Motif-driven Dense Subgraph Discovery in Directed and Labeled Networks

A. Erdem Sarıyüce

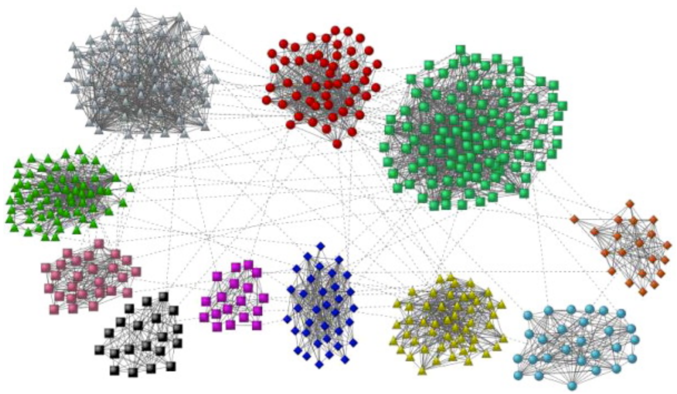
Assistant Professor

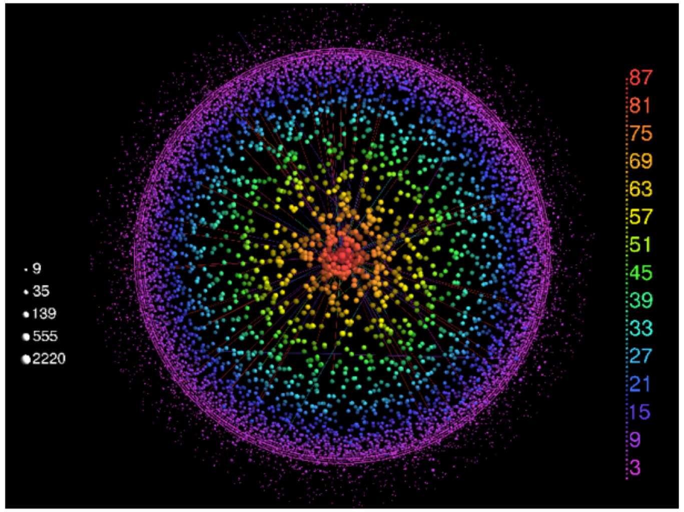
University at Buffalo The State University of New York

THE WEB CONFERENCE

Dense subgraph discovery

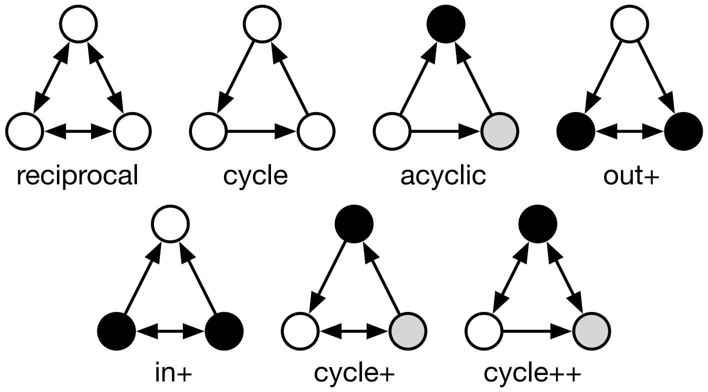
- Dense regions are unusual and interesting
  - Anomaly detection, community detection, visualization
- A good proxy for graph clustering
  - Exhibit good cuts [Gleich and C. Seshadhri, 2012]
- Literature is rich for simple, undirected networks
- What about heterogeneous networks?
  - Directed edges
  - Labeled nodes/edges
    - Categorical
    - Numerical
  - How to even define the density?



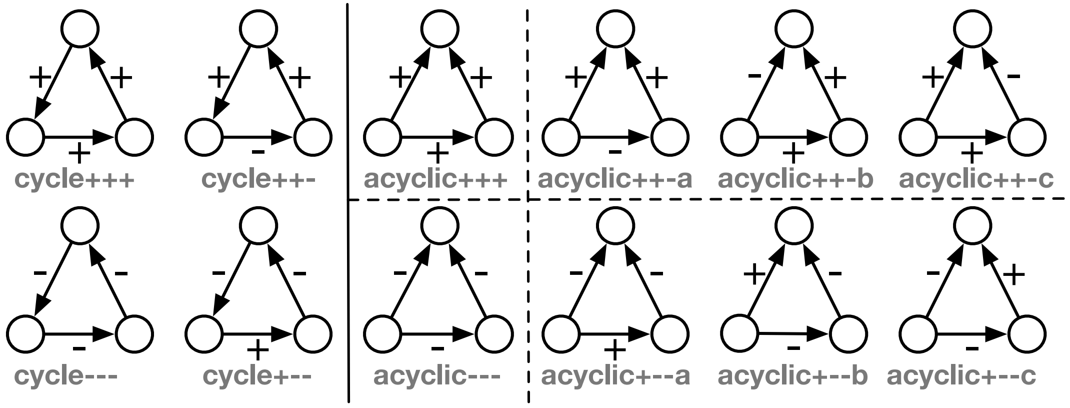


Motifs for help

- Fundamental building blocks in the organization and dynamics of real-world networks
- Captures higher-order relationships among multiple nodes
- Density is the avg. motif degree
  - Number-of-motifs / number-of-nodes



directed triangles



signed-directed triangles

- Extendible for heterogeneous networks
  - Pros: Customizable; dense subgraphs w.r.t. motif of interest
  - Cons: Spectrum is wide; hard to unify all in a framework

