Configuration Models of Random Hypergraphs

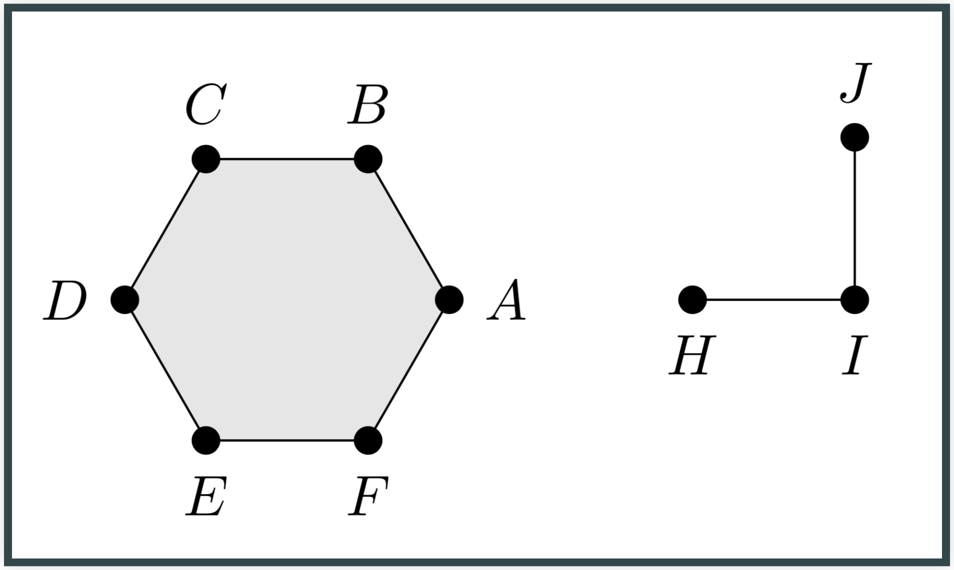
Phil Chodrow, *Massachusetts Institute of Technology*

Statistical Inference for Network Models, NetSci 2019

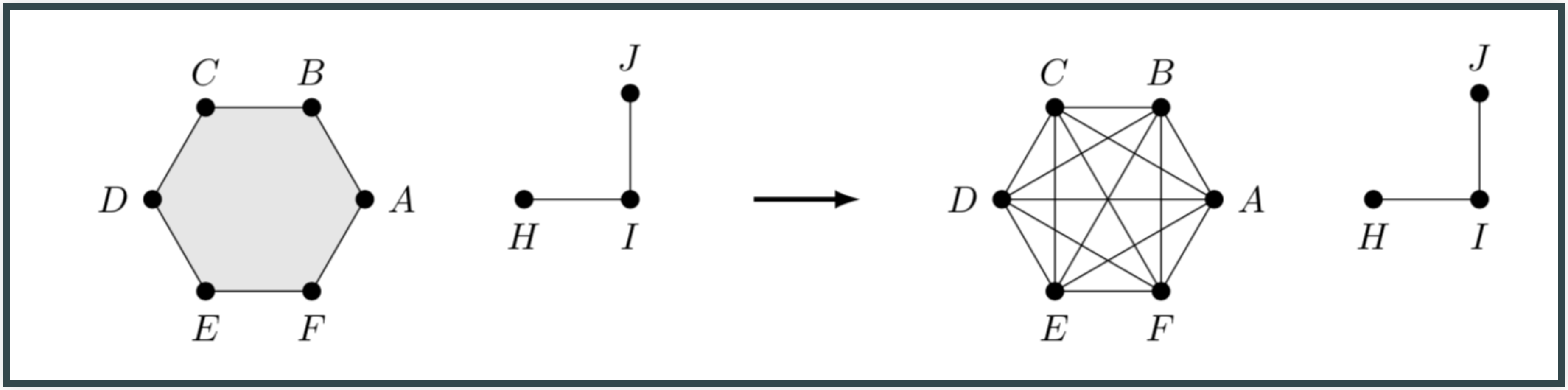
May 27th, 2019

Introduction

Polyadic Networks



- **Collaboration** (author/paper, sponsor/bill)
- **Communication** (receiver/email, participant/forum thread)
- **Composition** (word/sentence, ingredient/recipes, agent/group)

