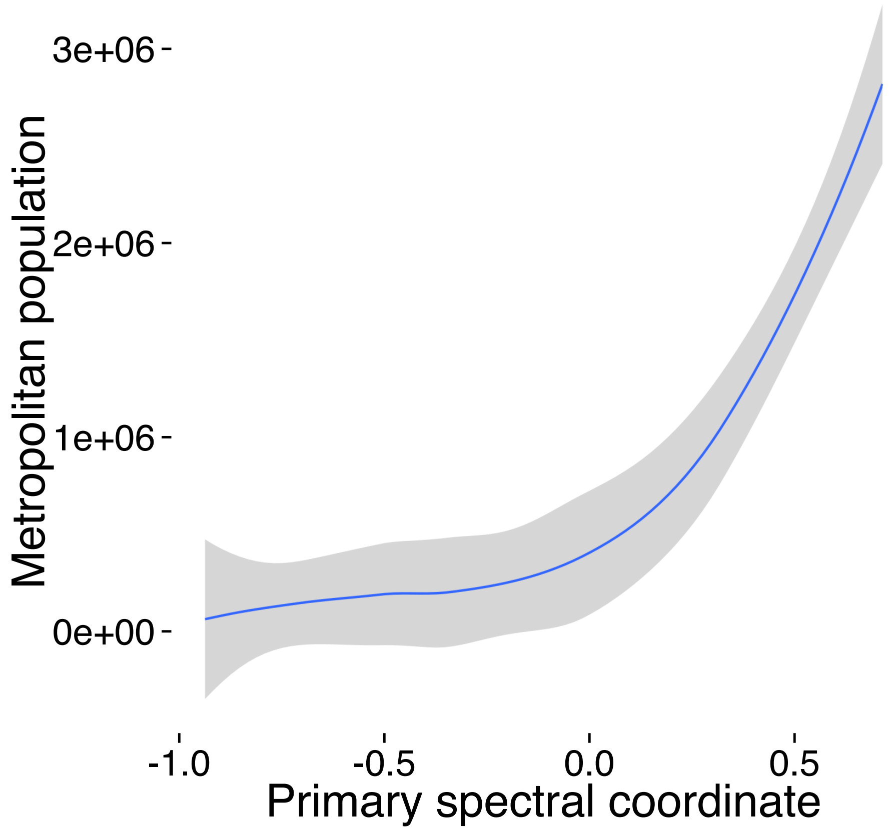


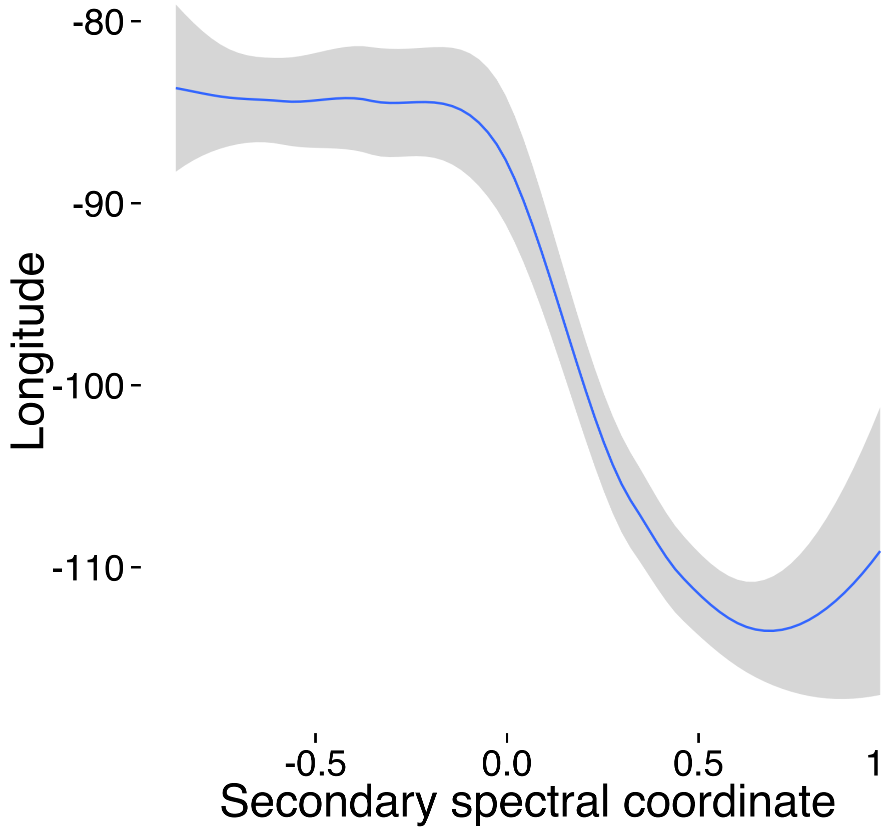
MOTIF SPECTRAL EMBEDDING



EDGE SPECTRAL EMBEDDING

Application 3. Transportation networks

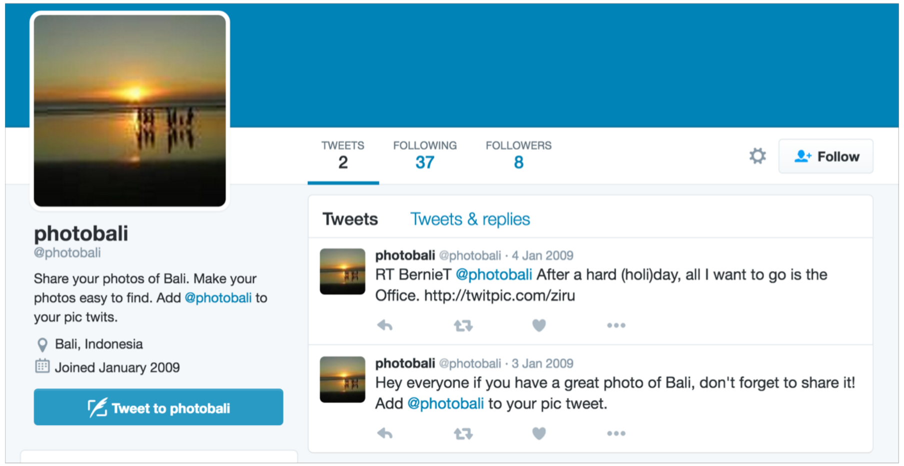