Idea: Participations of small motifs in larger motifs

- Given a pair of motifs $\boldsymbol{M}$ and $\boldsymbol{N}$ s.t. $\boldsymbol{M} \subset \boldsymbol{N}$, find the subgraphs where each $\boldsymbol{M}$ participates in many $\boldsymbol{N}$s
  - Inspired by core and truss decompositions

- $\boldsymbol{M}$ and $\boldsymbol{N}$ can have directed edges and categorical labels on nodes/edges
  - No numerical labels – future work

- Motif hypergraph:
  - $\boldsymbol{M}$s are the nodes
  - $\boldsymbol{N}$s are the hyperedges
  - An $\boldsymbol{M}$ is connected to an $\boldsymbol{N}$ iff $\boldsymbol{M} \subset \boldsymbol{N}$

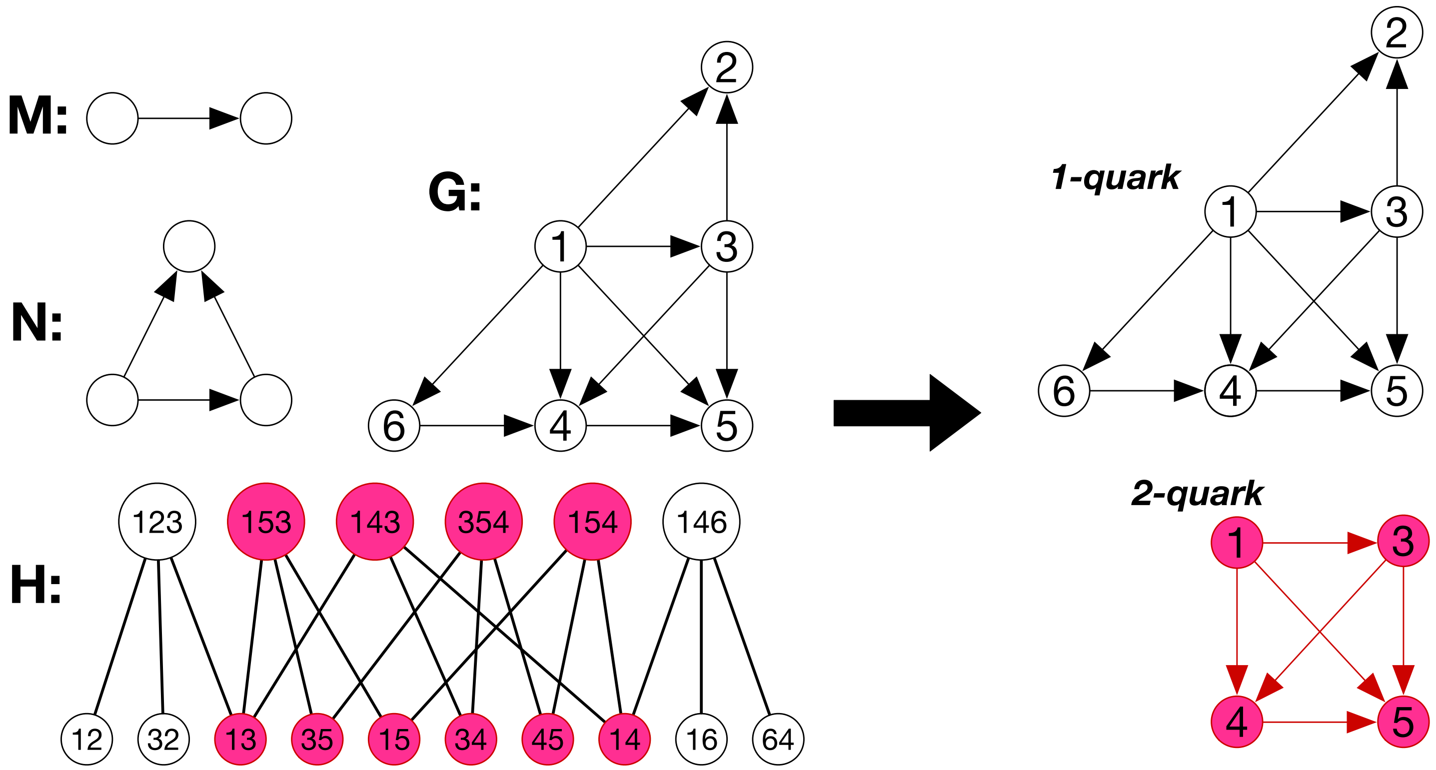
- Motif of interest is $\boldsymbol{N}$

Quark decomposition

- Given a graph $\boldsymbol{G}$ and motifs $\boldsymbol{M},\ \boldsymbol{N}\ (\boldsymbol{M} \subset \boldsymbol{N})$, let $\boldsymbol{H}$ be motif hypergraph,
  - A $\boldsymbol{k}$ -quark is a connected and maximal sub-hypergraph where each $\boldsymbol{M}$ instance participates in at least $\boldsymbol{k}$ number of $\boldsymbol{N}$ instances.
  - Quark number of an $\boldsymbol{M}$ is the largest value of $\boldsymbol{k}$ s.t. $\boldsymbol{M}$ belongs to a $\boldsymbol{k}$-quark.