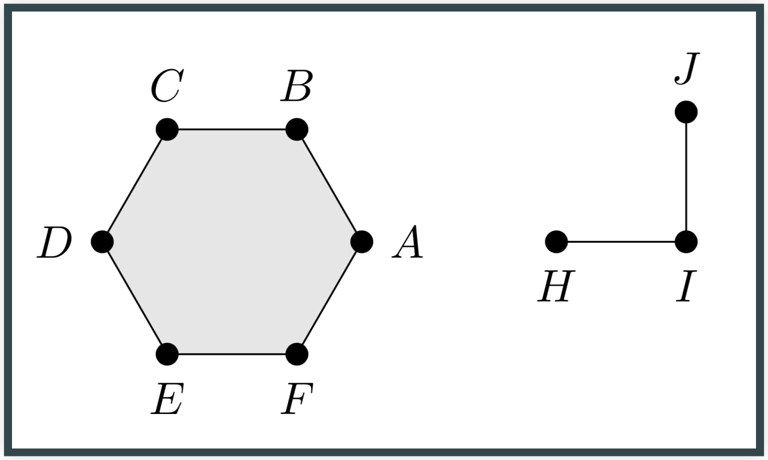


Historically, polyadic networks have often been treated through dyadic (graph) techniques.

Argument for Today

- Dyadic graph techniques can produce **unreliable results** on polyadic data sets.
- We should instead use natively polyadic **metrics** and **models**.
- Random hypergraph null models suggest **new perspectives** on some conventional wisdom in network science.

“Is This Data Set Degree-Assortative?”



- Dyadic: **Assortative!** $\rho = 0.99, p < 0.01.$
- Polyadic: **Disassortative!** $\rho = -0.25, p < 0.01.$

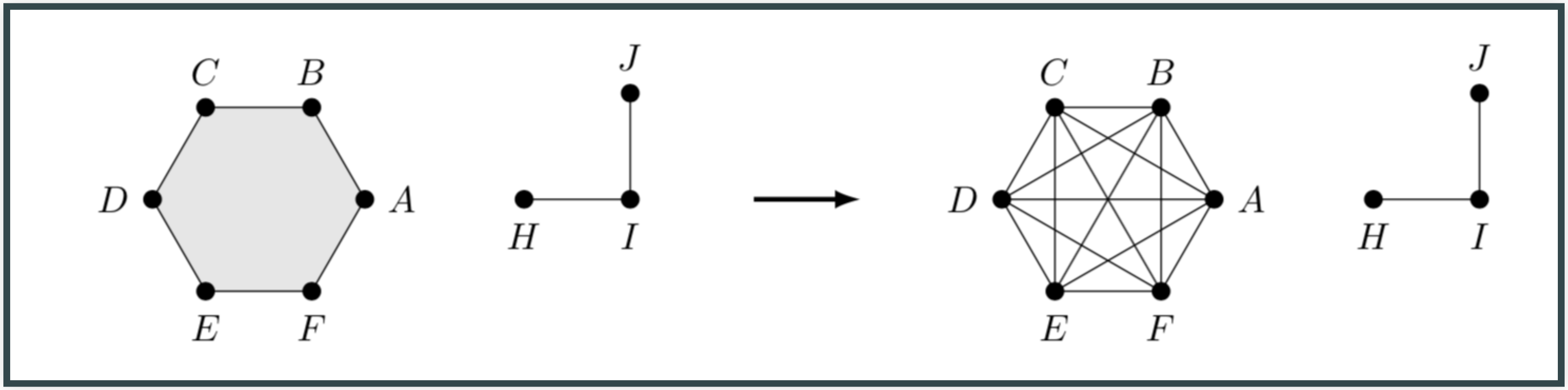
More Precisely...

