Motif-Driven Graph Analysis

Charalampos (Babis) Tsourakakis
Graph Exploitation Symposium
MIT Endicott House

Boston University, **Harvard University**
babis@seas.harvard.edu

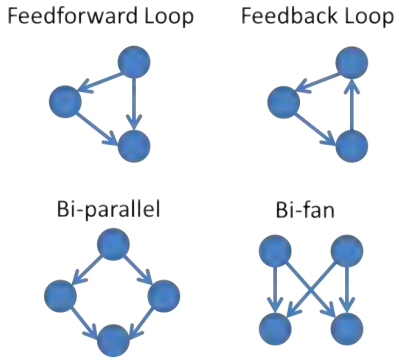
16 May 2017

Motivating Question

- *How can we effectively leverage motifs* for better cluster detection in graphs?*

*"Basic interaction patterns that recur throughout networks, much more often than in random networks. " (Uri Alon)

Examples of important motifs



Friends of friends tend to become friends themselves

Over-represented motifs usually represent functional units of biological processes in cells

Outline

0. **Related work**
1. **Motif-aware scalable dense subgraph discovery**, and **anomaly detection** (densest subgraph sparsifiers)
2. **Motif-aware graph clustering** (motif-expander graphs)
3. **Open problems**

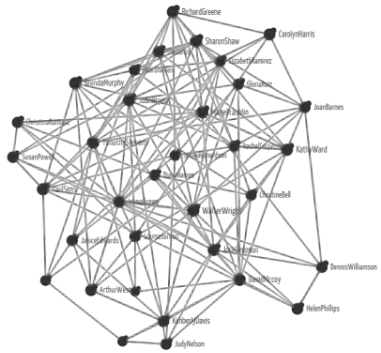
Joint work with:

Jakub Pachocki

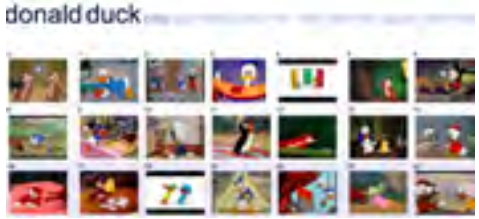
Michael Mitzenmacher

Today's graph mining challenges
Large near-clique and Community detection

Who-calls-whom



Topic clusters



	Security	**YouTube**
V	Humans	Videos
E	Phone call	Co-watching
Cluster	Suspicious Large-Near Clique	Thematic coherence Community

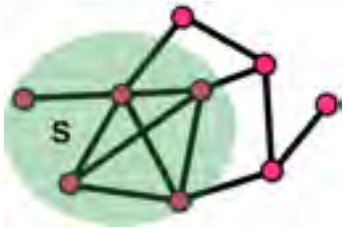
Dense Subgraph Discovery – Possible formulations

Edge density: $0 \leq \boxed{f_e(S) = \frac{e(S)}{\binom{|S|}{2}}} \leq 1$

Independent set Clique

- $\max_{S \subseteq V} f_e(S)$ is **ill-posed** since $f_e(\bullet\!\!-\!\!\bullet) = 1$
- $\max_{S \subseteq V} f_e(S)$ subject to $|S| =, \geq k$ is **NP-hard**

Related work – densest subgraph problem

Degree density: $\boxed{\rho(S) = \frac{e(S)}{|S|}}$. E.g., 

$\rho(S) = \frac{7}{5}$

- $\max_{S \subseteq V} \rho(S)$ **Poly-time solvable** (via max flows)
  [Goldberg, 1984, Gallo et al., 1989, Khuller et al., 2009]

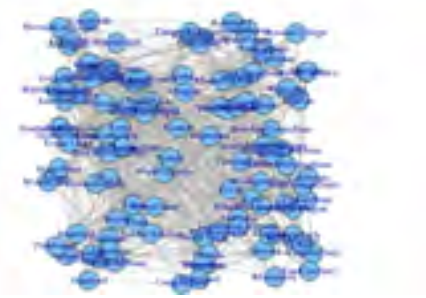
  2-approximation algorithm which uses linear space $O(n+m)$ and runs in linear time $O(n+m)$ due to Charikar

  "The densest subgraph problem **(DSP)** *lies at the core of large scale data mining"* [Bahmani et al., 2012]

- $\max_{S \subseteq V} \rho(S)$ subject to cardinality constraints is **NP-hard**

Related work – densest subgraph problem

Solving the **DSP** typically **does not** result in "clique"-like sets.



- FOOTBALL NETWORK with
  $n = 115$ vertices $\leftrightarrow$ teams
  $m = 613$ edges $\leftrightarrow$ games

- The optimal solution S^* to the **DSP** is the whole network with resulting degree density $\rho(S^*) = 5.3$ and edge density $f_e(S) = \frac{e(S)}{\binom{|S|}{2}} = 0.094$.

- There exists S' such that $|S'| = 18, f_e(S') = 0.48$. However $\rho(S') = 4.1$.

Related Work – Community Detection

- **Conductance** of a node set $S \subseteq V$ is defined as

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

where $w(S : \bar{S}) = \sum_{i \in S, j \in \bar{S}} w(i, j)$, and $\text{vol}(S) = \sum_{i \in S} deg(i)$.

- **Graph conductance**, is defined as $\phi(G) = \min_S \phi(S)$.
- **Expanders:** Intuitively, a graph that contains no set S with low conductance.
- **Intuition:** If $\phi(G)$ is low, there exists a community.