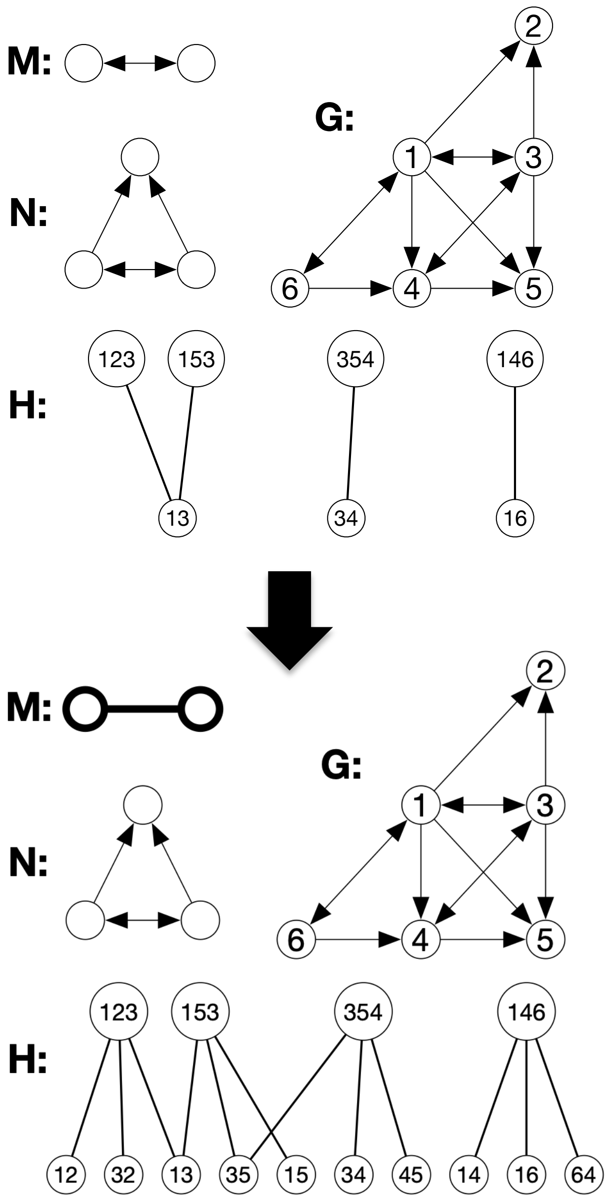
Quark decomposition

- Given a graph $\boldsymbol{G}$ and motifs $\boldsymbol{M}, \ \boldsymbol{N} \ (\boldsymbol{M} \subset \boldsymbol{N})$, let $\boldsymbol{H}$ be motif hypergraph,
  - A $\boldsymbol{k}$ -quark is a connected and maximal sub-hypergraph where each $\boldsymbol{M}$ instance participates in at least $\boldsymbol{k}$ number of $\boldsymbol{N}$ instances.
  - Quark number of an $\boldsymbol{M}$ is the largest value of $\boldsymbol{k}$ s.t. $\boldsymbol{M}$ belongs to a $\boldsymbol{k}$-quark.



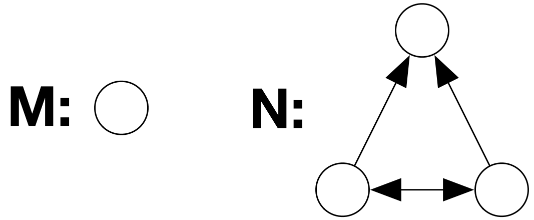
Limitations and practical instantiations

- What if there is only one M in N?
  - Size of each N in the motif hypergraph becomes one!
  - How to avoid?
- Consider M as vanilla
  - Labelless nodes/edges, directionless edges
- M is better to be an edge (or larger)
  - Overlapping subgraphs!