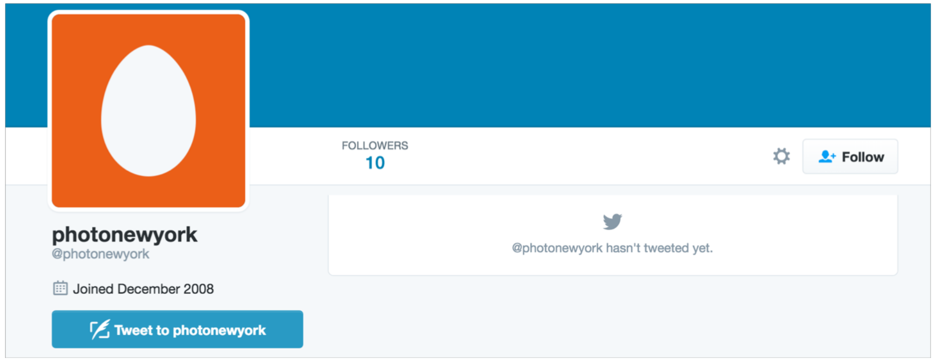


Applications 4, 5 & 6

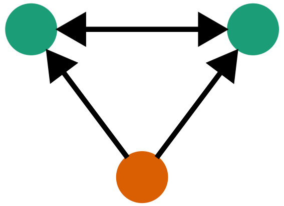
Just some extra fun things we found.

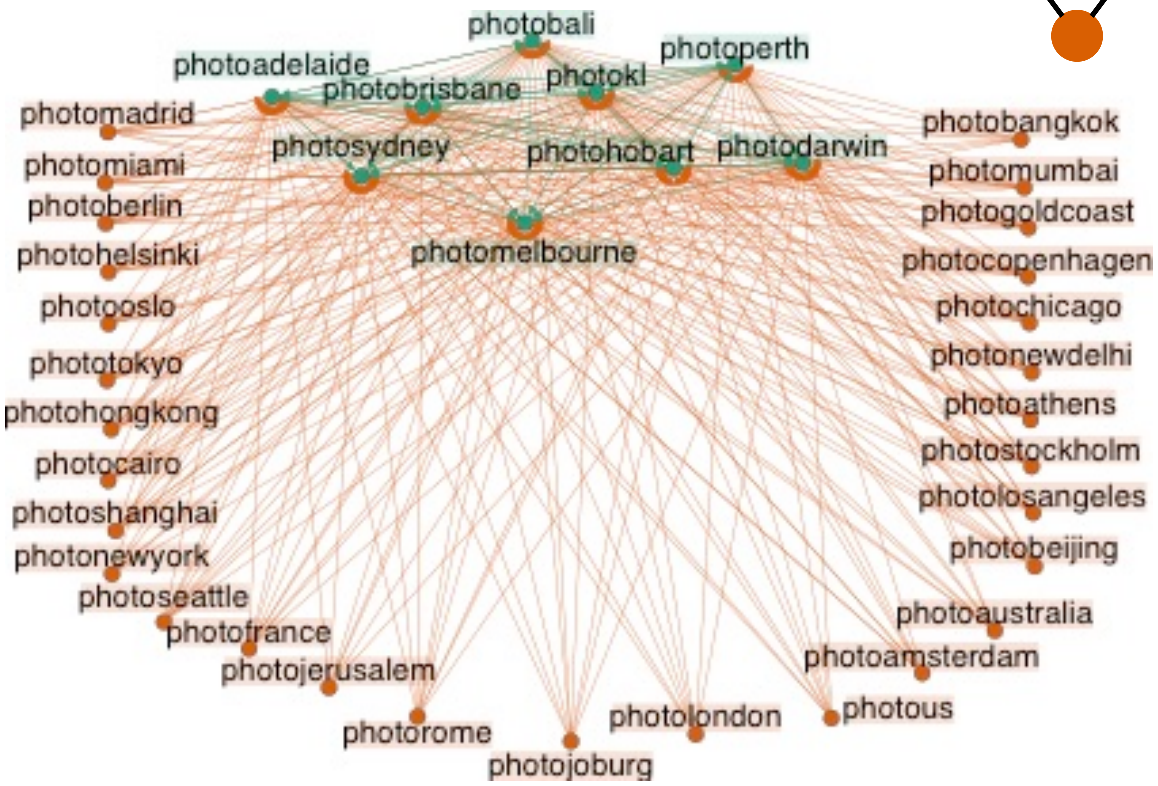
Application 4. Anomaly detection in social networks





