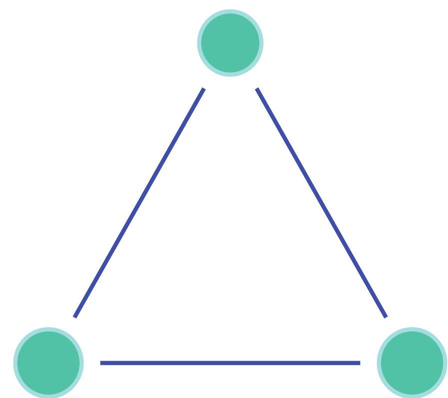
Triadic data analysis in temporal and higher-order networks

Austin Benson · Cornell University

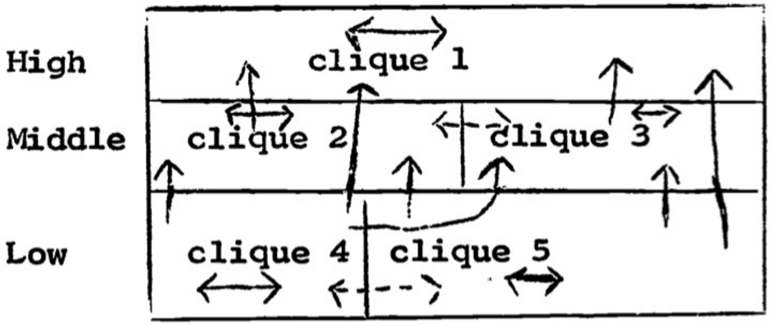
DynaMo@Networks 2021

The humble triangle is fundamental in network science.

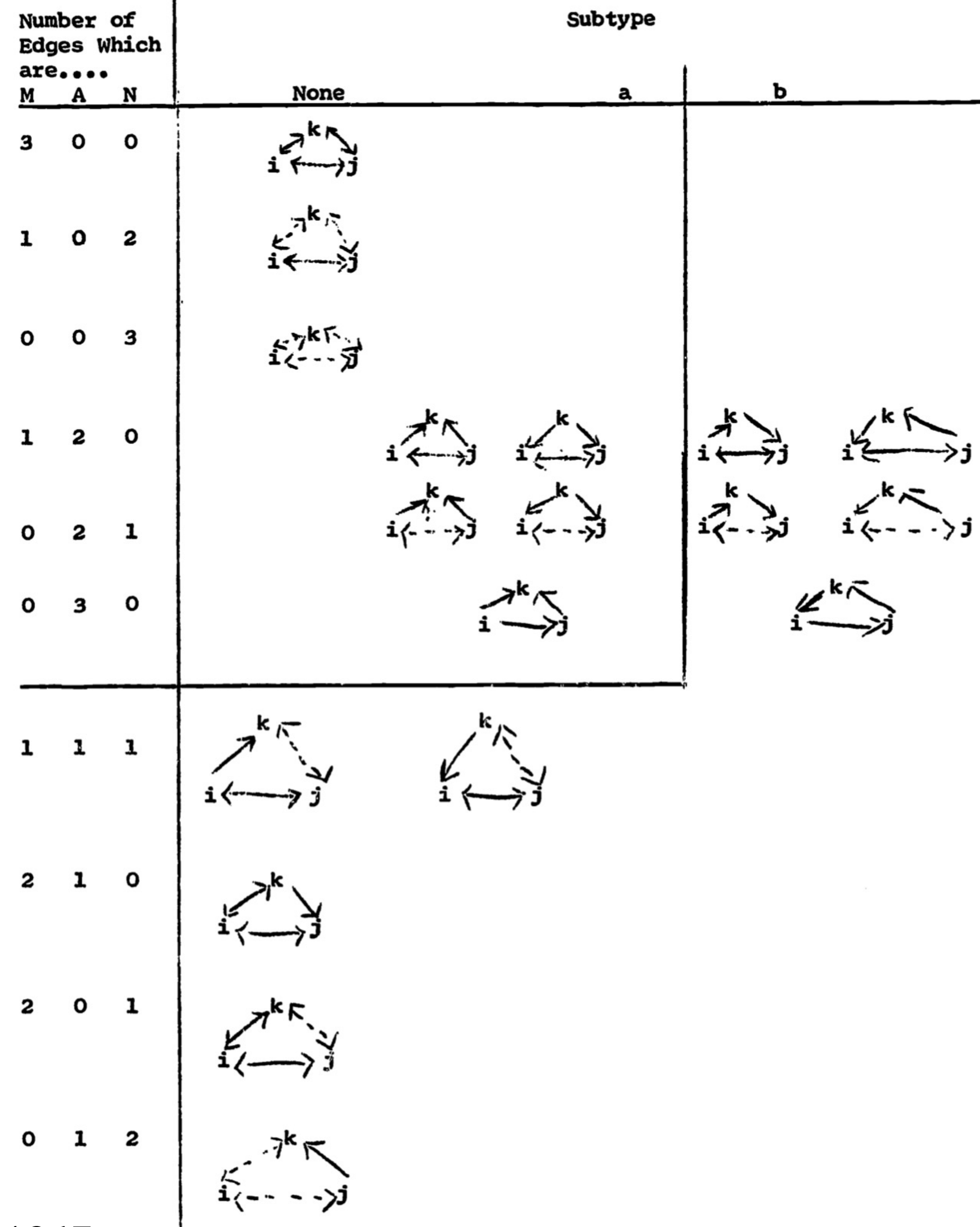


The Strength of Weak Ties, Granovetter, 1973.
Collective dynamics of 'small-world' networks. Watts & Strogatz, 1998.

Levels, Cliques and Relations



Triad			Year I Week 15			Year II Week 15		
M	A	N	expected	observed	difference	expected	observed	difference
2	1	0	75.4	68	--7.4	72.7	66	- 6.7
0	1	2	75.4	76	+ 0.6	72.7	62	-10.7
1	1	1	150.7	163	+12.3	145.3	157	+11.7
2	0	1	69.1	54	-15.1	51.9	25	-26.9
1	2	0b	41.1	39	- 2.1	50.9	27	-23.9
0	2	1b	41.1	37	- 4.1	50.9	31	-19.9
0	3	0b	7.5	3	- 4.5	11.9	4	- 7.9
Total			460.3	440	-20.3	456.3	372	-84.3



The Structure of Positive Interpersonal Relations in Small Groups, 1967.
James Davis and Samuel Leinhardt analyzing *triangles* to test a sociological theory of George Homans using data from Theodore Newcomb.

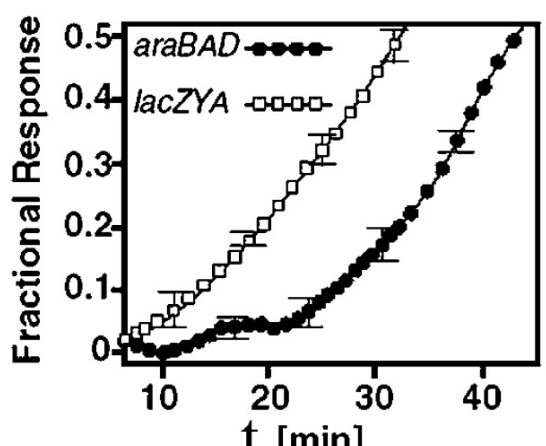
Network	Nodes	Edges	N_{real}	$N_{rand} \pm$ SD	Z score
Gene regulation (transcription)				**X ↓ Y ↓ Z**	**Feed-forward loop**
E. coli	424	519	40	7 ± 3	10
*S. cerevisiae**	685	1,052	70	11 ± 4	14

Network Motifs: Simple Building Blocks of Complex Networks, Milo et al., 2002.