Role confusion problem

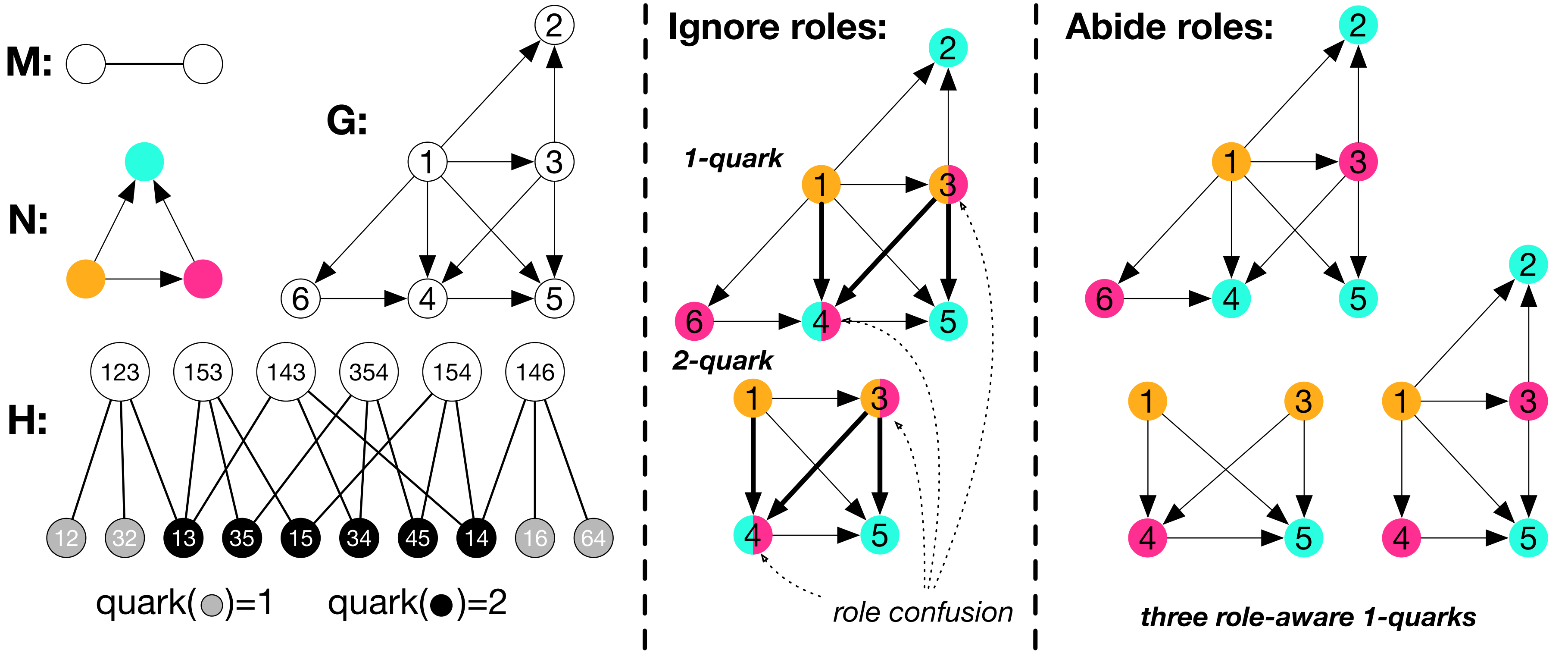
- What if $\boldsymbol{M}$ has different "roles" in $\boldsymbol{N}$s it's part of?
  - Orbits! [Pržulj, 2007]



- How to distinguish the participations where $\boldsymbol{M}$ is in different orbits?
  - Orbit degrees: Number of $\boldsymbol{N}$s that contain $\boldsymbol{M}$ s.t. $\boldsymbol{M}$ is in a specific orbit

- Role-aware $\boldsymbol{k}$-quark: $\boldsymbol{M}$'s orbit is the same in all the participations.
  - I.e., orbit degree of each $\boldsymbol{M}$ is at least $\boldsymbol{k}$

Role confusion problem



- Role-aware $\boldsymbol{k}$-quark: $\boldsymbol{M}$'s orbit is the same in all the participations.
  - I.e., orbit degree of each $\boldsymbol{M}$ is at least $\boldsymbol{k}$

Peeling algorithm works for quark decomposition!

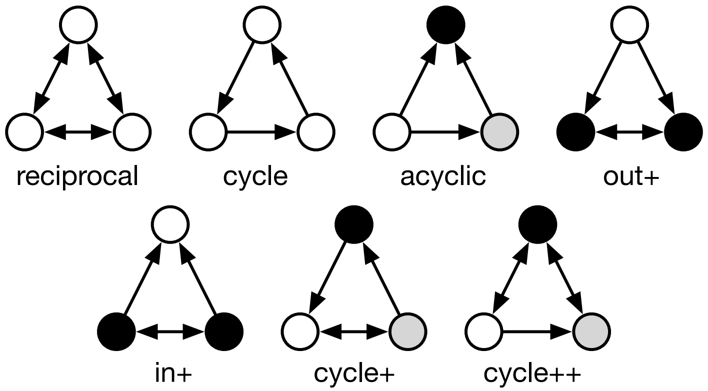
- Both quark and role-aware quark decompositions
- Subgraph and hierarchy construction included
- When $\boldsymbol{M}$ is a node or edge, time complexity is

$$O(\sum_{v \in V} d(v)^{|V_N|-1})$$

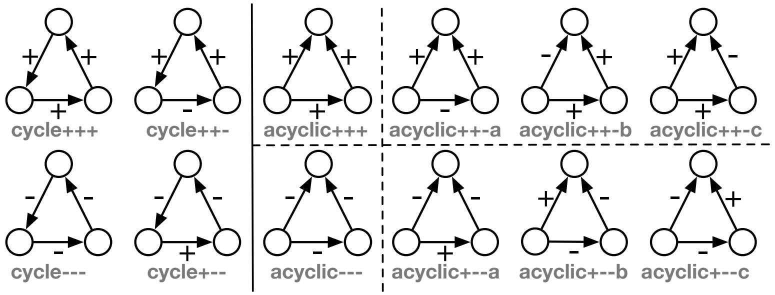
- Existing optimizations for peeling algorithms are applicable
  - Constructing subgraphs during the peeling
  - Parallel, local computations