Related Work – Community Detection

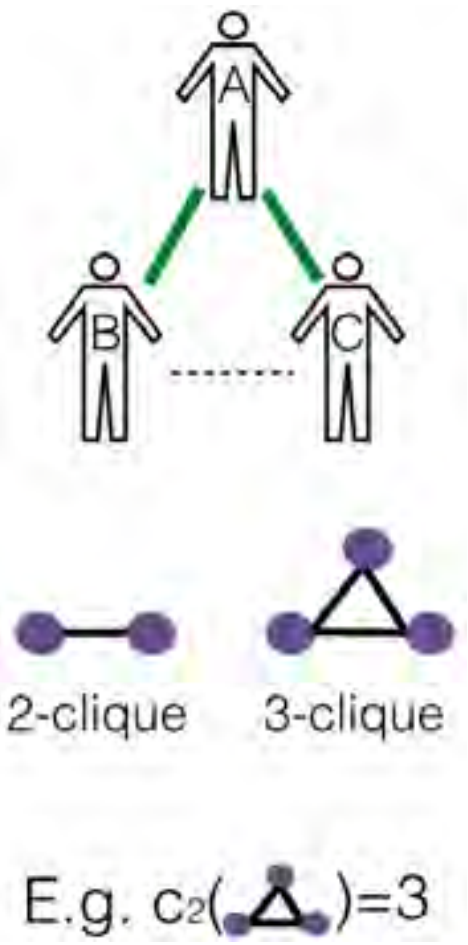
- **Spectral clustering (SC) – Cheeger Inequality:** Let $0 = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n \leq 2$ be the eigenvalues of the normalized Laplacian matrix.

$$\frac{\lambda_2}{2} \leq \phi(G) \leq \sqrt{2\lambda_2} \quad \text{[Alon and Milman, 1985]}$$

Community detections spans a very long line of research, some important algorithms:

- **MCL** [Dongen, 2000], **Infomap** [Rosvall and Bergstrom, 2008], **Girvan-Newman (GN)** [Girvan and Newman, 2002], **Louvain method** [Blondel et al., 2008], **Clauset-Newman-Moore (CNM)** [Clauset et al., yhtm], **Cfinder** [Adamcsek et al., 2006]

k-clique densest subgraph problem



For any $S \subseteq V$ let

$$c_k(S) = \#\text{ k-cliques induced by } S.$$

Define ***k*-clique density**

$$\rho_k(S) = \frac{c_k(S)}{s}, k \geq 2, s = |S|$$

Solve the ***k*-clique DSP**

$$\rho_k(S^*) = \rho_k^* = \max_{S \subseteq V} \rho_k(S)$$

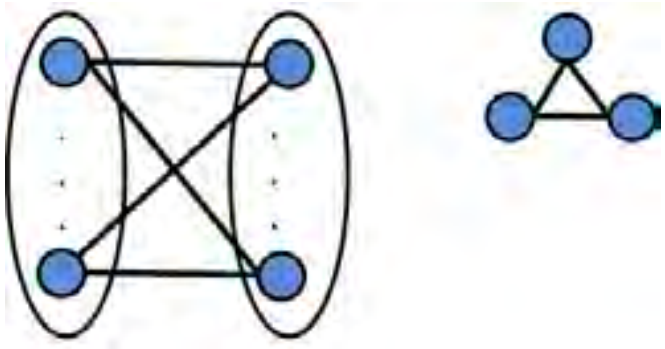
Triangle densest subgraph problem

We shall refer to the 3-clique DSP as the *triangle densest subgraph problem*.

$$\max_{S \subseteq V} \tau(S) = \frac{t(S)}{s}$$

.

How different can the densest subgraph be from the triangle densest subgraph? Radically different,e.g., $G = K_{n,n} \cup K_3$.



What happens on real-data? Can we solve the triangle DSP in polynomial time? The k-clique DSP?

Triangle densest subgraph problem

Theorem

There exists an algorithm which solves the TDSP and runs in $O\left(m^{3/2} + nt + \min\left(n, t\right)^3\right)$ *time.*

Furthermore,

Theorem

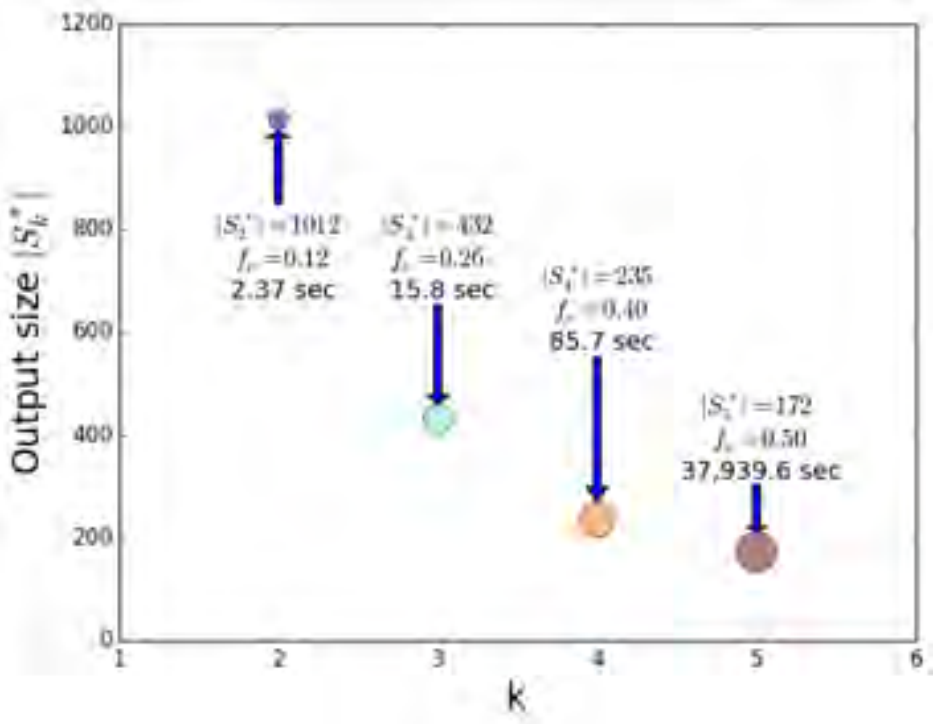
We can solve the k-clique DSP in polynomial time for any $k = \Theta(1)$.

