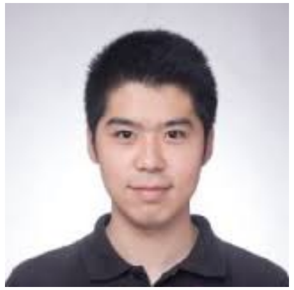
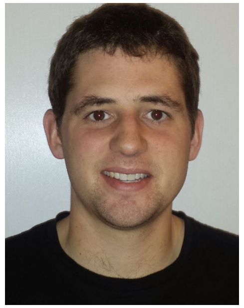
Local higher-order graph clustering



Hao Yin
Stanford University
yinh@stanford.edu



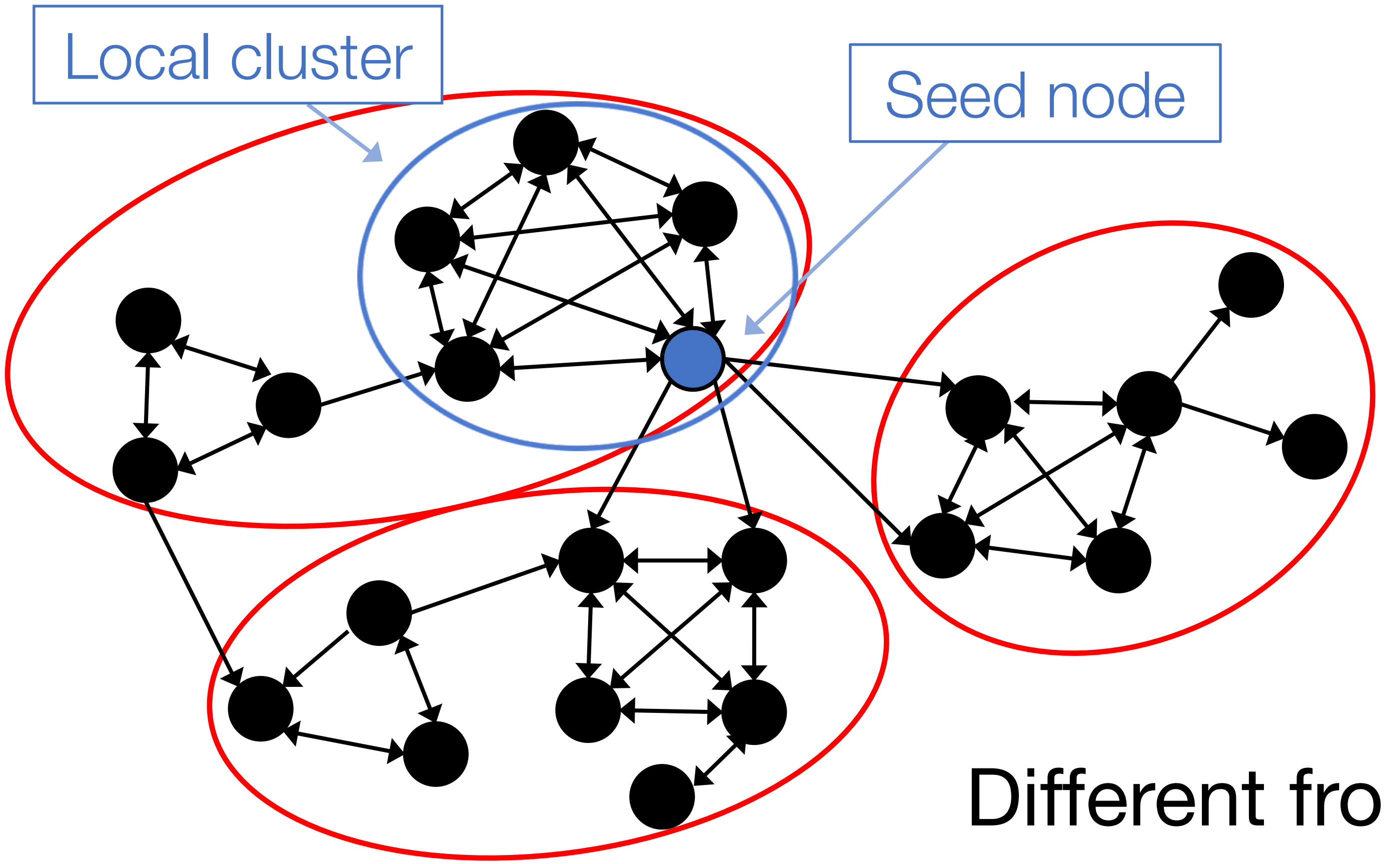
Austin R. Benson
Cornell University
arb@cornell.edu