Table 1. Structure and function of the coherent FFL types, with AND- and OR- gates at the Z promoter

	Coherent type 1		Coherent type 2		Coherent type 3		Coherent type 4	
Species	Structure	Abundance	Structure	Abundance	Structure	Abundance	Structure	Abundance
E. coli	X ↓ Y ↓ Z	28	X ⊥ Y ↓ Z	2	Y ↓ Y ⊥ Z	4	X ⊥ Y ⊥ Z	1
S. cerevisiae		26		5		0		0
Z Logic→	AND	OR	AND	OR	AND	OR	AND	OR
Steady-state Z(Sx,Sy)	$S_x \wedge S_y$	S_x	$\bar{S}_x \wedge S_y$	$\bar{S}_x$	$\bar{S}_x$	$\bar{S}_x \wedge \bar{S}_y$	S_x	$S_x \vee \bar{S}_y$
Response delay								
Sx on step	Delay	—	—	Delay	—	—	Delay	Delay
Sx off step	—	Delay	Delay	—	Delay	Delay	—	—
Inverted out	No	No	Yes	Yes	Yes	Yes	No	No

Structure and function of the feed-forward loop network motif, Mangan & Alon, 2003.

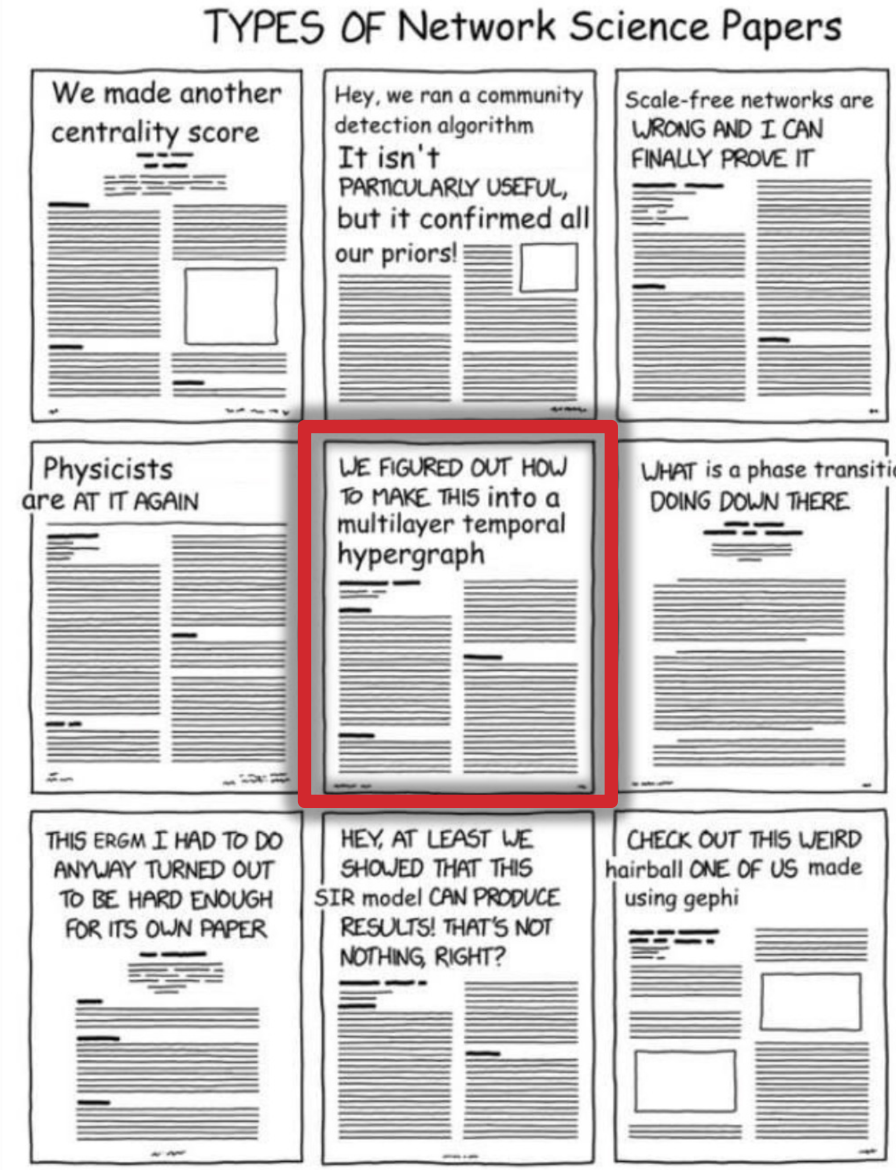


The Coherent Feedforward Loop Serves as a Sign-sensitive Delay Element in Transcription Networks, Mangan, Zaslaver, & Alon, 2003.

Modern network data is rich…

- Higher-order / multi-way interactions
- Temporal information
- Multilayer, multiplex, heterogeneous, attributed
- Features / covariates
- Large-scale with millions or billions of edges

Triangles are super useful for this rich data!



@zhangqian_rach

Triadic analysis for modern network data.

w/ R Abebe, M Schaub, J Kleinberg, A Jadbabaie

1. Open and closed triangles in temporal, higher-order interactions.
   Simplicial closure and higher-order link prediction, PNAS, 2018.

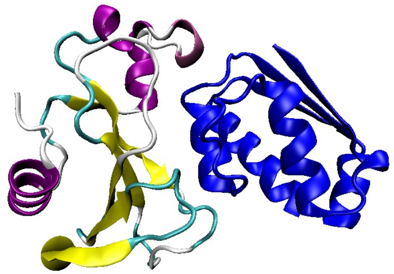
2. Triadic motifs in temporal networks.
   Motifs in temporal networks, WSDM 2017.
   Sampling methods for counting temporal motifs, WSDM, 2019.

Real-world systems are composed of “higher-order” interactions that we often reduce to pairwise ones.

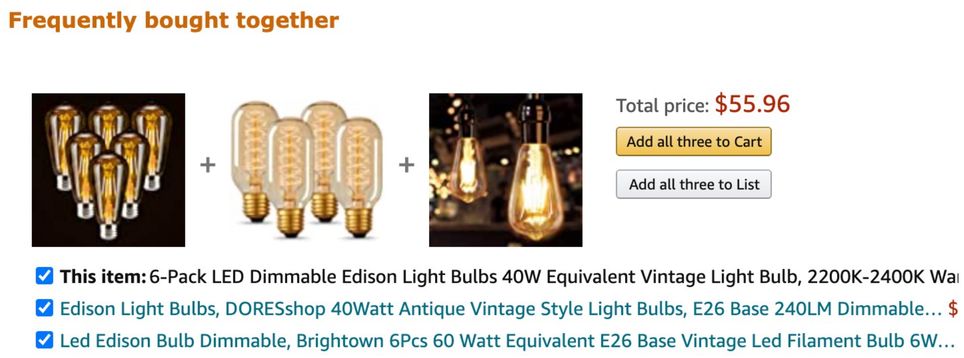
Communications
nodes are people/accounts
emails often have several recipients, not just one.