Experimental evaluation on heterogeneous networks

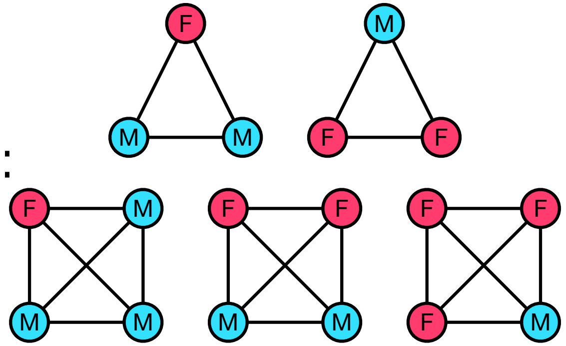
- Directed
  - $\boldsymbol{M}$ is edge
  - $\boldsymbol{N}$ is a triangle:



- Signed-directed
  - $\boldsymbol{M}$ is edge
  - $\boldsymbol{N}$ is a triangle:



- Node-labeled (genders)
  - $\boldsymbol{M}$ is edge or triangle
  - $\boldsymbol{N}$ is triangle or four-clique:

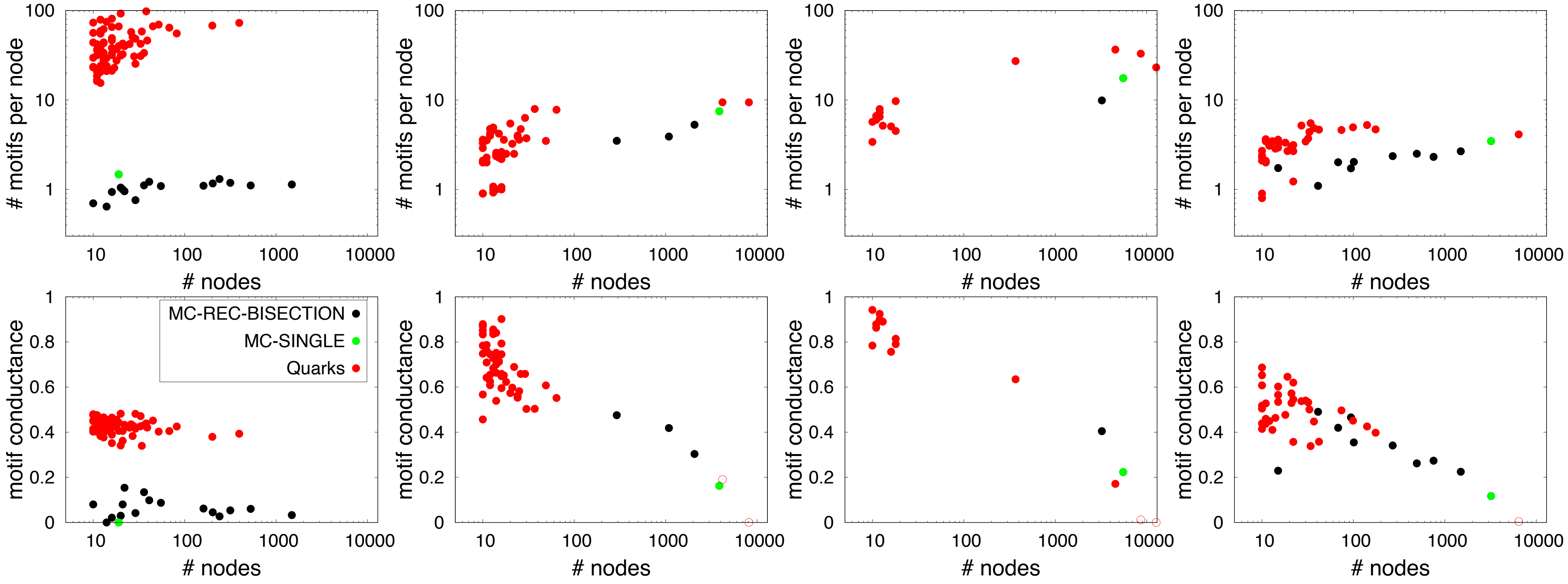


- Baselines:
  - Motif clustering
    - [Benson et al., 2016]
  - Cycle-truss and flow-truss
    - [Takaguchi and Yoshida, 2016]
  - Nucleus decomposition
    - [Sariyuce et al., 2015]

- Metrics
  - Motif conductance
  - Avg. motif degree
  - Edge density
    - For node-labeled

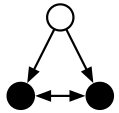
Quark decomposition vs. Motif clustering

- Motif clustering optimizes motif conductance, thus better
- Quark decomposition gives higher avg. motif degrees
- Motif clusters are big due to partitioning, quarks are smaller thanks to bottom-up dec.