Computation involves

- Enumerate the set C_k of k-cliques in G
- Maximum flow on an appropriate network $\mathcal{N}(\{s, t\} \cup V \cup C_k, A)$

Epinions social network

# nodes	75 877
# edges	405 739

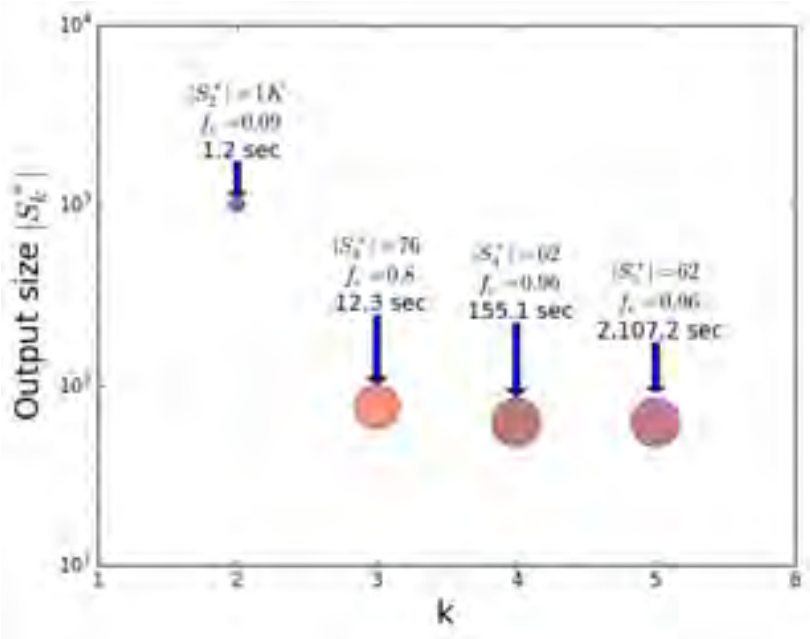


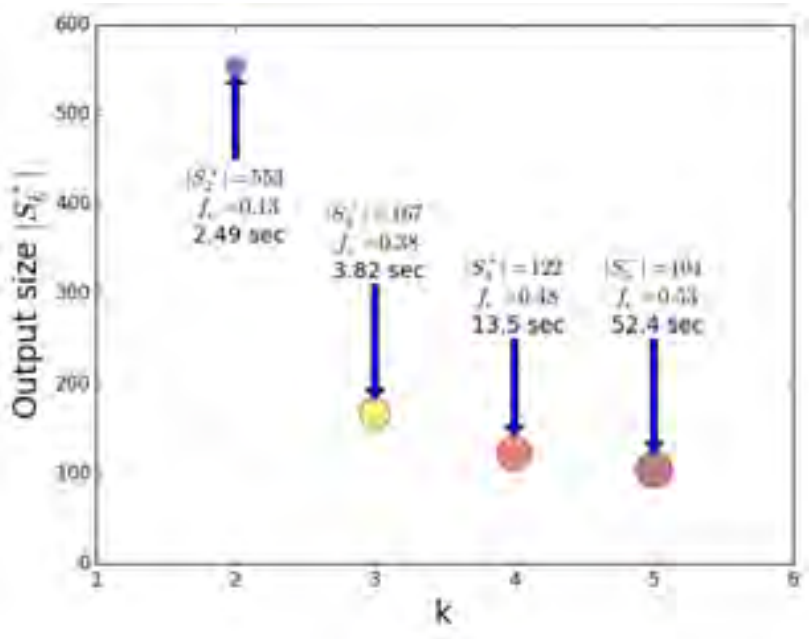
k	c_k	T_k (sec)
3-clique	1.6M	1.6
4-clique	5.8M	4.8
5-clique	17.4M	13.4