Physical proximity
nodes are people
people gather in groups

Commerce
nodes are products
several products can be purchased at once



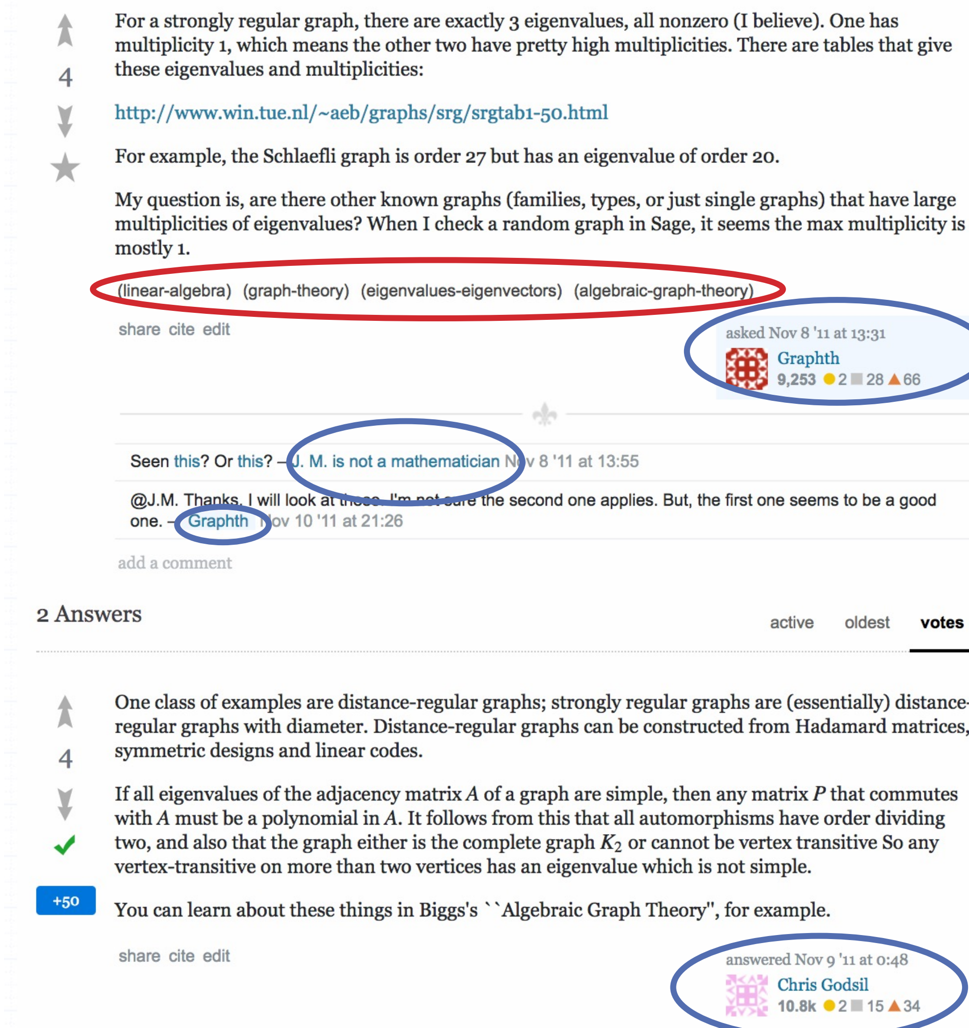
Cell biology
nodes are proteins
protein complexes may involve several proteins



We collected many datasets of timestamped hyperedges

bit.ly/sc-holp-data

1. Coauthorship in different domains.
2. Emails with multiple recipients.
3. Tags on Q&A forums.
4. Threads on Q&A forums.
5. Contact/proximity measurements.
6. Musical artist collaboration.
7. Substance makeup and classification codes applied to drugs the FDA examines.
8. U.S. Congress committee memberships and bill sponsorship.
9. Combinations of drugs seen in patients in ER visits.



https://math.stackexchange.com/q/80181

Thinking of higher-order data as a weighted projected graph with filled-in structures is a convenient viewpoint.

Data.