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

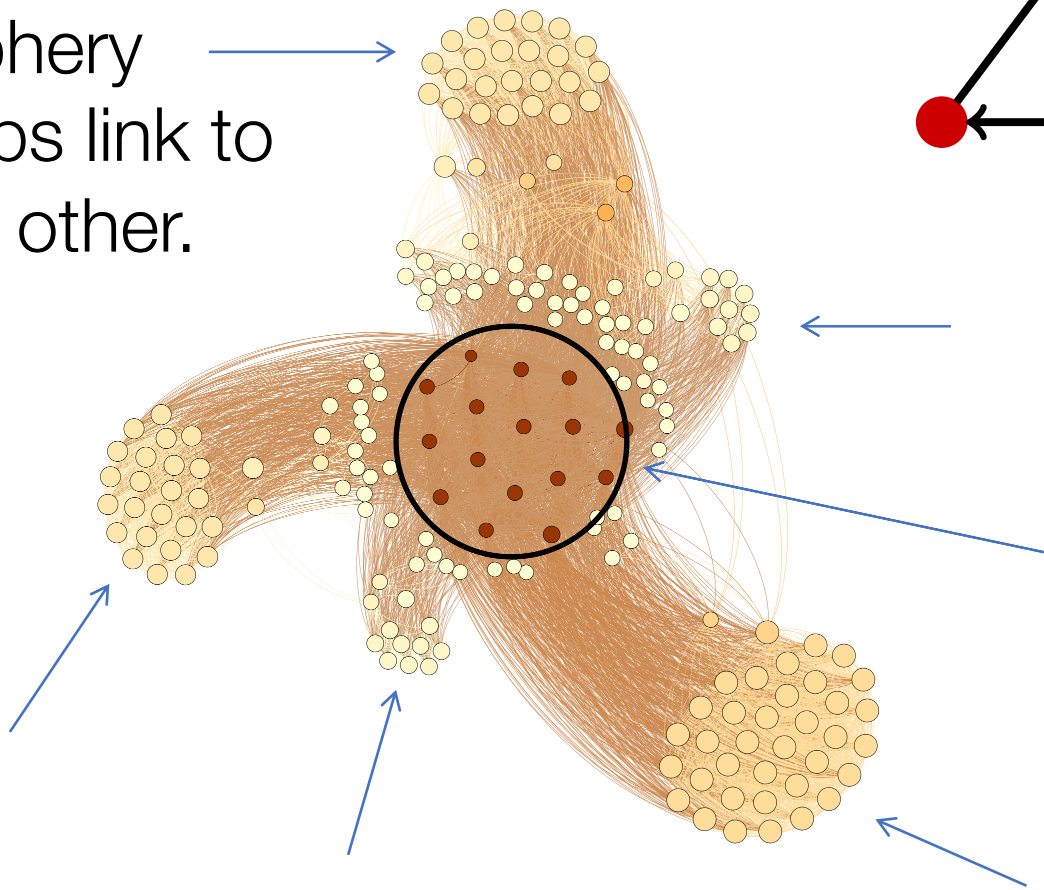


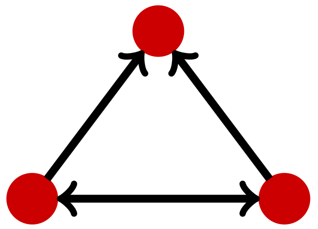


Anomalous cluster in the 1.4B edge Twitter graph.
All nodes are holding accounts for a company, and the orange nodes have incomplete profiles.

Application 5. Hierarchical structure in web graphs

Periphery groups link to each other.