Jure Leskovec
Stanford University
jure@cs.stanford.edu

David F. Gleich
Purdue University
dgleich@purdue.edu

* Code and data available at http://snap.stanford.edu/mappr

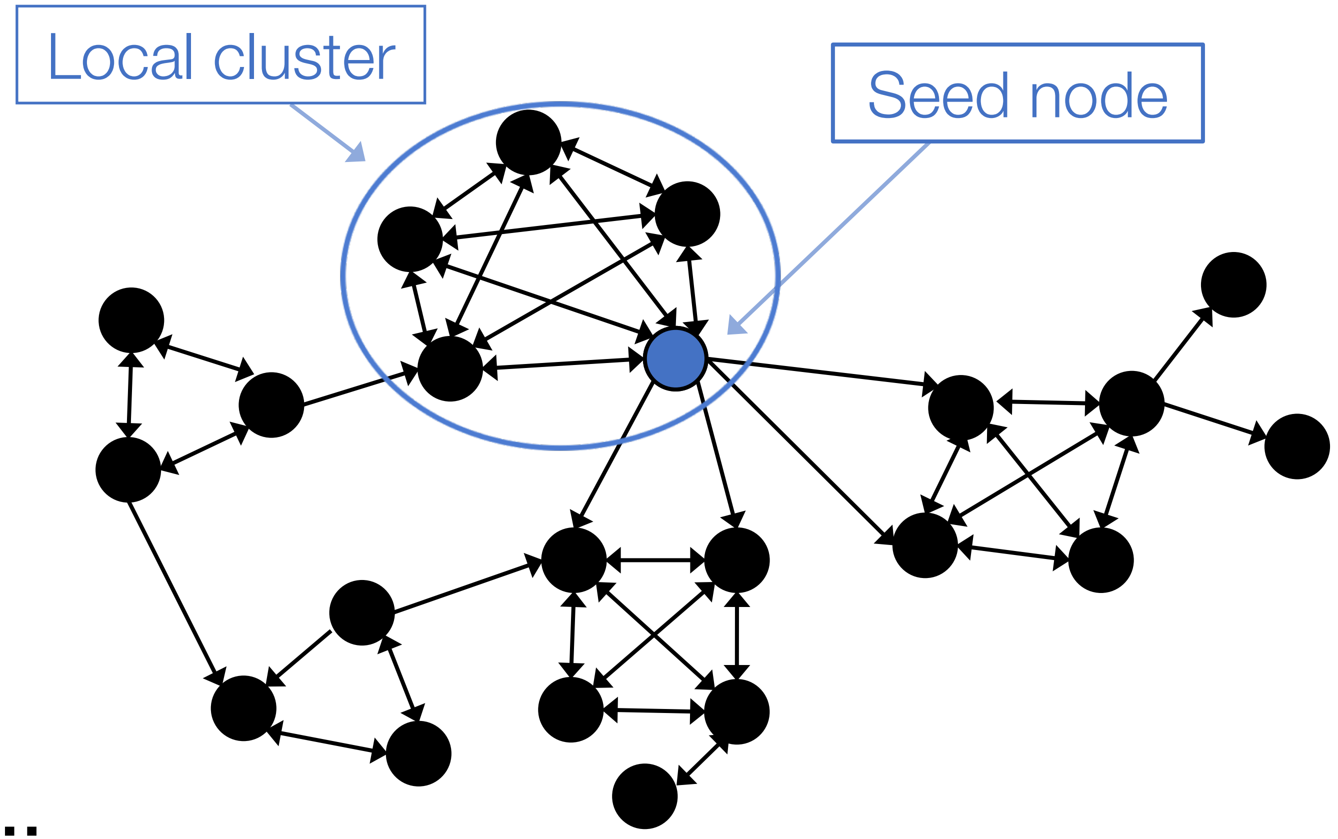
Background: Local clustering



Different from global clustering:

- Target community detection;
- Algorithm only explores a local neighborhood of seed node.

Background: Local clustering



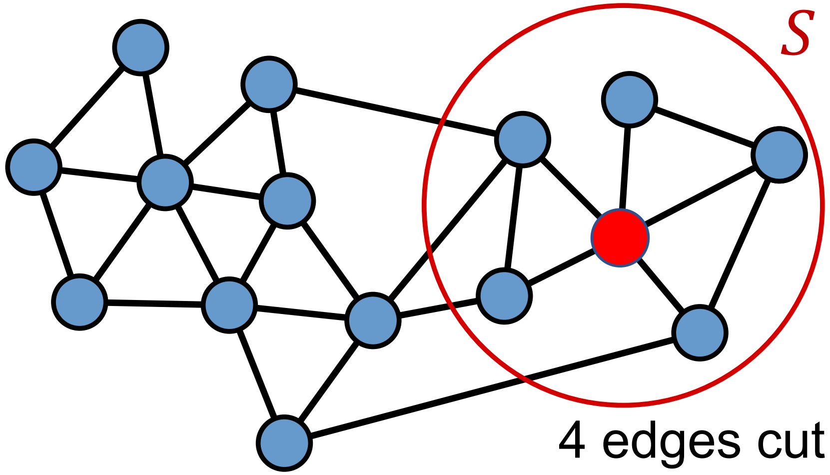
Used to…

- **find communities of an individual in social networks** [Jeub et al., *PRE*, 15].
- **find members of a protein complex in PPI networks** [Voevodski et al., *BMC Sys. Biol.*, 09].
- **find related videos in online media** [Gargi et al., *ICWSM*, 11].
- **and much more…** [Epasto et al., *WWW*, 14; Jiang et al., Sys. Biol., 09].