CA-Astro and Email networks

# nodes	18 772
# edges	198 050

# nodes	234 352
# edges	383 111



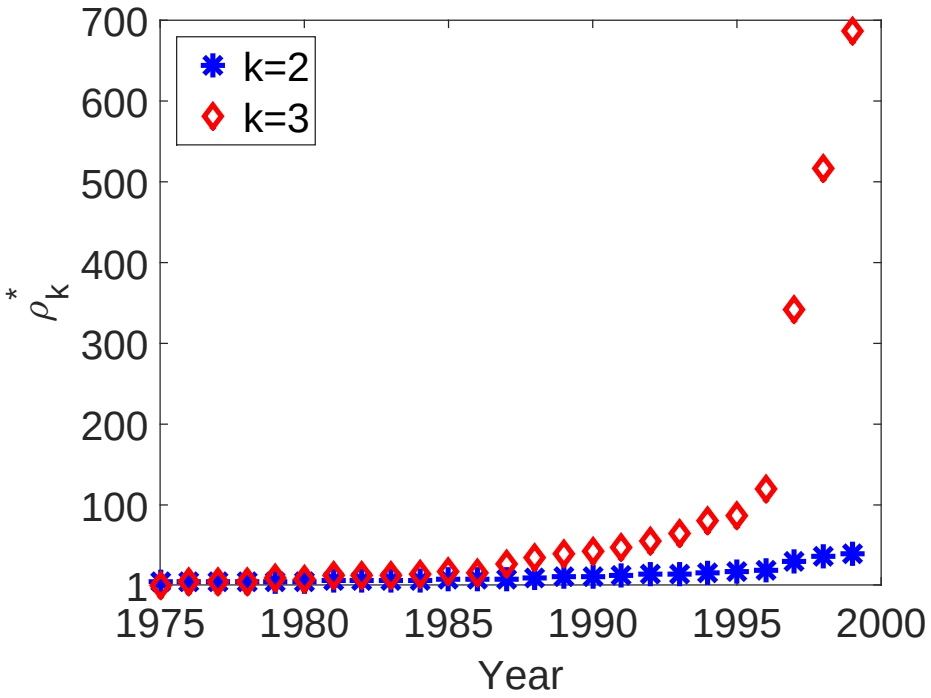


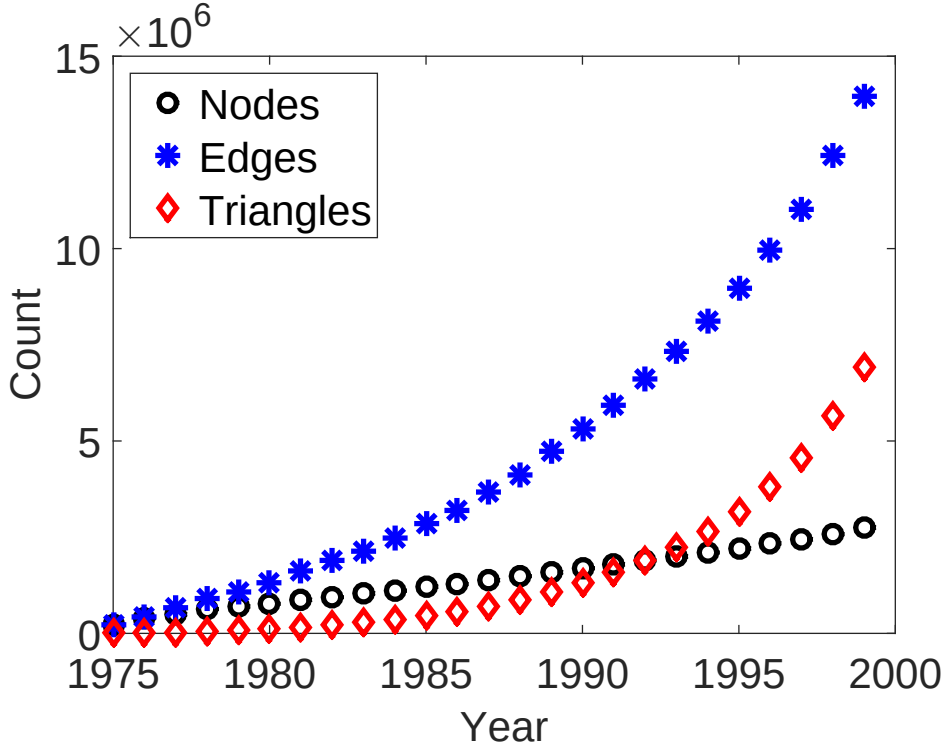
k	c_k	T_k (sec)
3-clique	1.4M	0.6
4-clique	9.5M	3.9
5-clique	65M	27.2

k	c_k	T_k (sec)
3-clique	0.4M	0.4
4-clique	1M	0.9
5-clique	2.6M	1.9

Time evolving networks

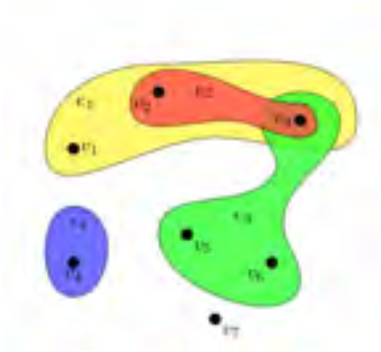
Patents citation network that spans 37 years, specifically from January 1, 1963 to December 30, 1999.





Densest subgraph sparsifiers

Definition: "A hypergraph is a generalization of a graph in which an edge can connect any number of vertices."



Densest subgraph sparsifier theorem

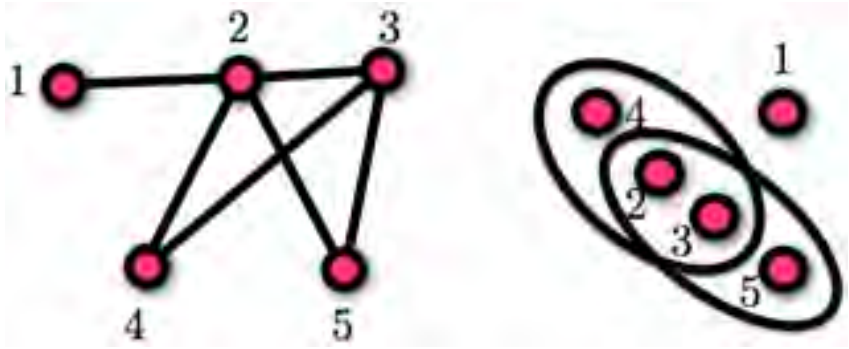
Let $\mathcal{H}(V, E_{\mathcal{H}})$ be a hypergraph, $\epsilon > 0$.

Let $E' \subseteq E_{\mathcal{H}}$ be a sample of $\frac{6n \log n}{\epsilon^2}$ edges chosen uniformly at random.

Solving the DSP on E' results in a $(1 + \epsilon)$ approximation to ρ^* **whp**.