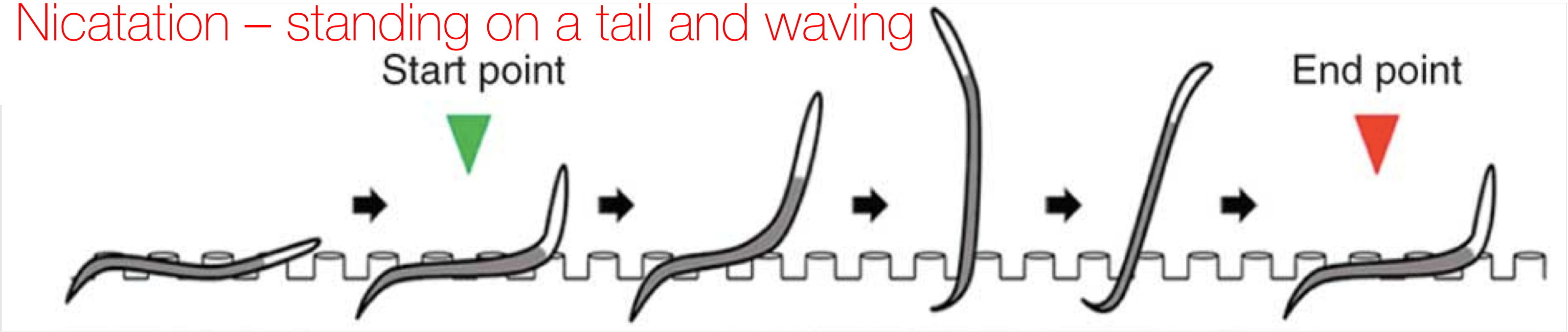


The "uplinked triangle" has been observed to occur much more frequently than in random graph models. [Milo+02]

Core group with large in-degree.

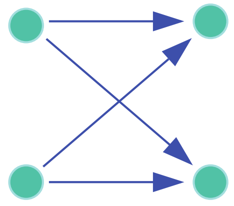
Application 6. Nictation control in a neural network

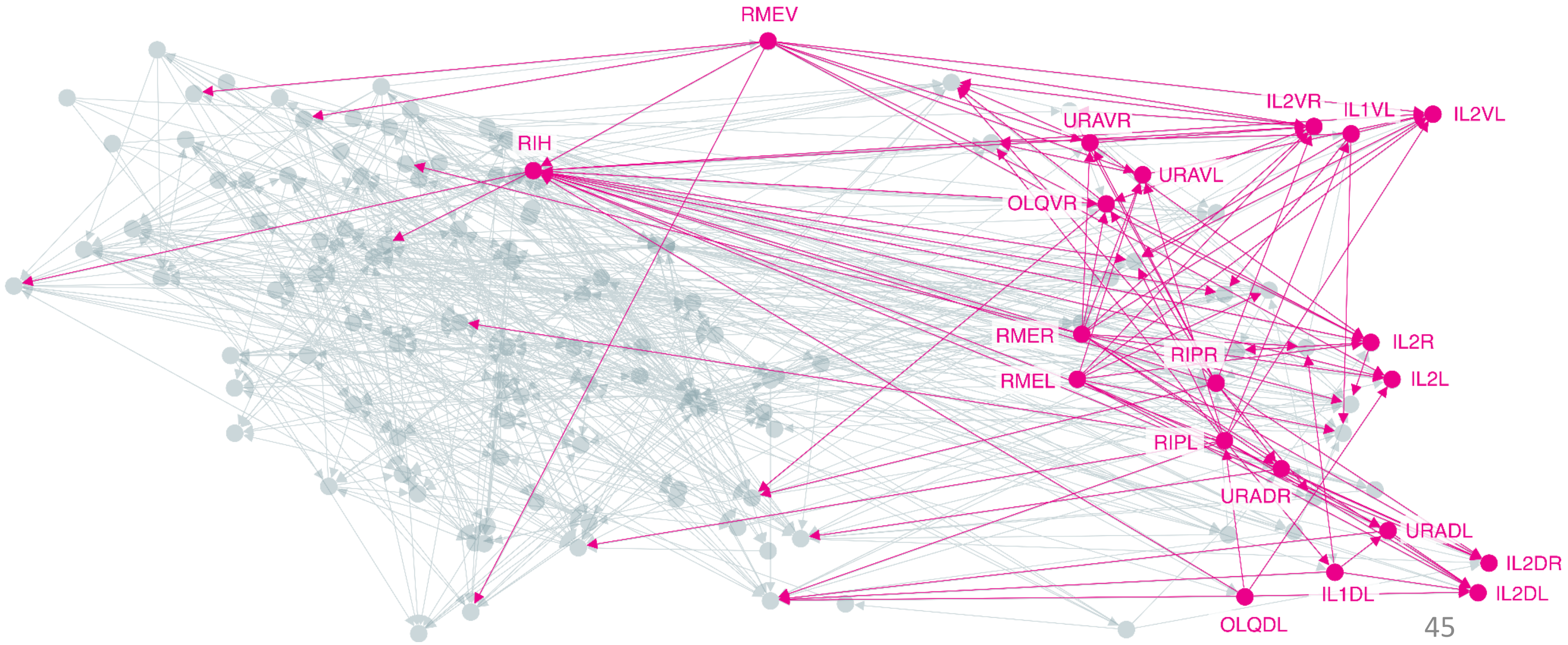
Nicatation – standing on a tail and waving



Nictation, a dispersal behavior of the nematode Caenorhabditis elegans, is regulated by IL2 neurons, Lee et al. Nature Neuroscience, 2011.