$t_1 : \{1, 2, 3, 4\}$

$t_2 : \{1, 3, 5\}$

$t_3 : \{1, 6\}$

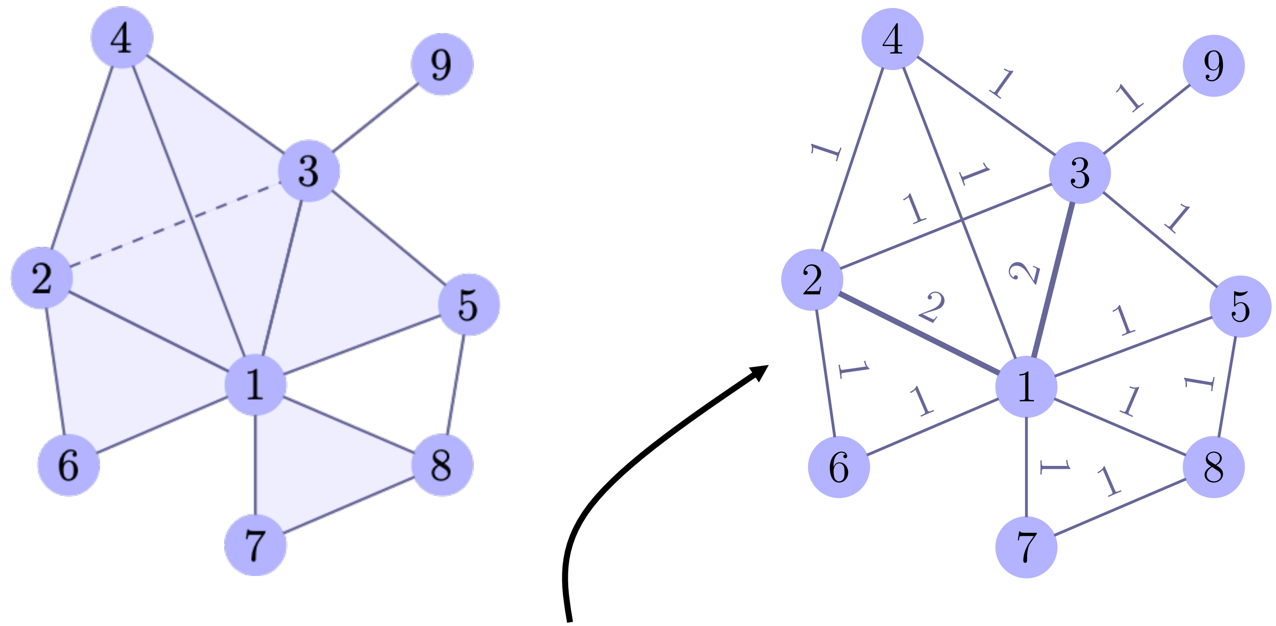
$t_4 : \{2, 6\}$

$t_5 : \{1, 7, 8\}$

$t_6 : \{3, 9\}$

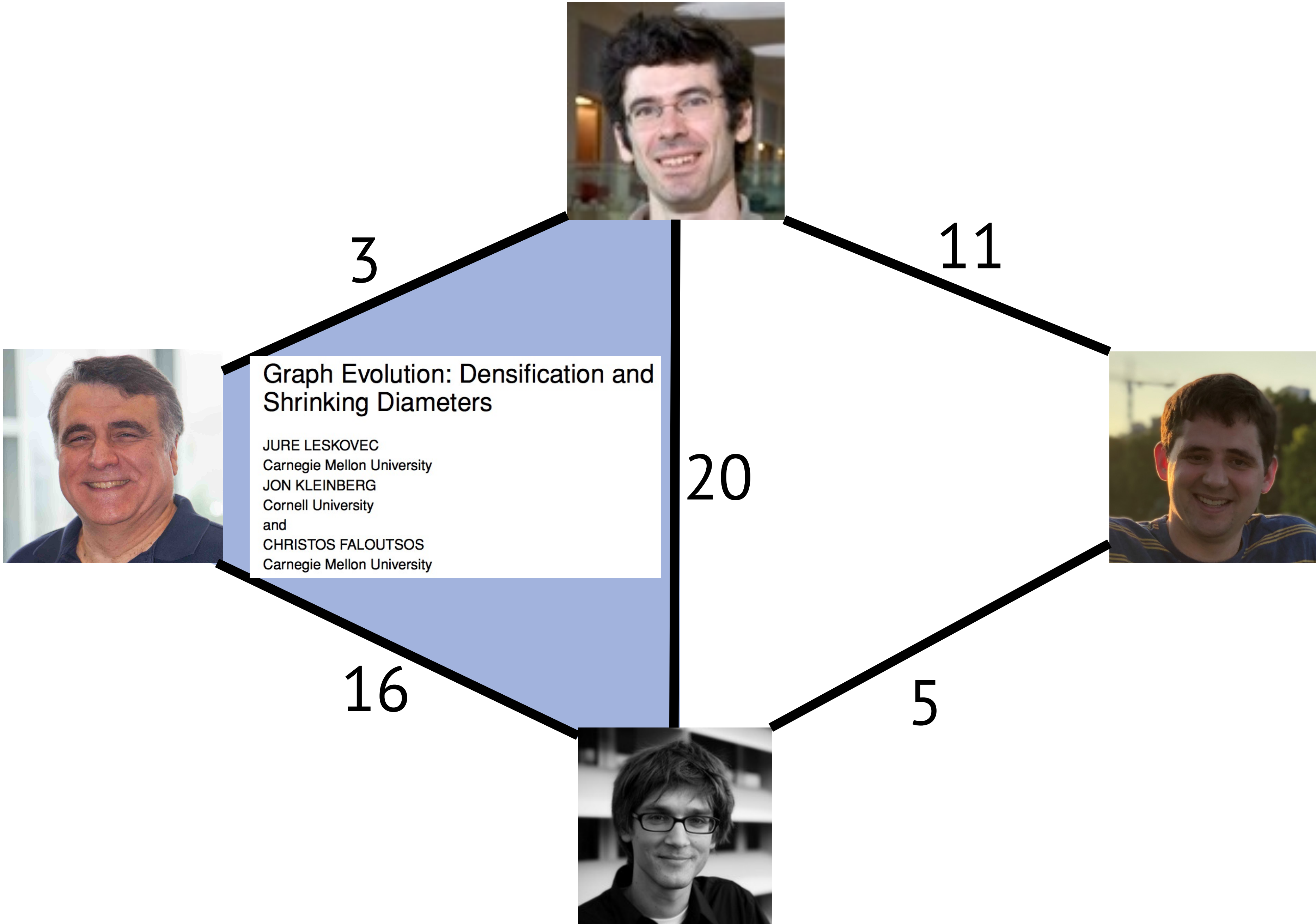
$t_7 : \{5, 8\}$

$t_8 : \{1, 2, 6\}$

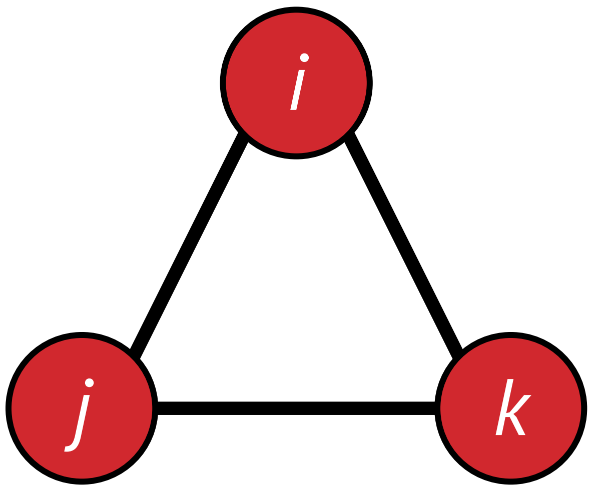


Projected graph ***W***.

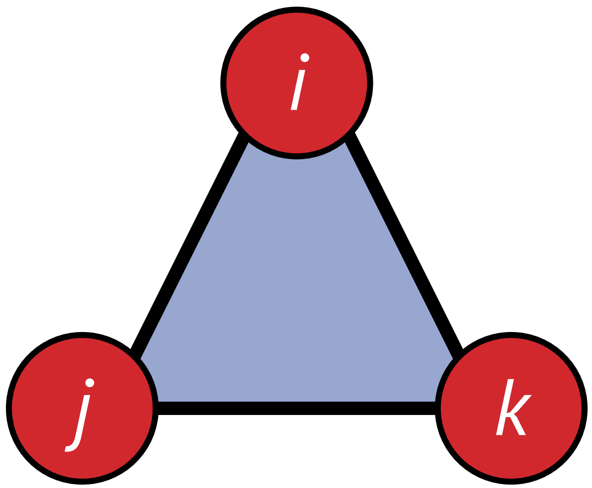
W_{ij} = # of hyperedges containing nodes *i* and *j*.


3
11
Graph Evolution: Densification and Shrinking Diameters
JURE LESKOVEC
Carnegie Mellon University
JON KLEINBERG
Cornell University
and
CHRISTOS FALOUTSOS
Carnegie Mellon University
20
16
5

What's more common in empirical data?



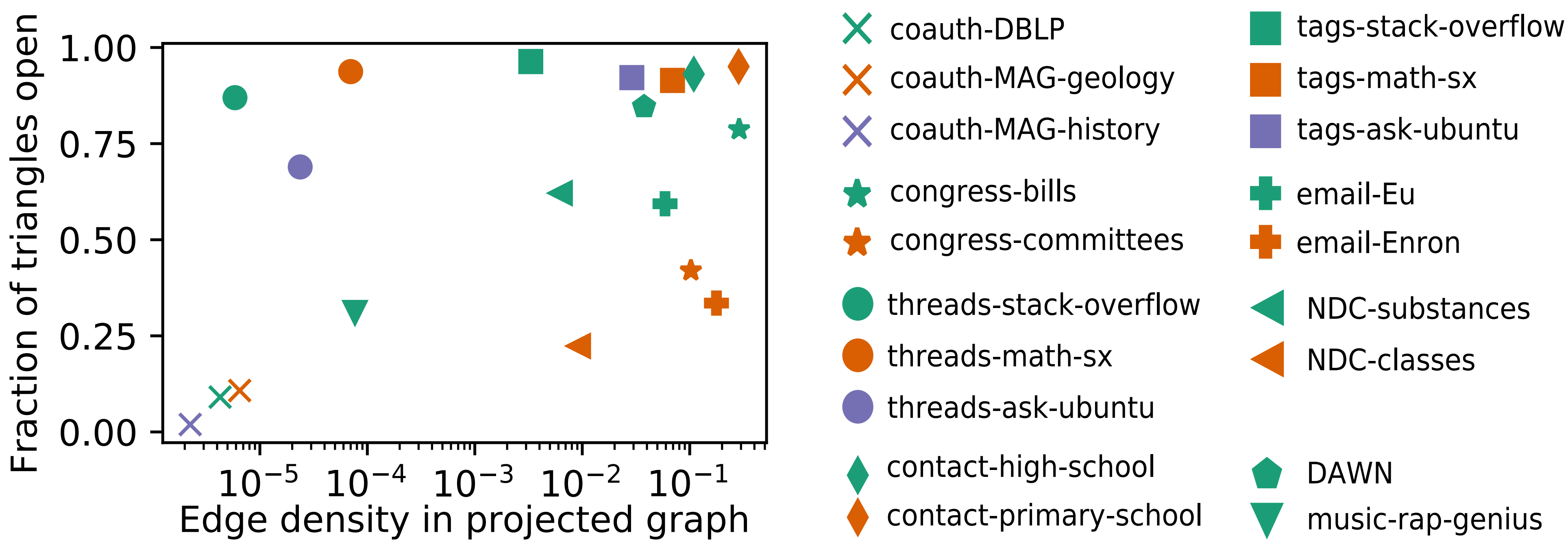
or



Open triangle
each pair has been in a hyperedge together but all 3 nodes have never been in the same hyperedge