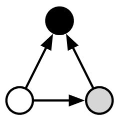


(a) *reciprocal* **(b)** *out+* **(c)** *acyclic* **(d)** *cycle++*

Food-web analysis

- Analysis with out+ 
- Quarks give consistently better classifications than motif clustering

out+	Metric	Quarks (7 subgraphs)	MC-K-MEANS w/ 4 clusters	MC-K-MEANS w/ 7 clusters
Class 1	ARI	**0.3627**	0.3005	0.1485
	F1	**0.4869**	0.4574	0.3794
	NMI	**0.5415**	0.5040	0.4843
	Purity	**0.5968**	0.5645	0.5161
Class 2	ARI	**0.3816**	0.3265	0.1871
	F1	**0.5675**	0.5380	0.4601
	NMI	**0.5206**	0.4822	0.4309
	Purity	**0.6452**	0.6129	0.5645

- Role-aware quark numbers find the preys, predators, and balancers with acyclic 
  - Predators: Birds (ducks, herons, greeb)
  - Preys: Clown goby, herbivorous shrimps, zooplankton
  - Balancer: Fishes (anchovy, sardines, mojarra)

Word-associations

- Diverse subgraphs obtained with different motifs
  - Not possible when directions ignored

direction-oblivious subgraph by (2,3) nucleus

astronomy cosmos earth moon planet planetarium planets sky solar-system space star stars sun universe

in+

god jupiter mars moon planets saturn space star stars uranus venus

darkness end endless eternal eternity ever everlasting finite for ever forever god infinite infinity lasting long love never perpetual space star stars universe

abroad away holiday holidays home sand spain sun sunshine vacation

cycle++

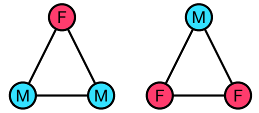
aeroplane air aircraft airport astronaut cosmos earth flight fly flying holst jet jupiter mars moon noise pilot plane planet planets rocket saturn sky solar-system space sphere sputnik star stars universe wing wings world

out+

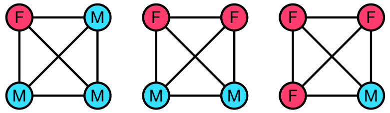
aeroplane air air-force aircraft flier fly glide kite parachute plane sky soar wing

Finding gender-balanced subgraphs

- Facebook100 dataset with genders as node-labels
- How to find gender-balanced dense subgraphs even when the graph is imbalanced?
  - Compared to label-oblivious nucleus dec.
- $\boldsymbol{M}$ is edge, $\boldsymbol{N}$ is triangle

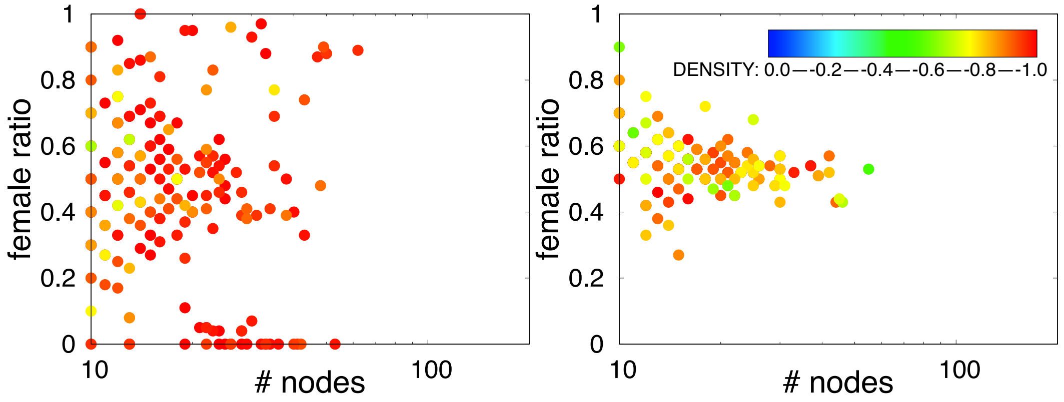


- $\boldsymbol{M}$ is triangle, $\boldsymbol{N}$ is four-clique



	$\lvert V \rvert$	$\lvert E \rvert$	$\frac{\lvert V_f \rvert}{\lvert V \rvert}$	edge, triangle (2,3)n	edge, triangle Quarks FMM	edge, triangle Quarks FFM	triangle, 4-clique (3,4)n	triangle, 4-clique Quarks FMMM	triangle, 4-clique Quarks FFMM	triangle, 4-clique Quarks FFFM
Mich67	3.7K	81.9K	25%	23.0%	45.0%	50.0%	24.5%	40.0%	45.0%	51.6%
Caltech36	769	16.7K	30%	39.4%	46.0%	52.0%	38.5%	43.1%	50.2%	52.8%
Carnegie49	6.6K	250.0K	37%	32.6%	49.0%	52.5%	38.5%	43.5%	49.5%	54.9%
MIT8	6.4K	251.3K	37%	38.8%	48.0%	52.1%	42.0%	44.3%	50.3%	53.9%
Stanford3	11.6K	568.3K	40%	46.8%	48.1%	49.0%	44.1%	45.4%	49.2%	55.4%
Cornell5	18.7K	790.8K	44%	44.3%	47.6%	51.8%	45.6%	43.7%	48.7%	54.9%
Penn94	41.6K	1.4M	44%	49.7%	48.4%	51.4%	52.1%	44.0%	49.8%	55.8%
UPenn7	14.9K	686.5K	44%	37.3%	48.8%	51.1%	46.4%	45.1%	50.4%	55.4%
Average of 18 networks:			$\boxed{40\%}$	42.5%	48.2%	$\boxed{51.5\%}$	44.1%	44.4%	49.7%	$\boxed{54.7\%}$

Female ratios



(a) (3, 4) **nuclei** **(b) Quarks by** FFFM

Density vs. female ratio for UPenn7

Conclusion & Future Work

- Principled approach for motif-driven dense subgraph discovery in directed and categorical-labeled networks
  - Successfully regularizes the motif degrees to quark numbers
- Role-aware variant considers the orbits and quantifies the roles systematically
- Versatile, efficient, and extendible
  - Code is available with detailed instructions for reproducibility!