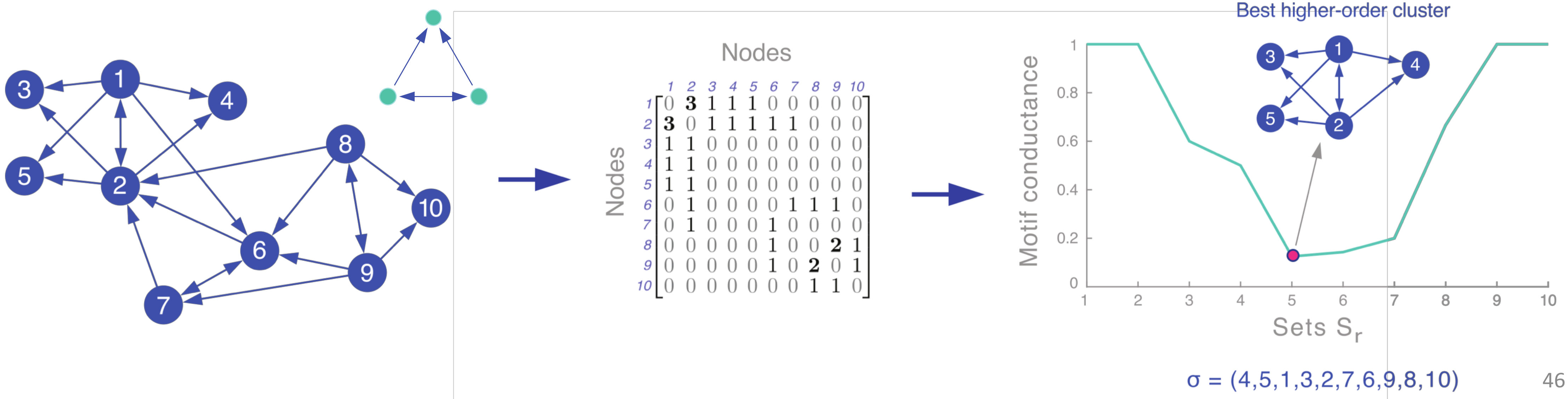
We find the control mechanism that explains nictation based on the bi-fan motif (Milo et al. found it over-expressed)





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.



Higher-order clustering

Benson, Gleich, & Leskovec, Higher-order organization of complex networks, *Science*, 2016

Code + data http://snap.stanford.edu/higher-order

Key takeaways

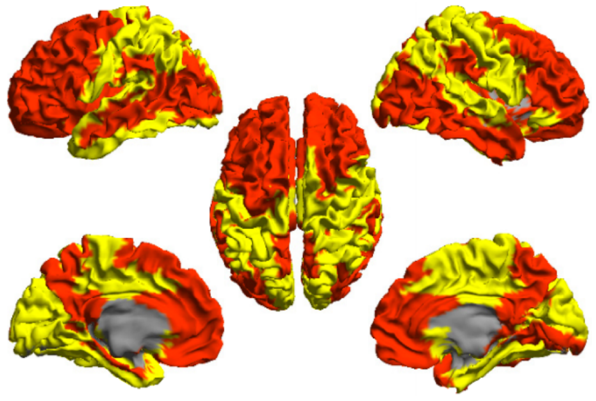
- Organizing graphs according to motifs reveals new insights into data

- 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

Phase Transfer Entropy directed brain networks

Fig. 4 The two clusters (in red and yellow) on the template brain obtained via the motif-based clustering algorithm after the $\pm\sigma$ sparsification based on the motif 78.

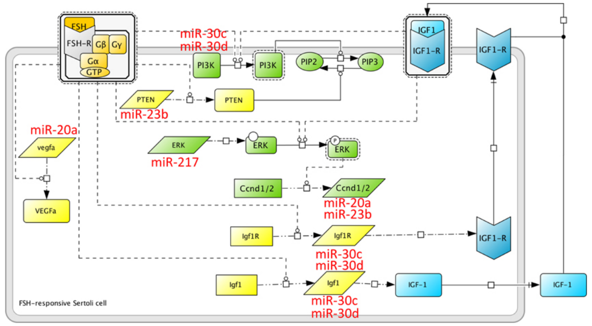


Intermission…

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]

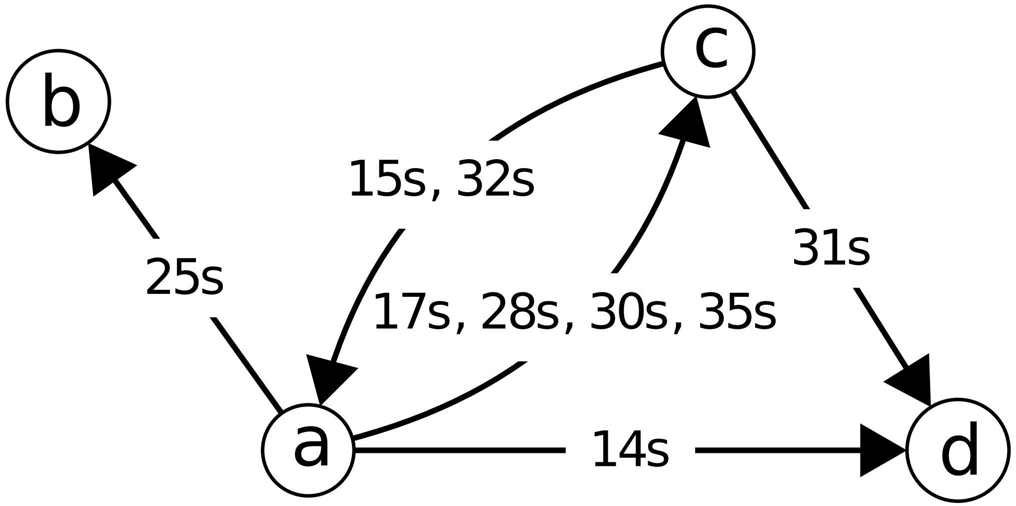
2. **Sequence of snapshot aggregates**
   Daily phone call graph [Araujo+14], Per-year co-authorship [Dunlavy+2010]