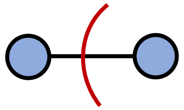
Background: Local clustering

- Existing methods find clusters with many internal edges and few external edges;
- Usually formulated as finding a cluster S with minimal (edge) conductance [Schaeffer, 07].

$$\phi(S) = \frac{\#(edges\ cut)}{Vol(S)}$$



4 edges cut
20 edge end points
edge conductance = 4 / 20

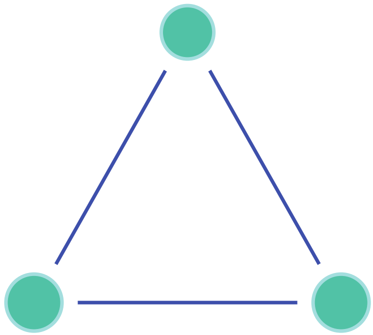
- $edges\ cut$: 
- $Vol(S) = \#(\text{edge end points in } S) = \sum_{u \in S} deg(u)$

However, edges are not the only interesting structures in networks!

Background: Higher-order structure

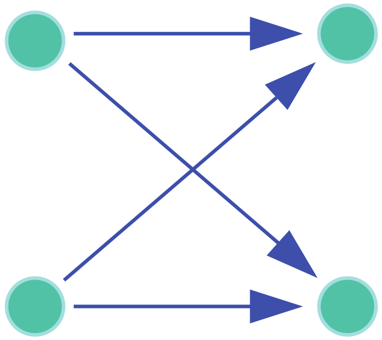
Higher-order connectivity patterns, or ***network motifs***, mediate complex networks.

Triangles in social networks.



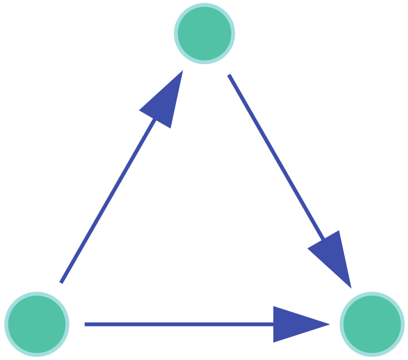
Rapoport, 1953;
Granovetter, 1973.

Bi-fans in neural connectivity networks.



Milo et al., 2002;
Benson et al., 2016.

Feed-forward loops in genetic transcription networks.

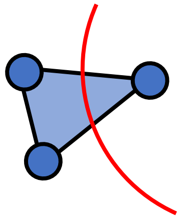


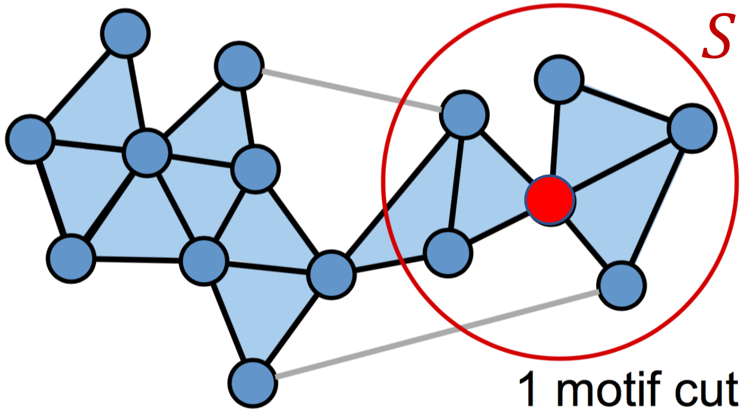
Mangan et al., 2003;
Alon, 2007.

Background: Motif-based graph clustering

Idea: Find a cluster S with minimal motif conductance [Benson et al., 16]

$$\phi_M(S) = \frac{\#(motifs\ cut)}{Vol_M(S)}$$

- $motifs\ cut$: 
- $Vol_M(S) = \#(\text{motif end points in } S)$



1 motif cut
11 motif end points
motif conductance = 1 / 11

Significant improvement in ground truth community detection and knowledge discovery [Benson et al., 16]

However, current motif-based clustering methods are global, and no motif-based local clustering method exists!

Our work: Local motif-based graph clustering

	Global method	Local method
Edge-based	[Fiedler, 1973]	[Anderson et al., 2006]
Motif-based	[Benson et al., 2016]	Our Work