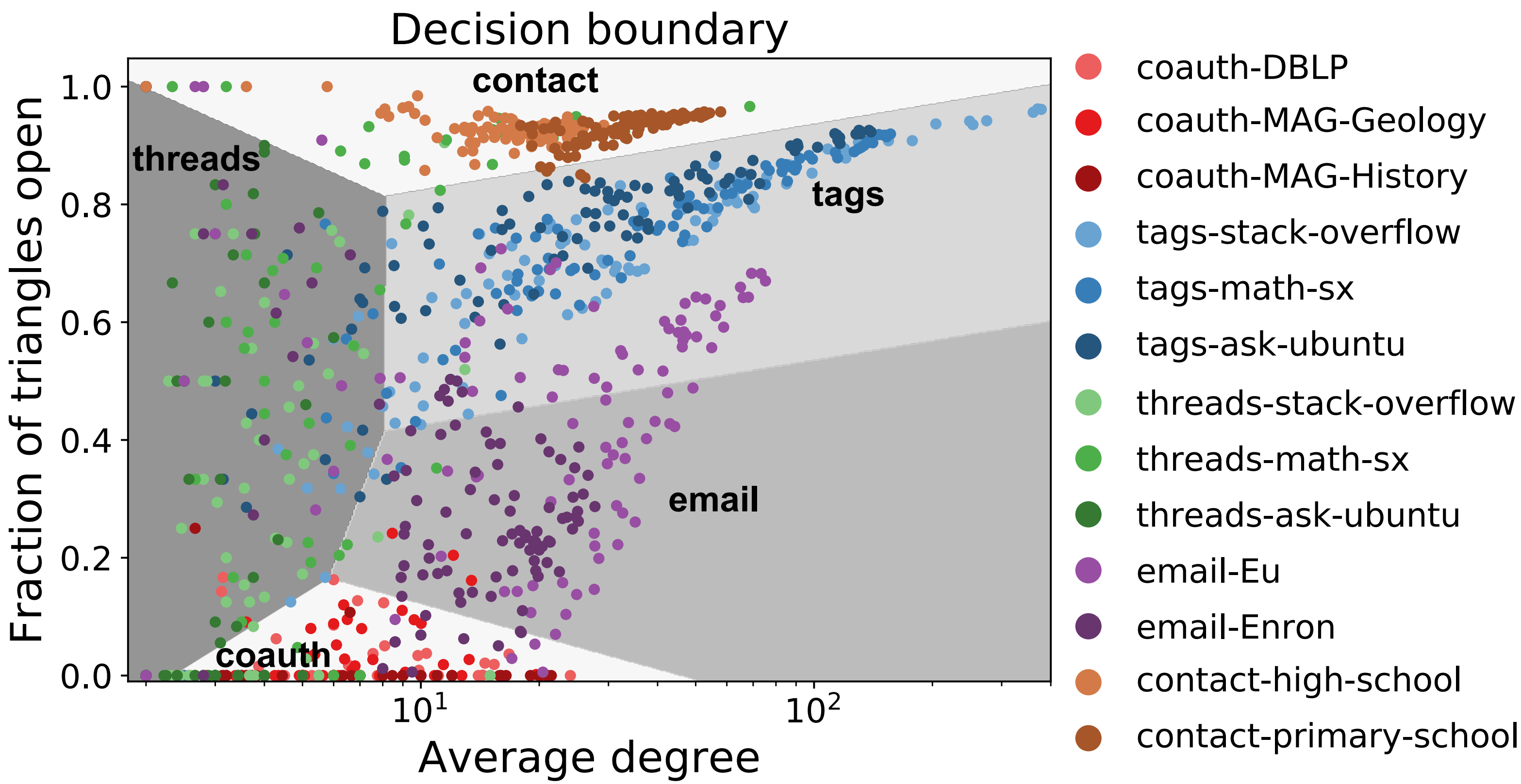
Closed triangle
there is some hyperedge that contains all 3 nodes

There is lots of variation in the fraction of triangles that are open, but datasets from the same domain are similar.



See also *Topological analysis of data* by Patania, Vaccarino, & Petri, 2017.

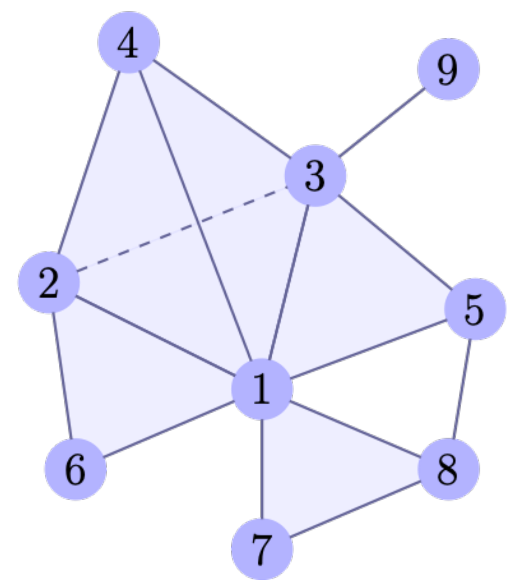
Dataset domain separation also occurs at the local level.

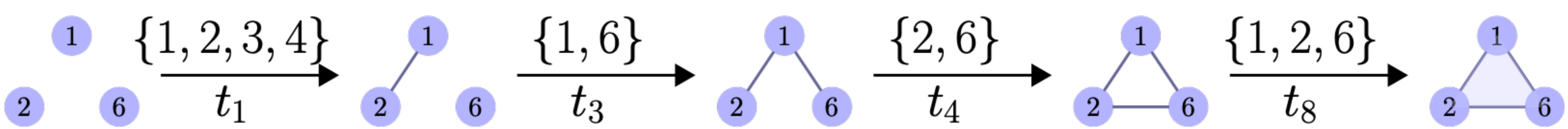


- Randomly sample 100 egonets per dataset and measure log of average degree and fraction of open triangles.
- Logistic regression model to predict domain (coauthorship, tags, threads, email, contact).
- 75% model accuracy vs. 21% with random guessing.

Triangles close over time.