Densest subgraph sparsifiers

Some hypergraphs of interest



- Hypergraphs $\rightarrow$ k-clique DSP

  Important corollary: Single-pass, $(1+\epsilon)$ semi-streaming algorithm! Just keep $O(n \log n/\epsilon^2)$ edges.

- Previous best known streaming algorithm required $O(\log n/\epsilon)$ passes [Bahmani et al., 2012].

Densest subgraph sparsifiers

Think of sampling each edge **wp** $p = \frac{6n \log n}{|E_{\mathcal{H}}|\epsilon^2}$.

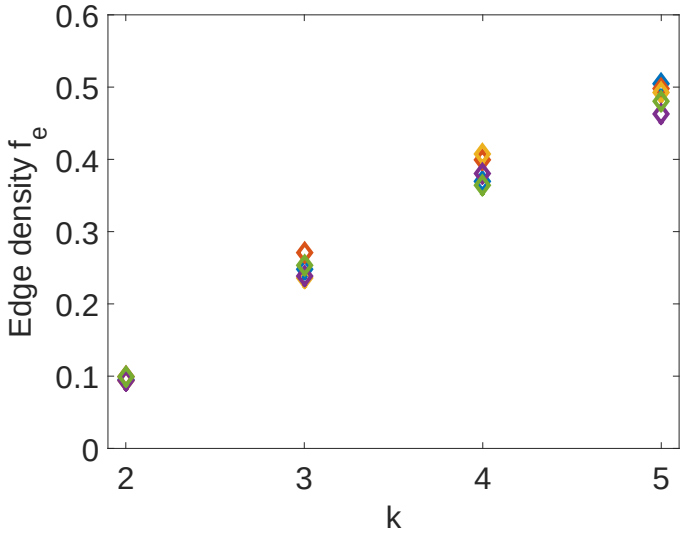
Expected space reduction is $O(\frac{1}{p})$.

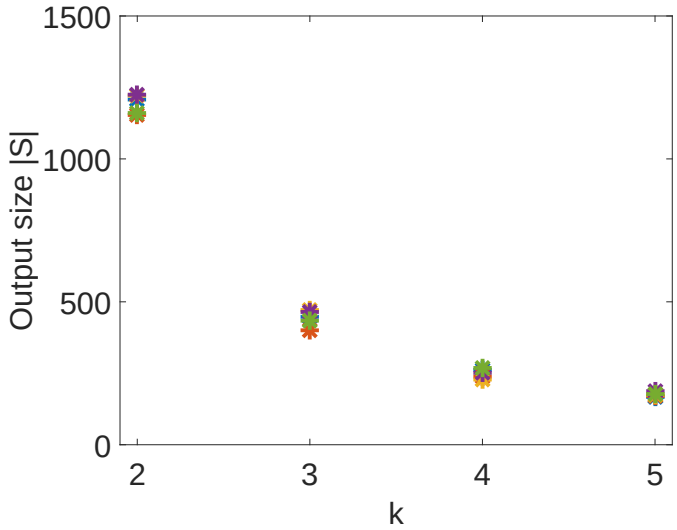
Expected speedup for maximum flow computation $O(\frac{1}{p^2})$

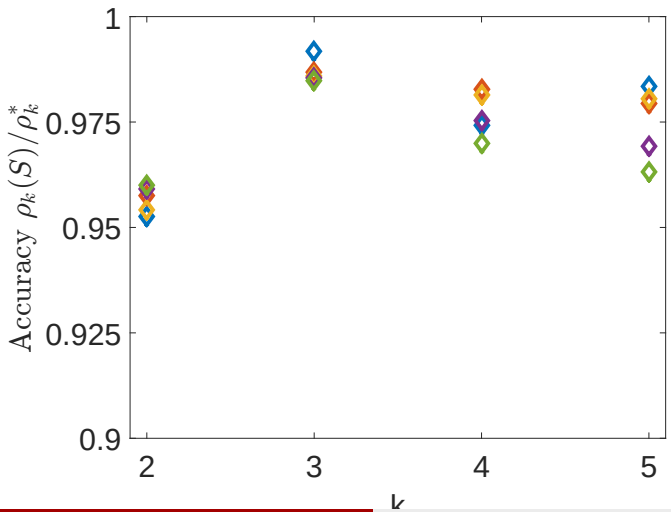
k	Avg. Speedup	Accuracy
2	$3\times$	$\geq 95\%$
3	$23.8\times$	$\geq 98\%$
4	$302\ \times$	$\geq 99\%$
5	$24\,000\times$	$\approx 100\%$