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

source	destination	timestamp
a	d	14s
c	a	15s
a	c	17s
a	b	25s
a	c	28s
a	c	30s
c	d	31s
c	a	32s
a	c	35s

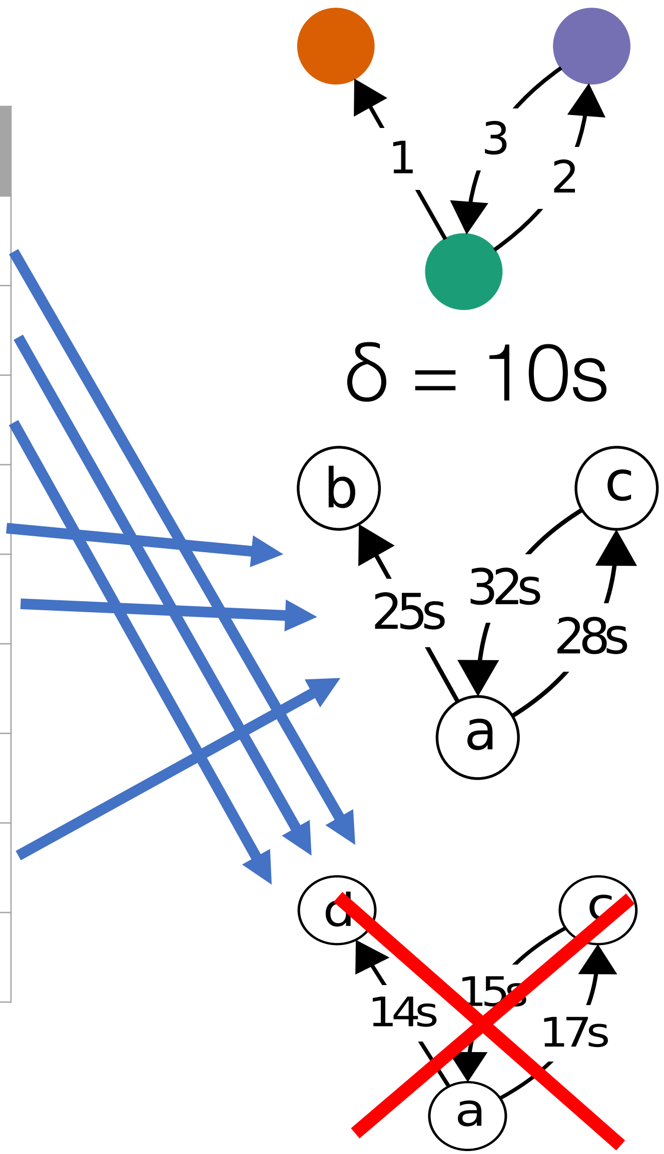


many timestamps between the same pair of nodes!

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

Temporal network motifs

source	destination	timestamp
a	d	14s
c	a	15s
a	c	17s
a	b	25s
a	c	28s
a	c	30s
c	d	31s
c	a	32s
a	c	35s



Temporal network motif

1. Directed multigraph with k edges
2. Edge ordering
3. Maximum time span δ

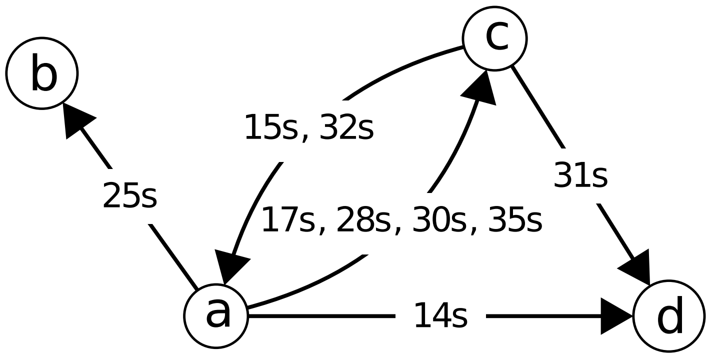
Motif instance k temporal edges that match the pattern that all occur within δ time

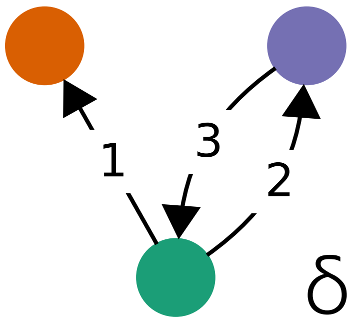
Wrong order!
(c, a) before (a, c)