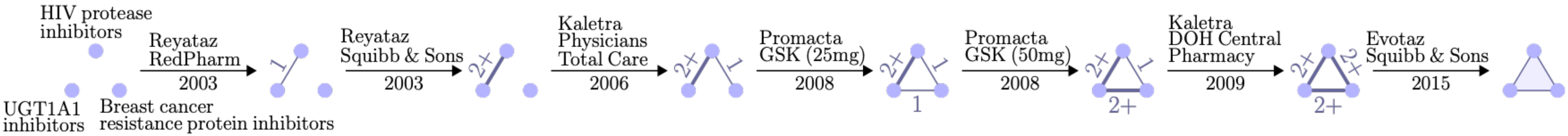
$t_1 : \{1, 2, 3, 4\}$
$t_2 : \{1, 3, 5\}$
$t_3 : \{1, 6\}$
$t_4 : \{2, 6\}$
$t_5 : \{1, 7, 8\}$
$t_6 : \{3, 9\}$
$t_7 : \{5, 8\}$
$t_8 : \{1, 2, 6\}$





Weak and strong ties are useful characterizations.

Substances in marketed drugs recorded in the National Drug Code directory.

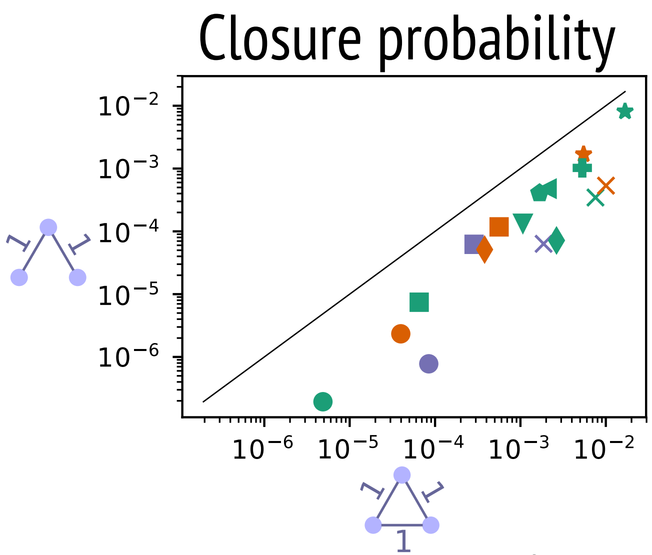


Bin weighted edges into "weak" and "strong ties" in the projected graph ***W***.
W_{ij} = # of simplices containing nodes *i* and *j*.

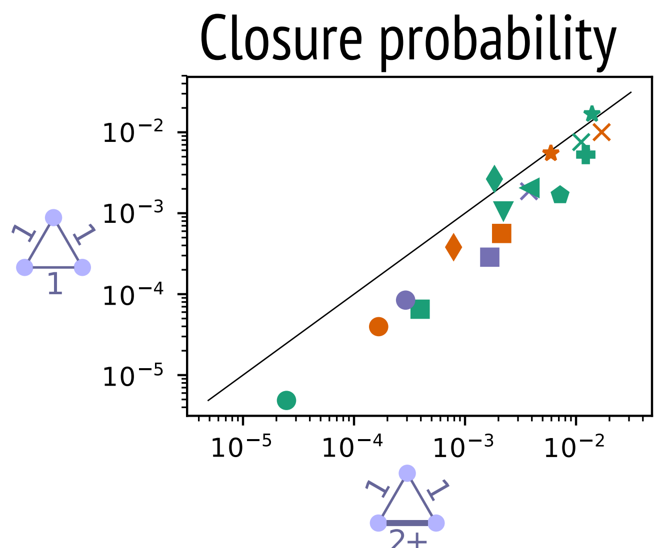
- **Weak ties.** $W_{ij} = 1$ (one hyperedge contains *i* and *j*)
- **Strong ties.** $W_{ij} \geq 2$ (at least hyperedges contain *i* and *j*)

Closure depends on structure in projected graph.

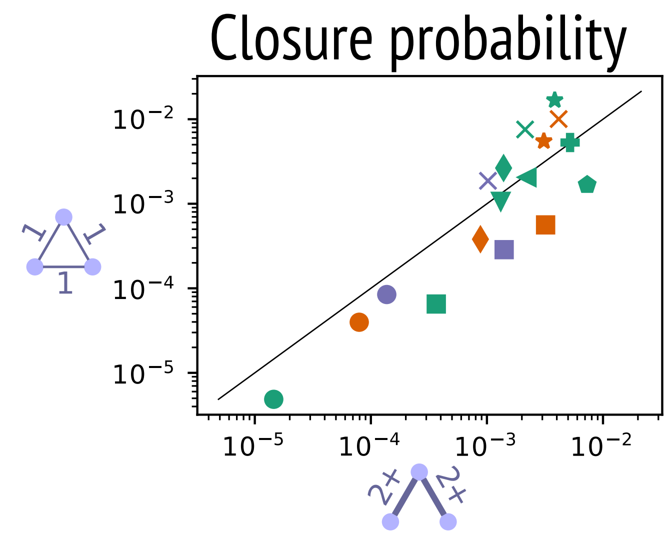
- First 80% of the data (in time) ⟶ record configurations of triplets not in closed triangle.
- Remainder of data ⟶ find fraction that are now closed triangles.



Increased edge density increases closure probability.



Increased tie strength increases closure probability.



Tension between edge density and tie strength.

We used this for a new higher-order link prediction task.

time

Data.

$t_1 : \{1, 2, 3, 4\}$

$t_2 : \{1, 3, 5\}$

$t_3 : \{1, 6\}$

$t_4 : \{2, 6\}$

$t_5 : \{1, 7, 8\}$

$t_6 : \{3, 9\}$

t

$t_7 : \{5, 8\}$

$t_8 : \{1, 2, 6\}$

- Observe simplices up to time t.
- Predict which groups of > 2 nodes will appear after time t.

We predict structure that graph models would not even consider!

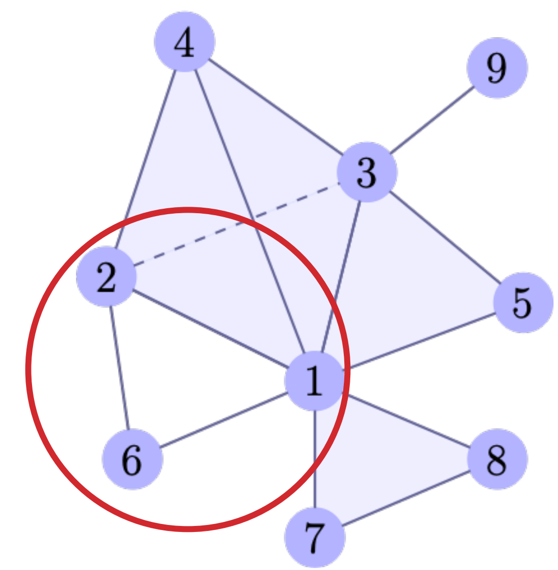
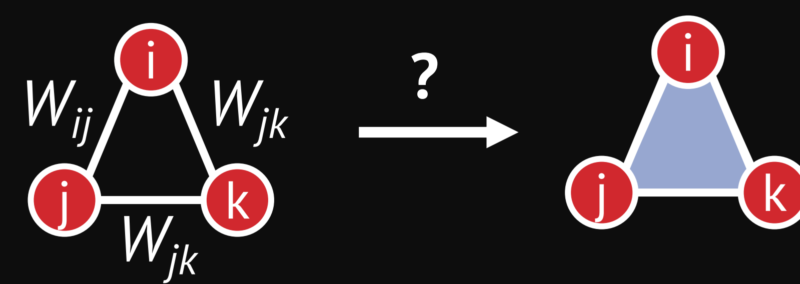


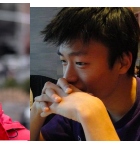
Table 2: Open triangle closure prediction performance based on several score functions: random (Rand.); harmonic, geometric, and arithmetic means of the 3 edge weights (Eqs. (19) to (21)); 3-way common neighbors (Common, Eq. (22)); 3-way Jaccard coefficient (Jaccard, Eq. (23)); 3-way Adamic-Adar (A-A, Eq. (24)); projected graph degree and simplicial degree preferential attachment (PGD-PA, Eq. (25) and SD-PA, Eq. (25)); unweighted and weighted Katz similarity (Katz, Eq. (29) and W-Katz, Eq. (30)); unweighted and weighted personalized PageRank (U-PPR, Eq. (34) and W-PPR, Eq. (35)); simplicial personalized PageRank (S-PPR, Eq. (42); the two missing entries are cases where computations did not finish within 2 weeks); and a feature-based supervised method logistic regression (Log. reg.). Performance is AUC-PR relative to the random baseline. The random baseline is listed in absolute terms and equals the fraction of open triangles that close.