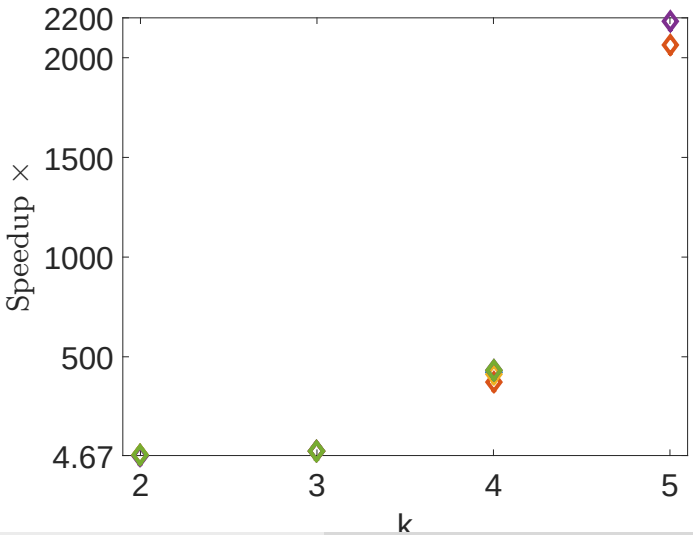
Zooming-in on Epinions network

EPINIONS graph, $n = 75\,877, m = 405\,739$







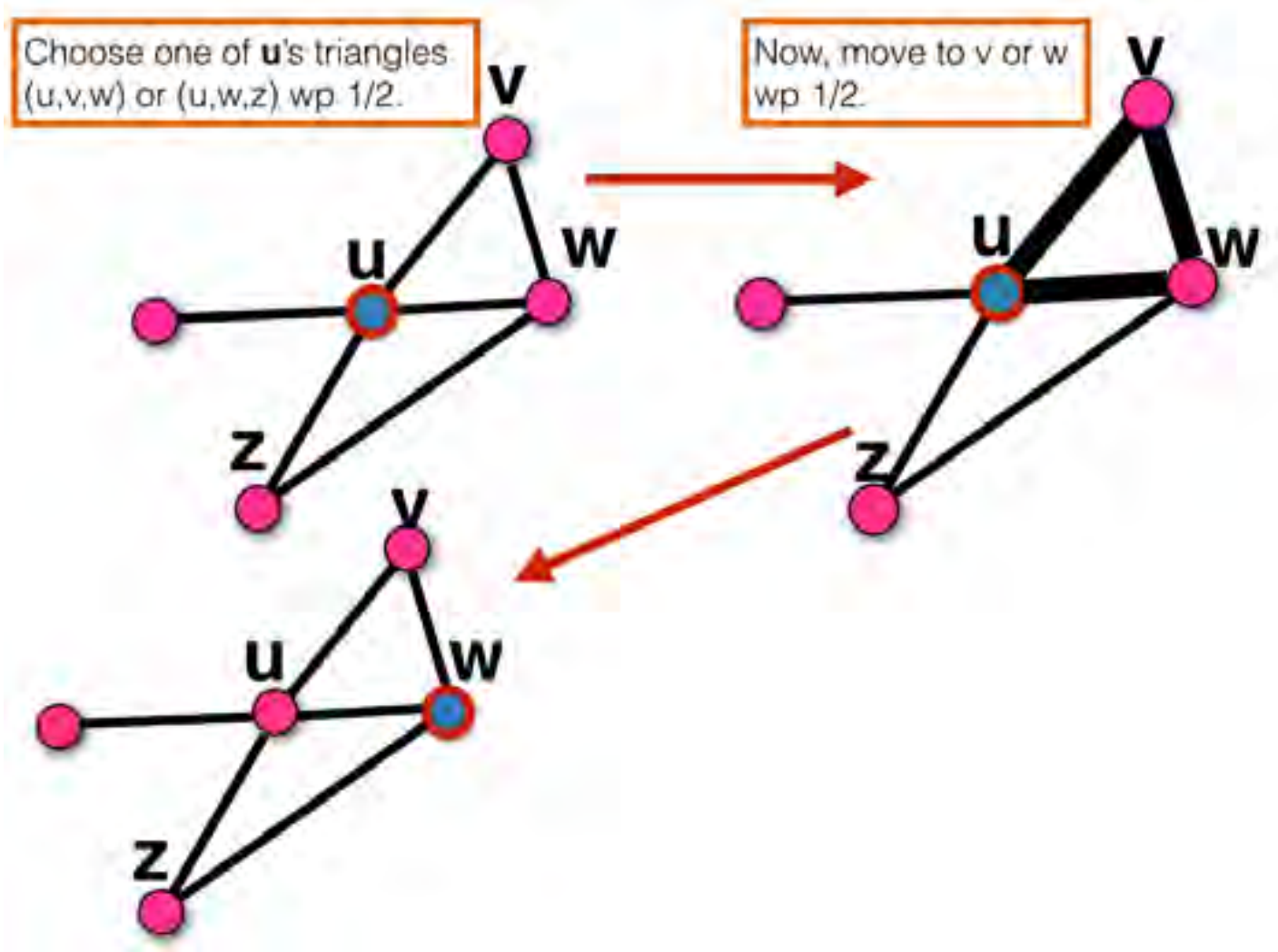


Community Detection – Our Contributions

1. We define a new motif-based random walk $\xrightarrow{\text{that results in}}$ new notion of motif-conductance

   Our RW framework formalizes heuristic formulations by Benson, Gleich, Leskovec (**Science 2016**)
2. Theoretical foundations **(motif-based expanders)**
3. Spectral motif-clustering (overlaps with **Science 2016**), and a well-performing heuristic TECTONIC (Triangle Connected Component Clustering)
4. Theoretical results for the planted partition model
5. Experiments on communities with groundtruth

Motif-Based Random Walk



Motif-Based Random Walk

- Let $S \subseteq V$ be any set of vertices
- $\phi_3(S)$ the probability of leaving S in one step of the walk conditioned on being at a vertex u chosen from S proportionally to the number of triangles $t(u)$ it participates in
- Define $t_i(u)$ is the number of triangles with i vertices in S (u included)
- $\phi_3(S) := \sum_{u \in S} \frac{t(u)}{\text{vol}_3(S)} \frac{0 \times t_3(u) + 0.5 \times t_2(u) + 1 \times t_1(u)}{t(u)} = \frac{t_2(S) + t_1(S)}{\text{vol}_3(S)}$
- Graph Triangle Conductance: $\phi_3(G) = \min_{S \subseteq V} \phi_3(S)$