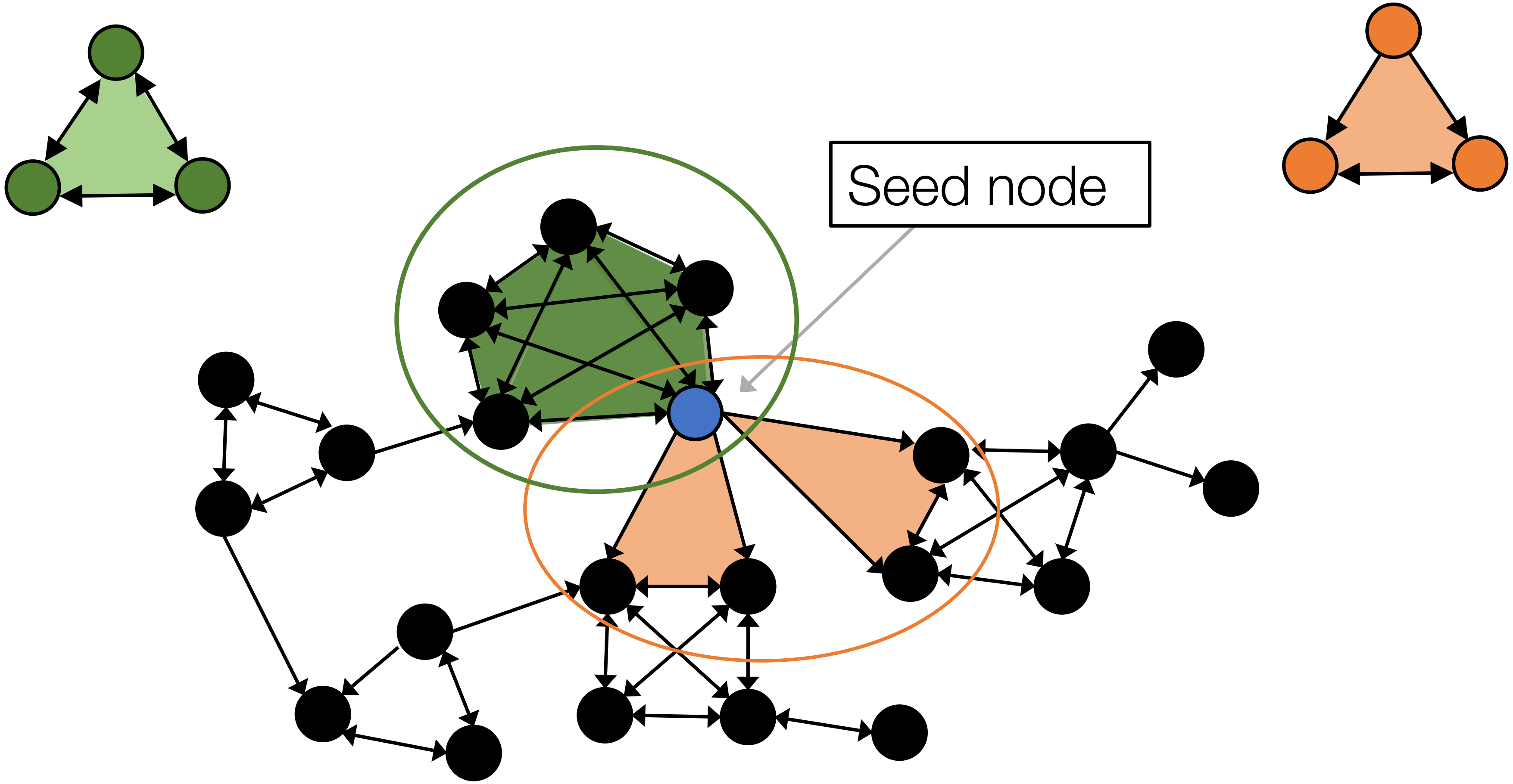
better results

scalable and faster

Problem:

- **Input:** a network, a seed node, and a motif.
- **Output:** a cluster containing the seed node with minimal motif conductance.

Our work: Local motif-based graph clustering



Different motifs give different local clustering results!

Our work: Local motif-based graph clustering

Challenge

- A generalization of the conductance minimization problem which is NP-hard [Wagner and Wagner, 1993].
- No existing methods for local clustering based on motif conductance.

Our approximate solution: MAPPR

Motif-based **A**pproximate **P**ersonalized **P**age**R**ank Algorithm

- A generalization of the APPR method [Andersen et al., 06].

MAPPR: Overview

- **Key ideas and steps:**
  1. Create a weighted graph with *weighted edge conductance* equals the *motif conductance* in the original graph;
  2. Find a cluster of minimal weighted edge conductance.

unweighted graph → weighted graph

$$\phi(S) = \frac{\#(edges\ cut)}{\sum_{u \in S} d(u)}$$

$$\phi_w(S) = \frac{\sum_{e \in cut(S)} w(e)}{\sum_{u \in S} d_w(u)}$$

$$d_w(u) = \sum_{u \in e} w(e)$$

MAPPR: Overview

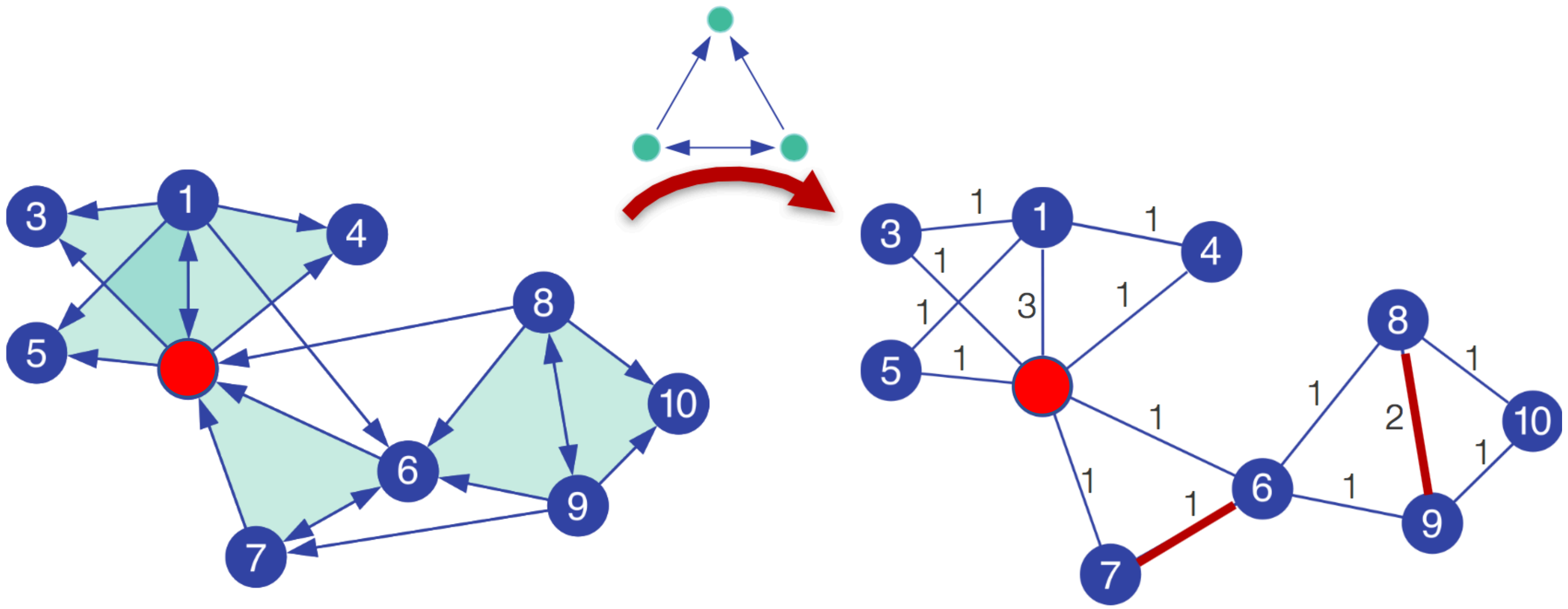
- **Key ideas and steps:**
  1. Create a weighted graph with *weighted edge conductance* equals the *motif conductance* in the original graph;
  2. Find a cluster of minimal weighted edge conductance.