Dataset	Rand.	Harm. mean	Geom. mean	Arith. mean	Common	Jaccard	A-A	PGD-PA	SD-PA	U-Katz	W-Katz	U-PPR	W-PPR	S-PPR	Log. reg.
coauth-DBLP	1.68e-03	1.49	1.59	1.50	1.33	1.84	1.60	0.74	0.74	0.97	1.51	1.62	1.83	1.21	3.37
coauth-MAG-History	7.16e-04	1.69	2.72	3.20	5.11	2.24	5.82	1.50	2.49	6.30	3.40	1.66	1.88	1.35	6.75
coauth-MAG-Geology	3.35e-03	2.01	1.97	1.69	2.43	1.84	2.71	1.31	0.97	1.99	1.74	1.06	1.26	0.94	4.74
music-rap-geni	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1.39	2.67
tags-stack-over	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	–	3.37
tags-math-sx	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1.86	13.99
tags-ask-ubunt	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1.19	7.48
threads-stack-o	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	–	1.53
threads-math-s	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.61	47.18
threads-ask-ubu	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1.78	9.82
NDC-substances	1.17e-03	4.90	3.27	2.90	5.92	3.36	5.97	4.76	4.46	5.35	2.93	1.39	1.83	1.86	8.17
NDC-classes	6.72e-03	4.43	3.38	1.82	1.27	1.19	0.99	0.94	2.14	0.92	1.34	0.78	0.91	2.45	0.62
DAWN	8.47e-03	4.43	3.86	2.13	4.73	3.76	4.77	3.76	1.45	4.61	2.04	1.57	1.37	1.55	2.86
congress-committees	6.99e-04	3.59	3.28	2.48	4.83	2.49	5.04	1.06	1.31	3.21	2.59	1.50	3.89	2.13	7.67
congress-bills	1.71e-04	0.93	0.90	0.88	0.65	1.23	0.66	0.60	0.55	0.60	0.78	3.16	1.07	6.01	107.19
email-Enron	1.40e-02	1.78	1.62	1.33	0.85	0.83	0.87	1.27	0.83	0.99	1.28	3.69	3.16	2.02	0.72
email-Eu	5.34e-03	1.98	2.15	1.78	1.28	2.69	1.37	0.88	1.55	1.01	1.79	1.59	1.75	1.26	3.47
contact-high-school	2.47e-03	3.86	4.16	2.54	1.92	3.61	2.00	0.96	1.13	1.72	2.53	1.39	2.41	0.78	2.86
contact-primary-school	2.59e-03	5.63	6.40	3.96	2.98	2.95	3.21	0.92	0.94	1.63	4.02	1.41	4.31	0.93	6.91

$$\text{score}_p(i,j,k) = (W_{ij}^p + W_{jk}^p + W_{ik}^p)^{1/p}$$



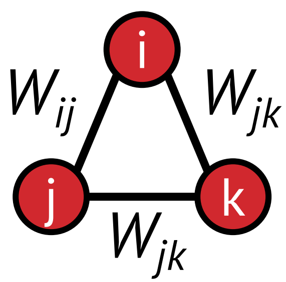
Finding the top-k weighted triangles in large graphs required new algorithms.


w/ R Kumar, P Liu, M Charikar

Retrieving Top Weighted Triangles in Graphs, WSDM, 2020.

$$\mathrm{score}_p(i,j,k) = (W_{ij}^p + W_{jk}^p + W_{ik}^p)^{1/p}$$



Finding top 1000 triangles

dataset	# nodes	# edges	time (existing)	time (ours)
reddit-reply	8.4M	435M	1.1 hours	5 seconds
spotify	3.6M	1.9B	> 24 hours	31 seconds

Triadic analysis for modern network data.

w/ A Paranjape, J Leskovec, P Liu, M Charikar

1. Open and closed triangles in temporal, higher-order interactions.
   Simplicial closure and higher-order link prediction, PNAS, 2018.

2. Triadic motifs in temporal networks.
   Motifs in temporal networks, WSDM 2017.
   Sampling methods for counting temporal motifs, WSDM, 2019.

Temporal network data is extremely common.

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

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

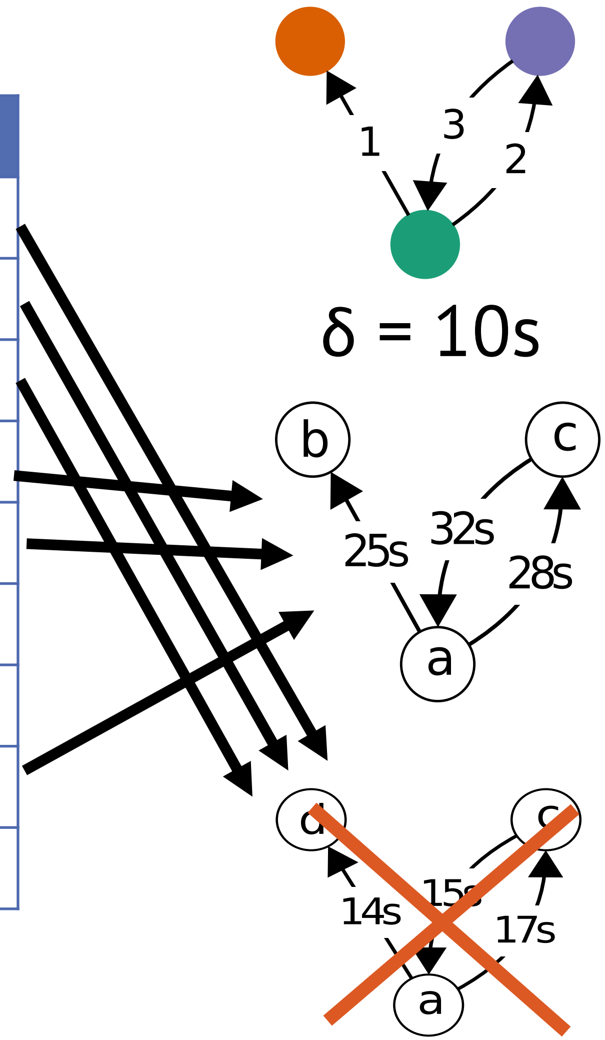
Payment systems
credit card transactions, cryptocurrencies, Venmo

Technical infrastructure
packets over the Internet, messages over supercomputer

We developed a model for temporal motifs.

Motifs in Temporal Networks, WSDM, 2017.