- Hierarchy structure had limited success
  - Further analysis of hierarchy w.r.t a given motif
- Extension for networks with numerical node/edge labels
  - While incorporating the ordering

Paper, slides, talk, code: http://sariyuce.com/WWW21

Questions: erdem@buffalo.edu

Thanks!

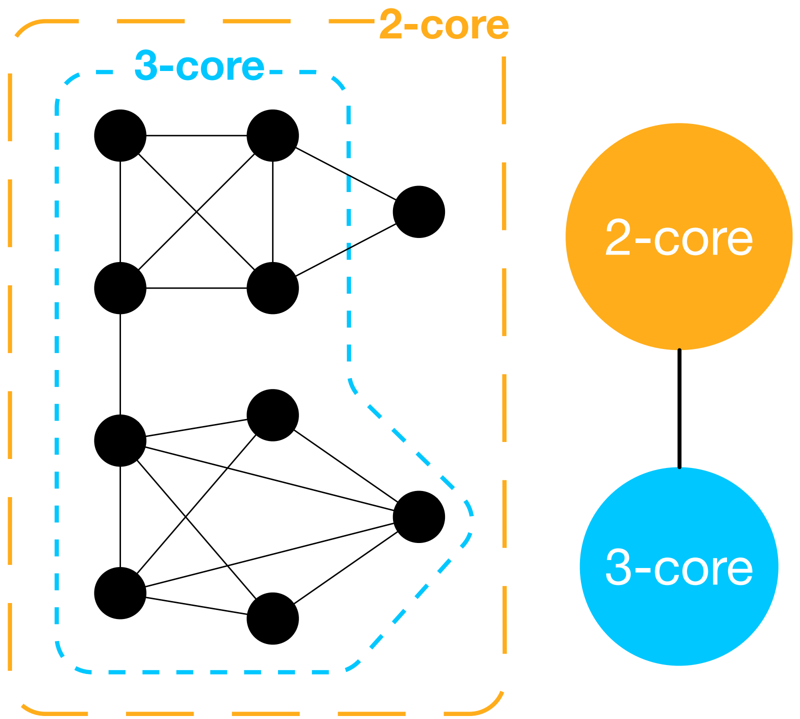
THE WEB CONFERENCE

University at Buffalo The State University of New York

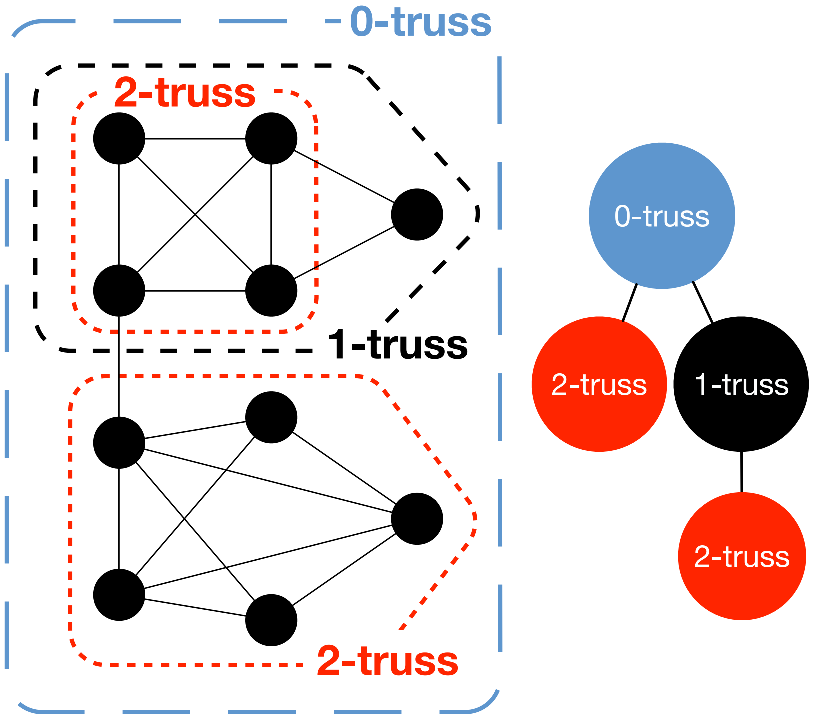
How to model dense subgraphs?

- Two effective models for simple, undirected networks
  - With hierarchical relations

- ***k*-core:** Every vertex has at least k edges
  - [Seidman ‘83], [Matula & Beck ‘83]