- **Properties:**
  - Runtime guarantee
    - ➢ Procedure stops upon finding a good cluster, no need to explore the rest of graph.
  - Quality guarantee
    - ➢ Finds a near-optimal cluster regarding motif conductance.

MAPPR Step I: Weighted graph

- Create a weighted graph with

$$w(i, j) = \#\text{motif instances containing nodes } i \text{ and } j.$$



Graph G Weighted graph $W^{(M)}$

- The motif conductance (approximately) equals the **weighted edge conductance** in this weighted graph [Benson et at.,16].

MAPPR Step II: APPR vector

- Compute an approximate PPR vector for this weighted graph.
  - The PPR vector p is the stationary distribution of a random walk which at each step it "teleports" back to the seed with some probability.
  - $p(u)$ measures an "integrated closeness" of node u to the seed.
  - On a weighted graph, we choose each edge with **probability proportional to its weight**.
  - We adapted the approximate PPR algorithm [Anderson et al., 06] for weighted graphs.

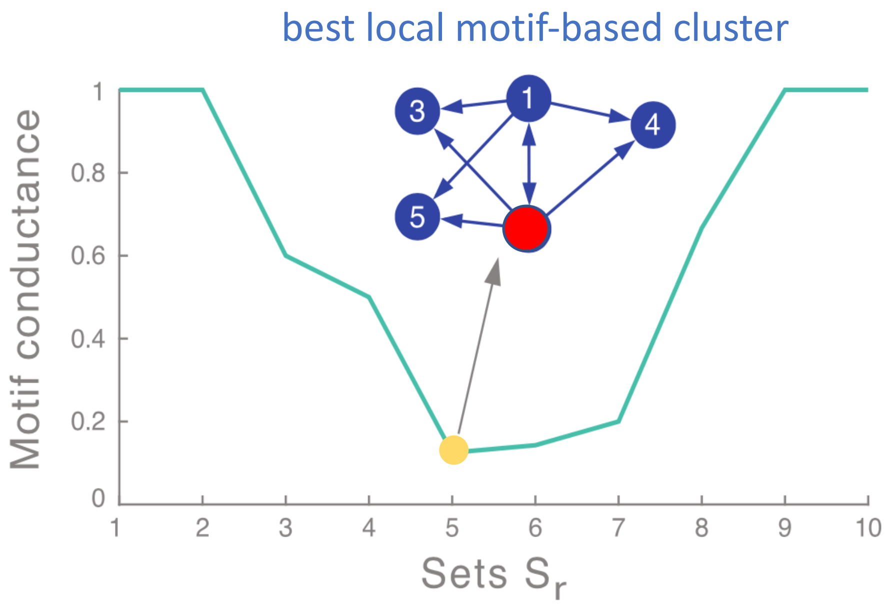
MAPPR Step III: Sweep

- Use the sweep procedure on APPR vector p to output the set with minimal weighted edge conductance [Anderson et al., 06].

1) Sort nodes by $p(u)/d_w(u) : u_1, u_2, \dots \mathrm{u}_{[\mathrm{p}]}$;
2) Compute the conductance of each

$$S_r = \{\, u_1, u_2, \dots, u_r \,\};$$

3) Output the S_r with minimal weighted edge conductance.



MAPPR: Runtime

- ## Theory

After motif counting, for each seed, the algorithm finishes in time proportional to the ***output cluster size***!

- No dependence on graph size!

Key part of proof: Interpret integer-weighted edges as parallel unweighted edges, then apply the previous analysis [Anderson et al., 06].

- ## Practice

Takes < 2 seconds / seed on 2 billion edge graphs!

- Global motif-based method takes 2 hours.

MAPPR: Quality

- ## Theory

For any unknown target community T, MAPPR seeded with most nodes in T would output a cluster S with

$$\phi_M(S) \leq \tilde{O}\left(\min\left(\sqrt{\phi_M(T)}, \phi_M(T)/\sqrt{\eta}\right)\right).$$

- η is the inverse mixing time of the subgraph induced by T.
- Guaranteed to find a near-optimal cluster.
- Results inherited from classic APPR analysis [Anderson et al., 06, Zhu et al., 13].