Paranjape, Benson, & Leskovec, *WSDM*, 2017

Algorithmic challenge of temporal motifs

Given a temporal network and a temporal network motif,



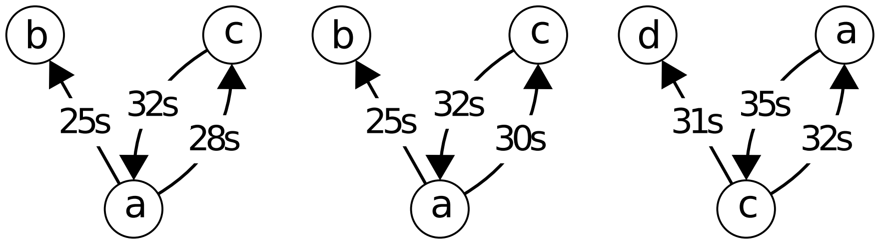


$\delta = 10s$

count the number of motif instances in the network.

3

the 3 instances

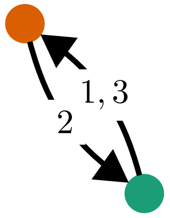


Summary of new algorithms

In a network with *m* temporal edges and *T* static triangles and a motif with *k* temporal edges.

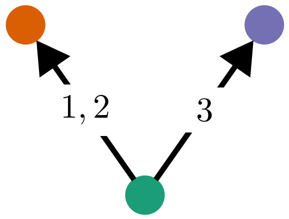
1. General algorithm for any motif.
   faster than $O(m^k)$ brute force approach

   2-nodes, k temporal edges. $O(k^2 m)$,
   linear time in size of data for const. k

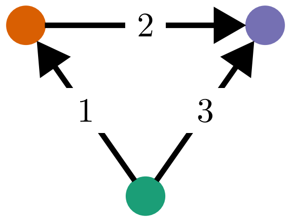


Optimized algorithms for special cases

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

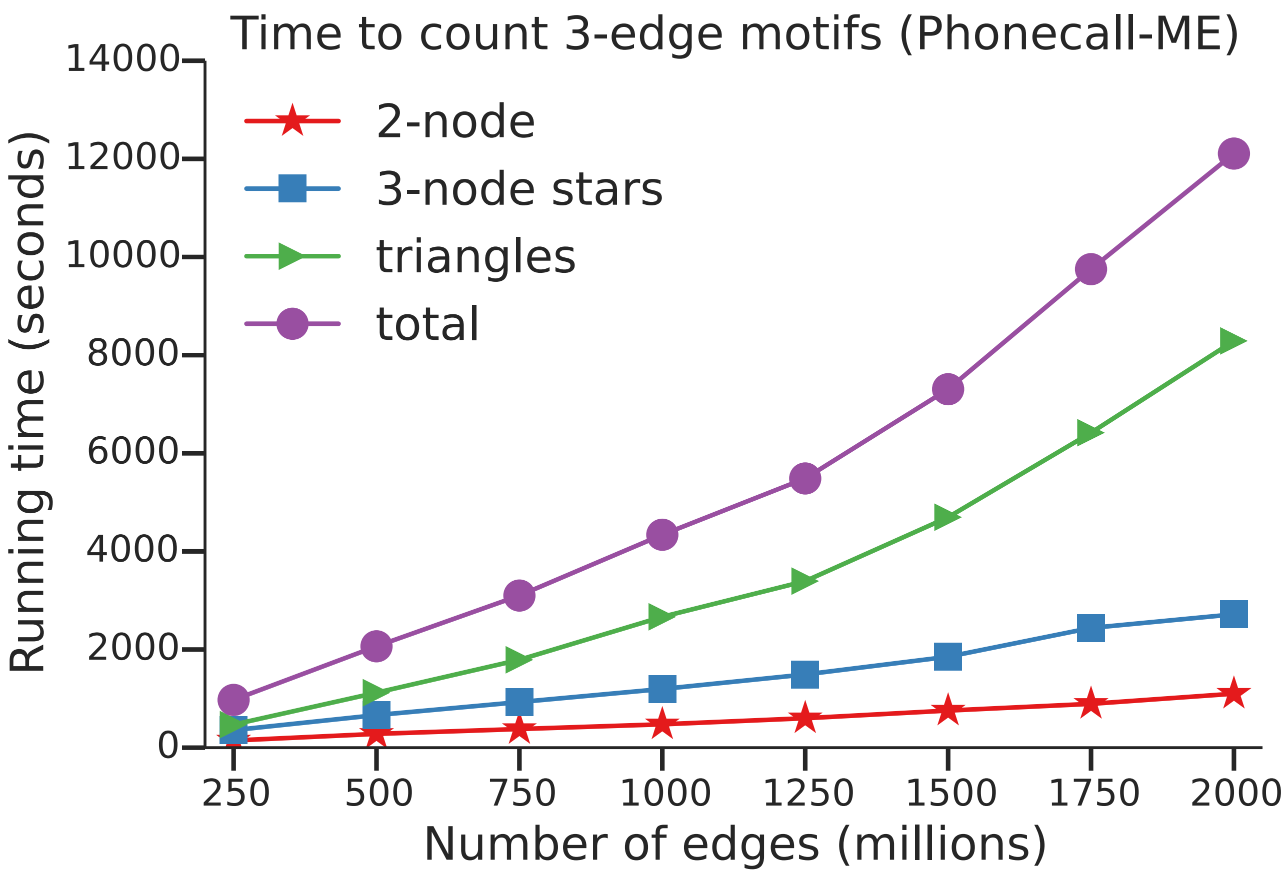


3. 3 nodes, 3 temporal edges, triangles. $O(T^{1/2}m)$
   faster than previous state-of-the-art $O(Tm)$