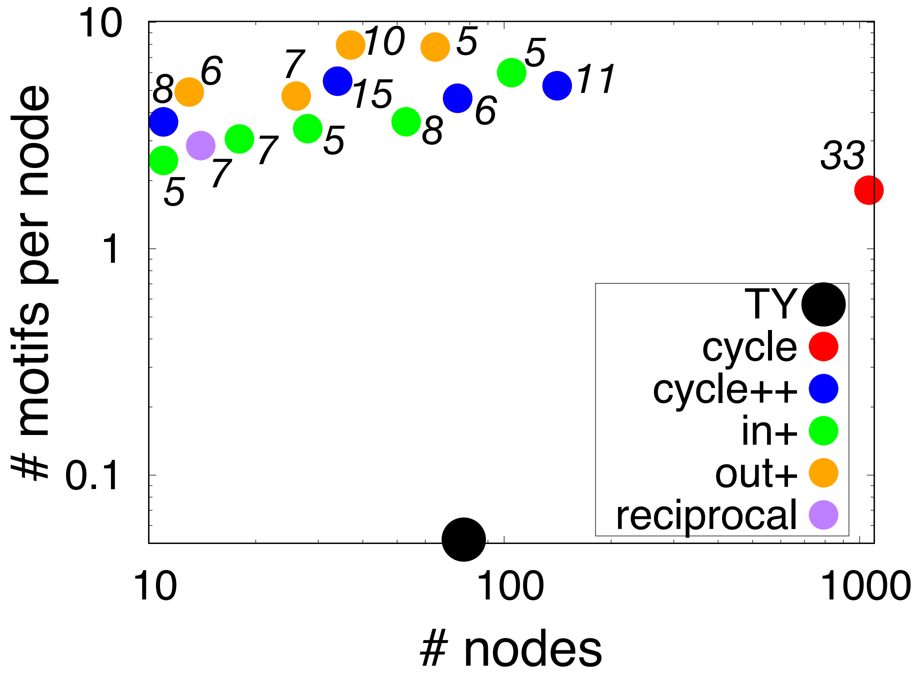


- ***k*-truss:** Every edge has at least k triangles
  - [Cohen ‘08]

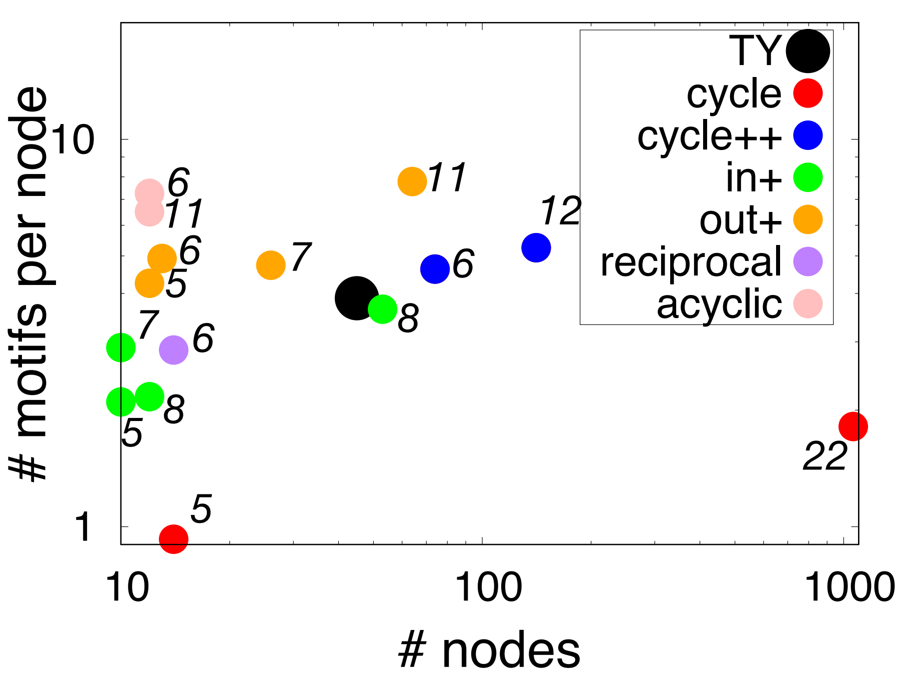


Quarks vs. Cycle- & Flow-truss

- Higher avg. motif degrees with quarks
- Almost all the nodes in cycle- & flow-trusses are found, in various types
  - Considering each bidirectional edge atomically (instead of two unidirectional edges) highlights the diversity



(a) cycle-truss vs. quarks



(b) flow-truss vs. quarks

Runtime comparison with motif clustering

- Motif clustering with a single optimal cluster
  - Quark decomposition finds all the $\boldsymbol{k}$-quarks
- Quark decomposition is mostly faster, for all motifs; up to 10x speedups
- Motif clustering is mostly faster for en-Wikipedia and wiki-Talk
  - Spectral clustering is heavy, cost increases when multiple clusters found

	cycle		acyclic		*out+*		*in+*		*cycle+*		*cycle++*	
	Q	M	Q	M	Q	M	Q	M	Q	M	Q	M
web-ND	**0.34**	3.31	**4.26**	16.8	**0.62**	6.3	**2.11**	8.54	**0.53**	10.01	**0.78**	9.86
amzn	**0.74**	3.54	**3.29**	79	**2.25**	132	**1.92**	105	**1.18**	5.29	**3.23**	107
wiki	28.9	**14.0**	112	**18.2**	**10.9**	16.4	21.1	**17.7**	20.5	**20.2**	47.8	**16.8**
soc-p	**23.6**	79	**66.9**	99	**37.0**	119	**34.2**	139	**48.9**	129	**98.1**	128
live-j	**37.4**	200	**180**	943	**118**	1135	**126**	1438	**112**	828	**289**	2248
en-w	900	**501**	7746	**864**	1511	**799**	1709	**677**	**398**	724	2223	**677**