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

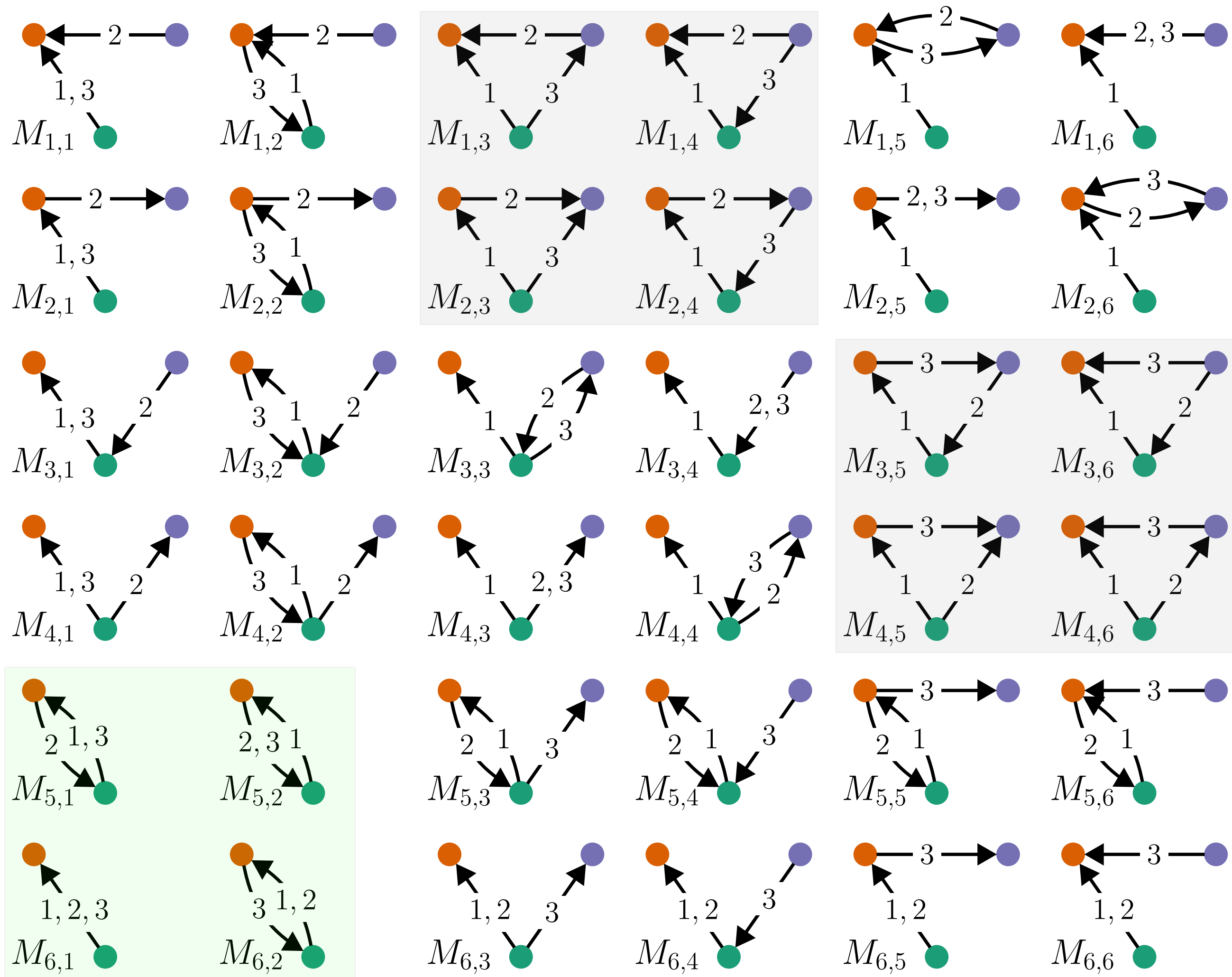
Motif instance

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

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

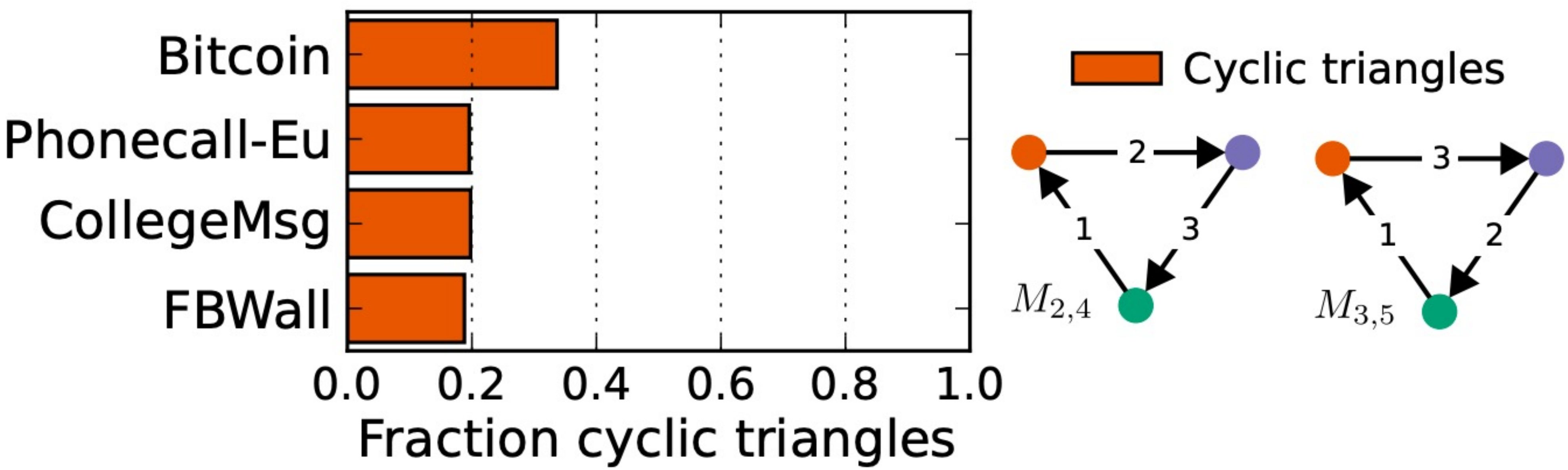
See also *Temporal Network Motifs: Models, Limitations, Evaluation*, Liu, Guarrasi, & Sarıyüce, 2021.

We also developed fast counting algorithms.

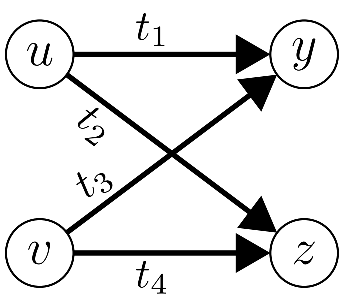


It takes ~2.5 hours to count all instances of these motifs in a 2B edge phone call network (single threaded).

Cyclic triangles are much more frequent in payment networks than in social networks.



Sampling algorithms let us go even faster for large datasets and more complicated motifs.



		running time (seconds)			
dataset	# temporal edges	exact	sampling	par. sampling	
EquinixChicago	345M	481.2	45.50	5.666	1.3%
RedditComments	636M	**X**	6739	2262	–

δ = 1 day, 16 threads

Triadic data analysis in temporal and higher-order networks

THANKS! Austin Benson
http://cs.cornell.edu/~arb
@austinbenson
arb@cs.cornell.edu

Santa Fe, NM

Lots of data available at https://www.cs.cornell.edu/~arb/data/

Supported by ARO MURI, ARO Award W911NF19- 1-0057, NSF Award DMS-1830274, and JP Morgan Chase & Co.