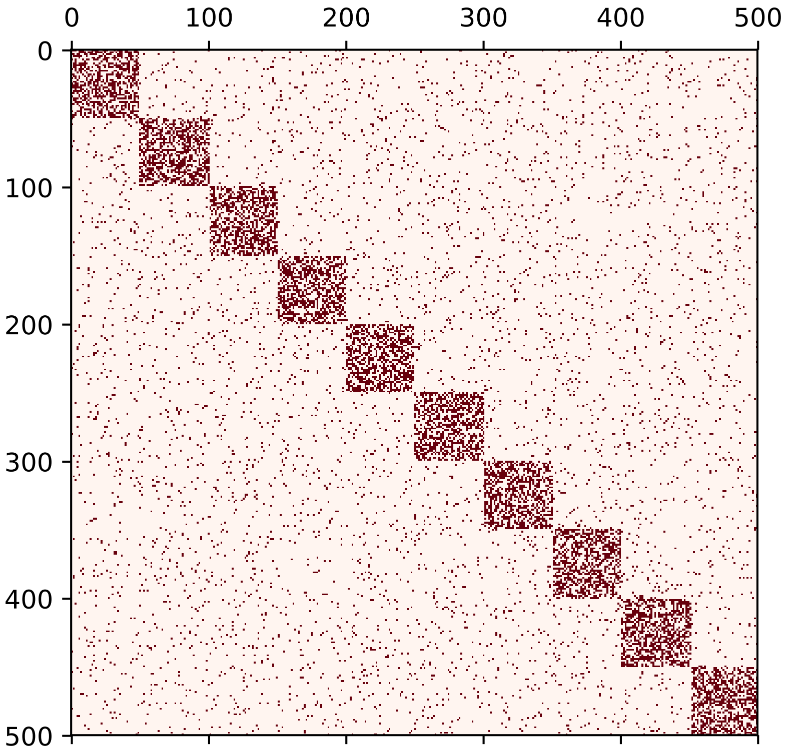
- ## Practice

- Better recovers ground truth communities!

MAPPR: Better discovery on synthetic networks

Random graph models with planted community structure:

1. Planted partition model



MAPPR: Better discovery on synthetic networks

Random graph models with planted community structure:

1. Planted partition model

2. Lancichinetti-Fortunato-Radicchi (LFR) model [Lancichinetti et al., 08, 09]

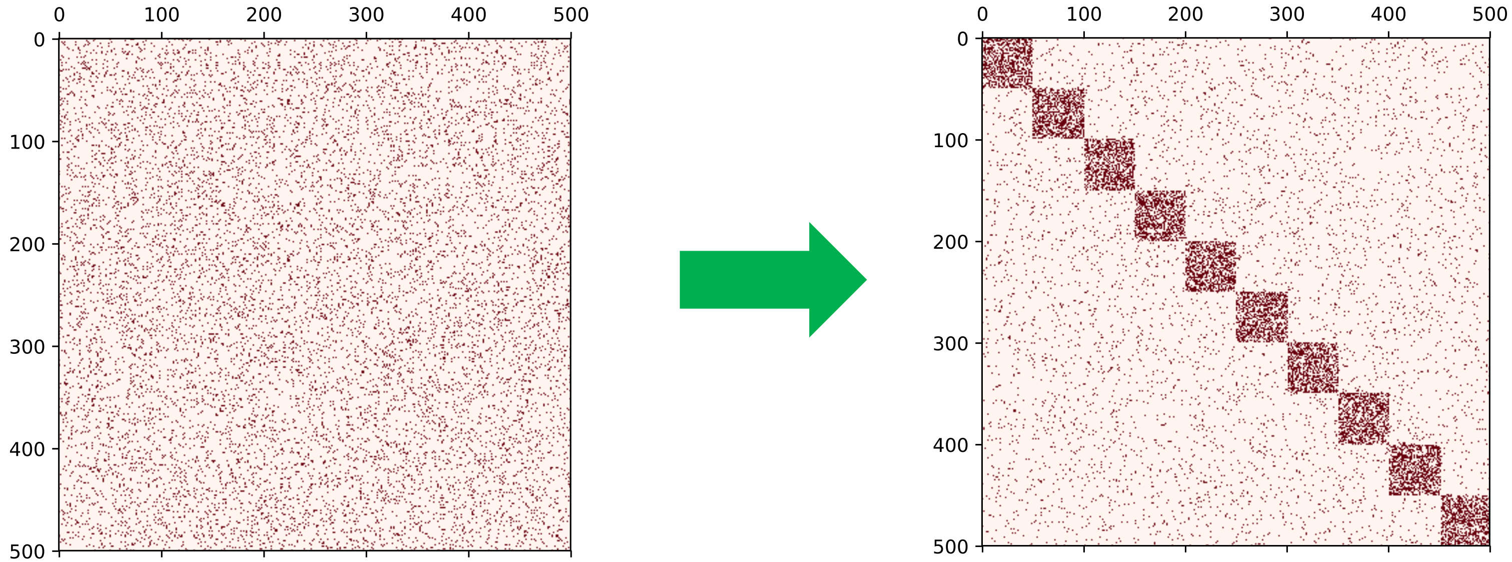
   A variant of planted partition model with

   - power-law degree distribution,
   - community overlapping,
   - power-law community size distribution, etc.