Triangle Spectral Clustering

- **Observation:** Minimizing triangle conductance in G is equivalent to minimizing conductance in a hypergraph $\mathcal{F}$
  One hyperedge $e \in \mathcal{F}$ for each triangle $(u, v, w) \in G$
  Problematic approach from an algorithmic perspective
- **Idea**: No need to create a hypergraph! Just reweight each edge in G by the number of triangles it participates in.
- **Next,** apply your favorite clustering algorithm on $H(V, E, w)$, the weighted version of G
- **Theorem:** Cheeger's clustering algorithm on $H(V, E, w)$ outputs a cut $(S : \bar{S})$ such that

$$\frac{\lambda_2(H)}{2} \leq \phi_3(G) \leq \sqrt{2\lambda_2(H)}.$$

Triangle Spectral Clustering

- Our work and the Science paper by Benson, Gleich, and Leskovec appeared independently at the same time and share the algorithmic contribution of performing efficiently motif-based clustering

  Contributions (diff): of our work that don't appear in Science

  1. The RW interpretation of the graph reweighting scheme, that provides a principled approach to define the notion of conductance for other motifs;
  2. Theoretical foundations for motif-based graph clustering (expanders)
  3. Planted Partition model
  4. TECTONIC heuristic
  5. Experimental evaluation on real-world networks with ground-truth communities.

TECTONIC Heuristic

Input: Graph $G(V, E)$, threshold $\theta > 0$

1. Count $t(u, v)$ for each $(u, v) \in E$
2. Reweight each edge $(u, v) \in E$ by $w(u, v) \leftarrow \frac{t(u,v)}{deg(u)+deg(v)}$
3. Remove all edges (u, v) with weight $w(u, v) < \theta$
4. Output the resulting connected components

Inspired by hierarchical clustering, since

$$\frac{t(u, v)}{deg(u) + deg(v)} < \theta \Leftrightarrow dist^2(A^{(u)}, A^{(v)}) > \theta',$$

for $\theta = \frac{1}{2}\left(1 - \frac{\theta'}{deg(u)+deg(v)}\right)$.

Experimental Results

Groundtruth communities (available at SNAP)

p precision, r recall, T run time in seconds

Quality: MCL, TECTONIC

Speed: TECTONIC

Method	Amazon			DBLP			YouTube		
	p	r	T	p	r	T	p	r	T
MCL	95.6	90.1	736.54	55.1	81.7	1 166	39.9	60.6	19 187.1
Louvaine	50.0	14.7	9.00	50.20	12.13	10.38	50.13	27.55	55.8
CFinder	-	-	$> 5h$	-	-	$> 5h$	-	-	$> 5h$
GN	-	-	$> 5h$	-	-	$> 5h$	-	-	$> 5h$
CNM	-	-	$> 5h$	-	-	$> 5h$	-	-	$> 5h$
Infomap	50.0	14.8	63.0	50.16	12.13	64.0	50.00	27.6	204
SC	-	-	$> 5h$	-	-	$> 5h$	-	-	$> 5h$
tSC	-	-	$> 5h$	-	-	$> 5h$	-	-	$> 5h$
Thres. 0	85.2	96.0	4.62	4.0	100.0	1.65	22.5	70.8	6.92
Thres. 1	94.1	81.1	4.61	12.0	91.4	1.65	36.1	59.7	6.92
Thres. 2	97.1	67.7	4.62	23.0	81.6	1.65	45.0	53.9	6.92
Thres. 3	98.0	52.4	4.62	35.7	71.4	1.65	49.6	50.3	6.93
TECTONIC	94.9	91.3	4.62	48.3	79.1	1.65	66.7	43.3	6.92

Research directions

- Random graph models. Can we use stochastic graph models to explain some of our findings?
- Data-aware graph algorithms. Can we exploit idiosyncrasies of real-world networks to solve computationally challenging problems, including NP-hard problems?
- Motif-aware semi-supervised learning (SSL). Can we improve graph based SSL using motifs?
- Variations on reweighting: Use $w(e) = 1 + \alpha t(e)$ for some parameter α. How do we set optimally weights?

Thank you! Questions?

email: babis@seas.harvard.edu

github: https://github.com/tsourolampis

web page: http://people.seas.harvard.edu/~babis/

references I

Adamcsek, B., Palla, G., Farkas, I. J., Derényi, I., and Vicsek, T. (2006).
CFinder: locating cliques and overlapping modules in biological networks.
Bioinformatics, 22(8):1021–1023.

Bahmani, B., Kumar, R., and Vassilvitskii, S. (2012).
Densest subgraph in streaming and MapReduce.
Proceedings of the VLDB Endowment, 5(5):454–465.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008).
Fast unfolding of communities in large networks.
Journal of statistical mechanics: theory and experiment, 2008(10):P10008.

references II

Clauset, A., Newman, M. E., and Moore, C. (2004).
Implementation at
https://www.cs.unm.edu/~aaron/research/fastmodularity.htm).

Finding community structure in very large networks.

Physical review E, 70(6):066111.

Dongen, S. v. (2000).

Graph clustering by flow simulation.

Girvan, M. and Newman, M. E. J. (2002).

Community structure in social and biological networks.