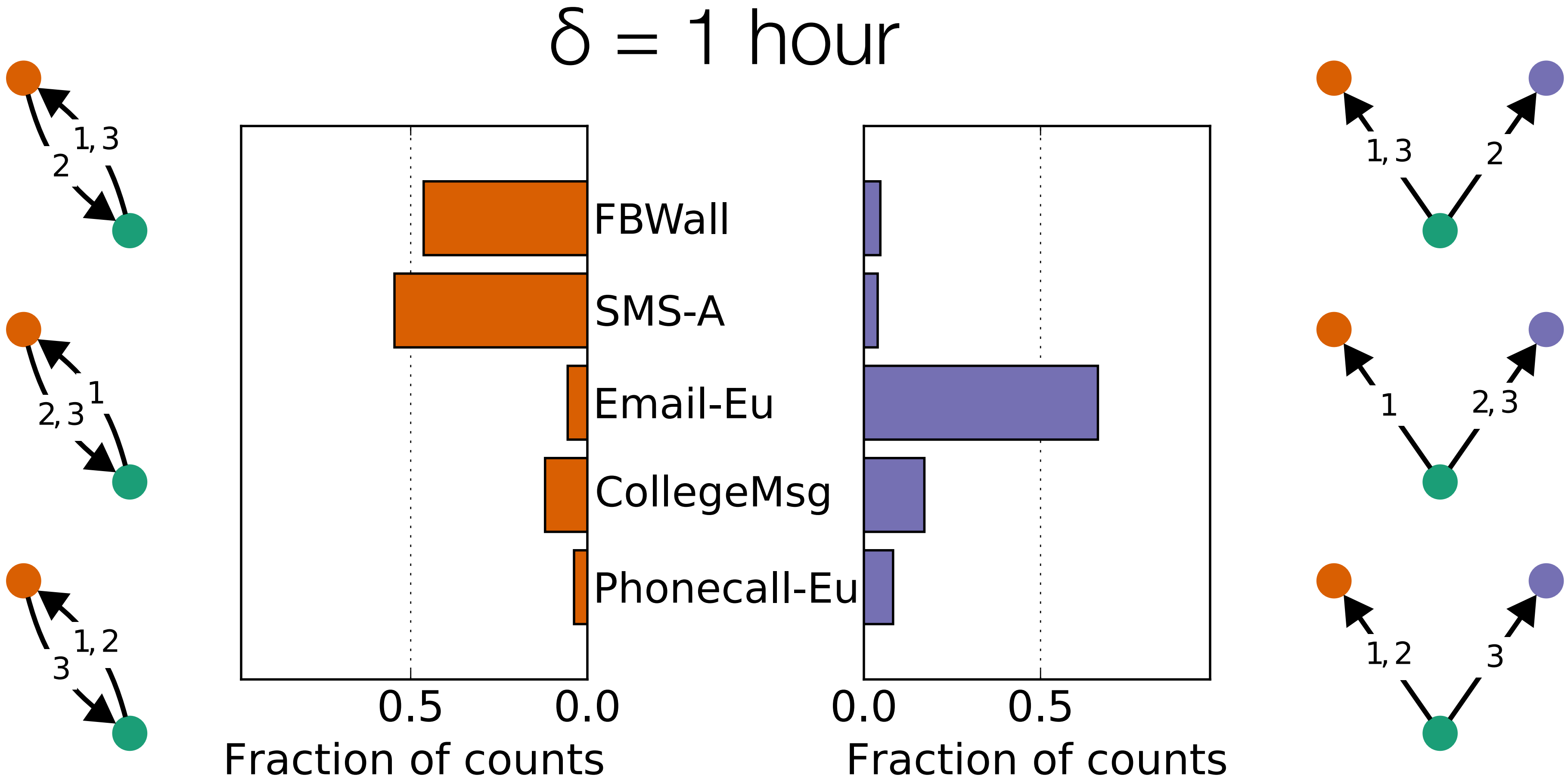


New algorithms let us analyze large datasets

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



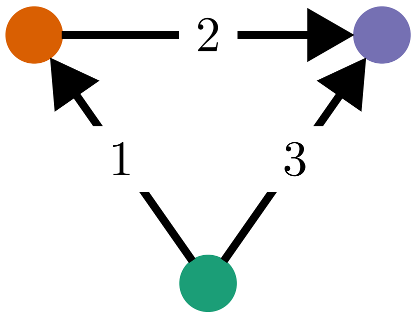
Temporal network motifs

Paranjape, Benson, & Leskovec, Motifs in Temporal Networks, WSDM, 2017.

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.