MAPPR: Better discovery on synthetic networks

Experiment Procedure:

- Seeded at every node, compute the F_1 score of the MAPPR cluster with the ground truth cluster, and take the average.

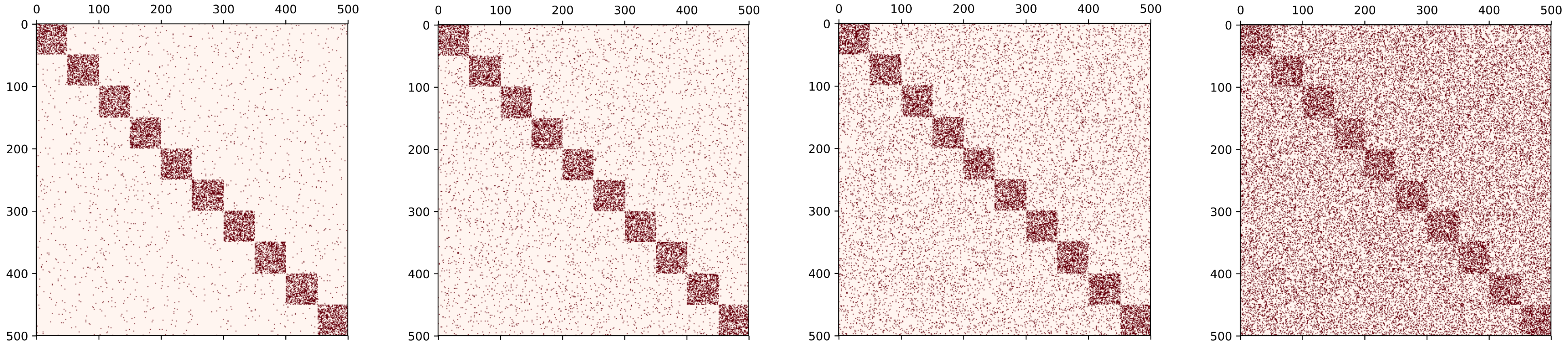


MAPPR: Better discovery on synthetic networks

Experiment Procedure:

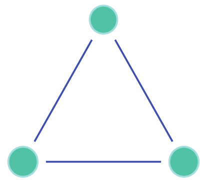
- Seeded at every node, compute the F_1 score of the MAPPR cluster with the ground truth cluster, and take the average.
- Repeat this experiment under different mixing level

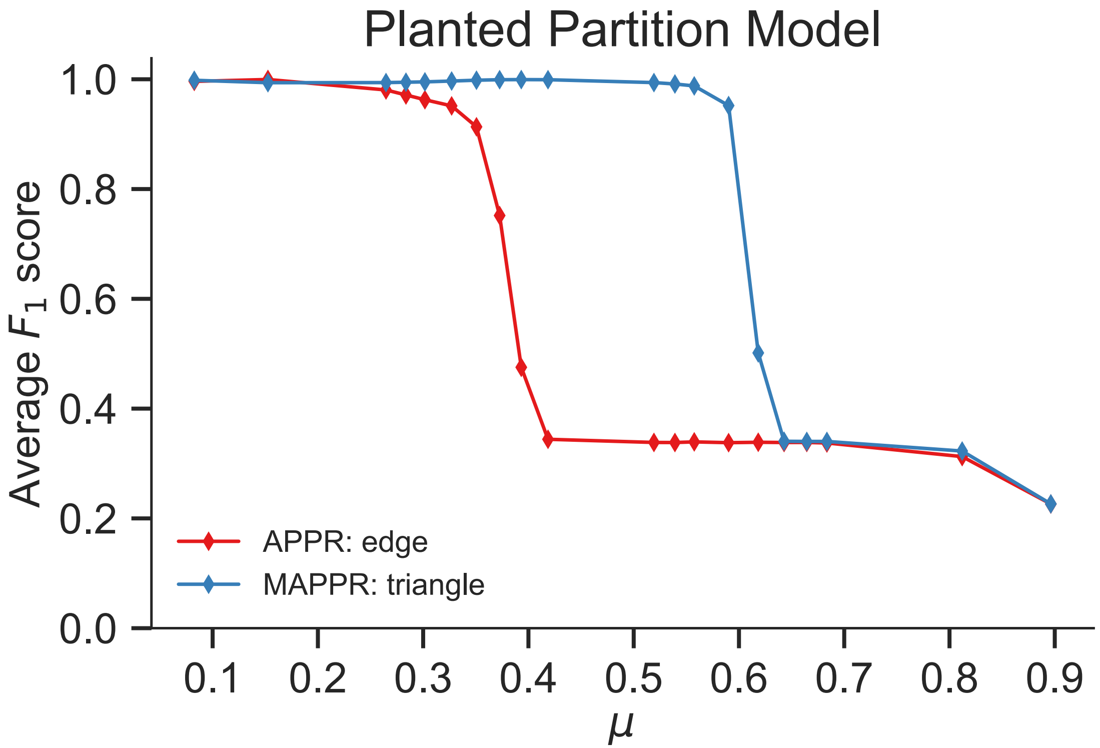
μ : fraction of edges across the ground truth communities.