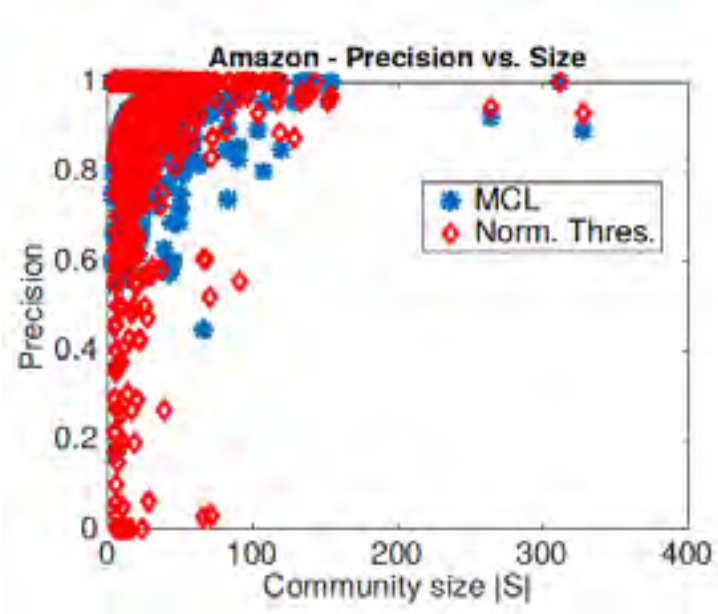
Proceedings of the National Academy of Sciences, 99(12):7821–7826.

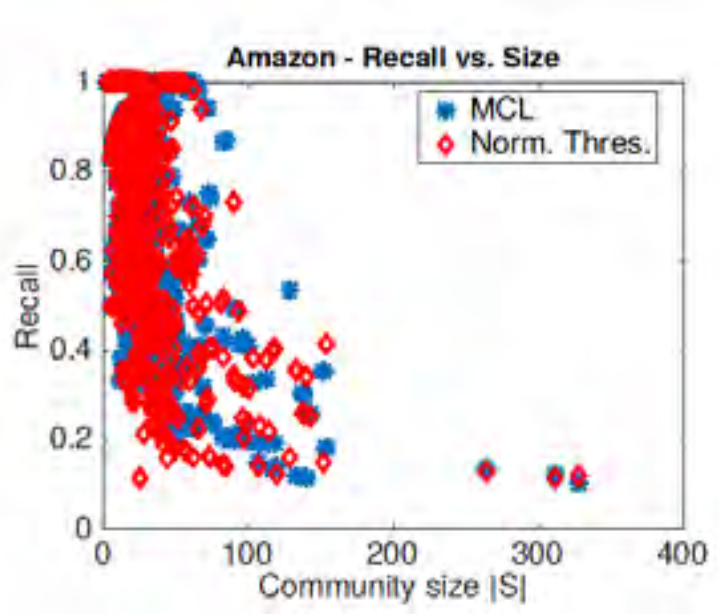
Rosvall, M. and Bergstrom, C. T. (2008).

Maps of random walks on complex networks reveal community structure.

Proceedings of the National Academy of Sciences, 105(4):1118–1123.

Experimental Results – MCL and TECTONIC

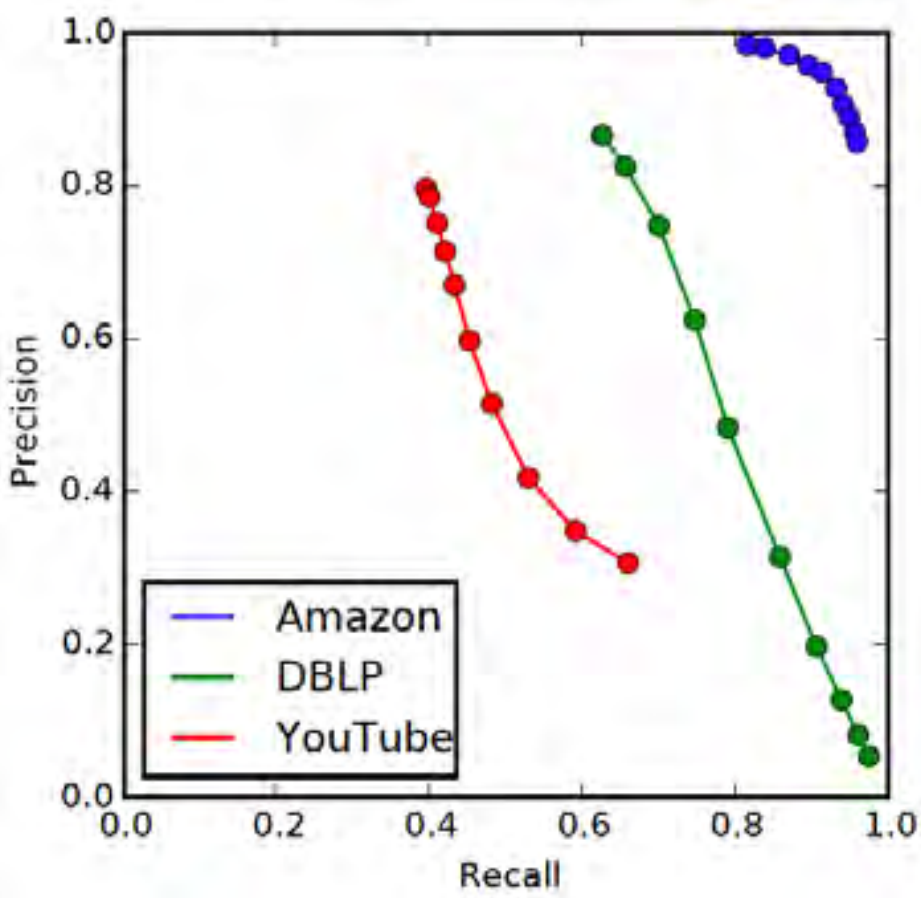




Community size vs. precision (a), and recall (b) for the Amazon graph (top 5000 groundtruth communities)

As groundtruth community size increases, getting good recall results becomes increasingly hard.

Experimental Results – TECTONIC's parameter θ



Precision vs. recall as parameter θ takes values in $\{0.01, 0.02, 0.03, .., 0.1\}$ (behavior monotonic)