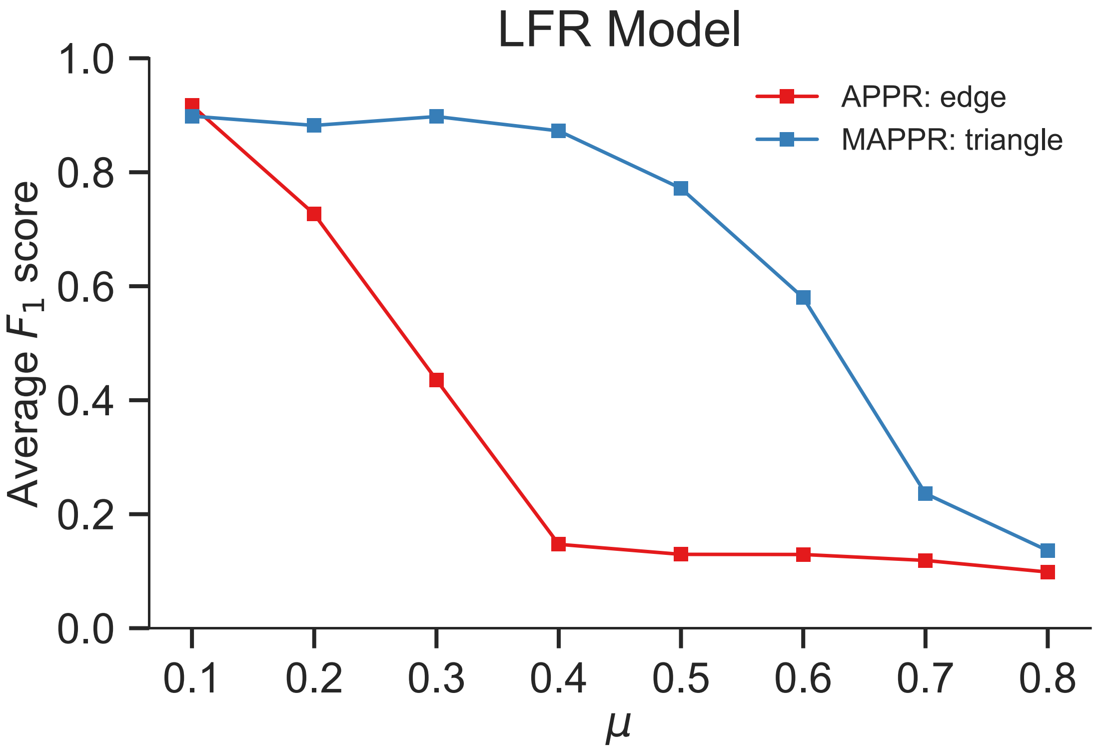


$\mu = 0.2$ $\mu = 0.4$ $\mu = 0.6$ $\mu = 0.8$

MAPPR: Better discovery on synthetic networks







- A large region of μ where MAPPR with triangle motif outperforms the edge-based method!
- Intuition: Edges across different communities are less likely to form a triangle.

MAPPR: Better discovery on real-world networks

email-Eu network

- Nodes are people at a research institute. Each person belongs to a department.
- Edges are email correspondence.
- Given a person as a seed, can we recover other members of his/her department?

Graph Statistics:	
Nodes:	1K
Edges:	25.6K
Departments:	28
Depart. Sizes:	10 -- 109

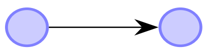
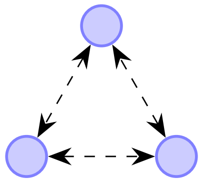
New dataset!

`http://snap.stanford.edu/data/email-Eu-core.html`

MAPPR: Better discovery on real-world networks

email-Eu network

- Nodes are people at a research institute. Each person belongs to a department.
- Edges are email correspondence.
- Given a person as a seed, can we recover other members of his/her department?

Method		Average F_1
APPR		0.40
MAPPR		0.50

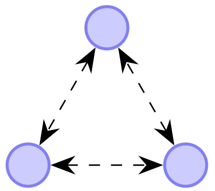
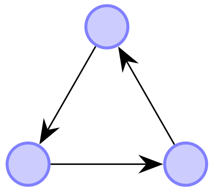
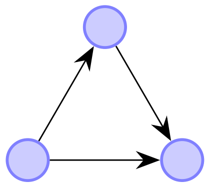
A 25% improvement!

❖ Triangle motif better discovers ground truth communities!

MAPPR: Better discovery on real-world networks

email-Eu network

- Nodes are people at a research institute. Each person belongs to a department.
- Edges are email correspondence.
- Given a person as a seed, can we recover other members of his/her department?

Motif used in MAPPR	Average F_1
	0.50
	0.45
	0.47

❖ Both feed-forward loop and cycle are important in discovering communities structure in communication network!

More in the paper: Finding good seeds

- Idea: A vertex with low 1-hop neighborhood motif conductance has lower motif conductance in its MAPPR cluster.

- Theory: Vertex 1-hop neighborhoods have low motif conductance.
  - Related with *higher-order clustering coefficient* [Yin et al., 17].

Recap

- Proposed the MAPPR algorithm for local motif-based graph clustering
  - ✓ Runtime guarantee
  - ✓ Quality guarantee
  - ✓ Better recovery in synthetic models and real-world datasets
- Finding Good Seeds

Local higher-order graph clustering

Hao Yin, Austin R. Benson, Jure Leskovec, David F. Gleich

* Code and data available at http://snap.stanford.edu/mappr