*“Is the observed data more **assortative** by **degree** than would be expected **at random**?”*

We need three ingredients:

- A data representation that will determine the node **degrees**.
- A measure of **assortativity**.
- A **null model**.

The Dyadic Approach (“Assortative!”)



- **Data representation:** Least productive nodes have highest degrees.
- **Measurement:** Single interaction gets counted 15 times in ρ.
- **Null model:** Dyadic nulls models compare to counterfactuals with 17 two-author papers.

Hypergraph Configuration Models

Hypergraphs

A *hypergraph* $G = (V, E)$ consists of a vertex set V and an edge-set E. E is a *multiset of subsets of* V. Multi-edges are ok. The node degree is the number of incident edges (**not** number of neighbors.)

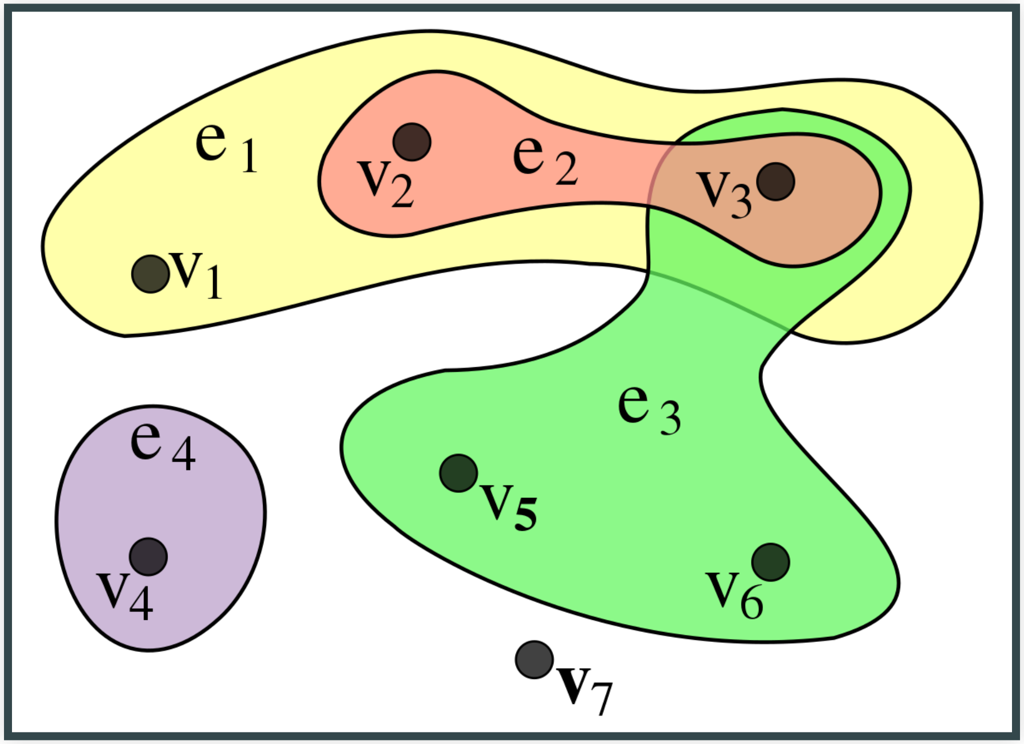


Image: Wikipedia

Configuration Models for Hypergraphs

Observe G with degree sequence $\deg(G) = \mathbf{d} \in \mathbb{R}^n$
and edge dimension sequence $\dim(G) = \mathbf{k} \in \mathbb{R}^m$.

Null space: all graphs with the same degree sequence
and edge dimension sequence.

The **vertex-labeled hypergraph configuration model** $\eta_{\mathbf{d},\mathbf{k}}$ is the uniform distribution on this null space.

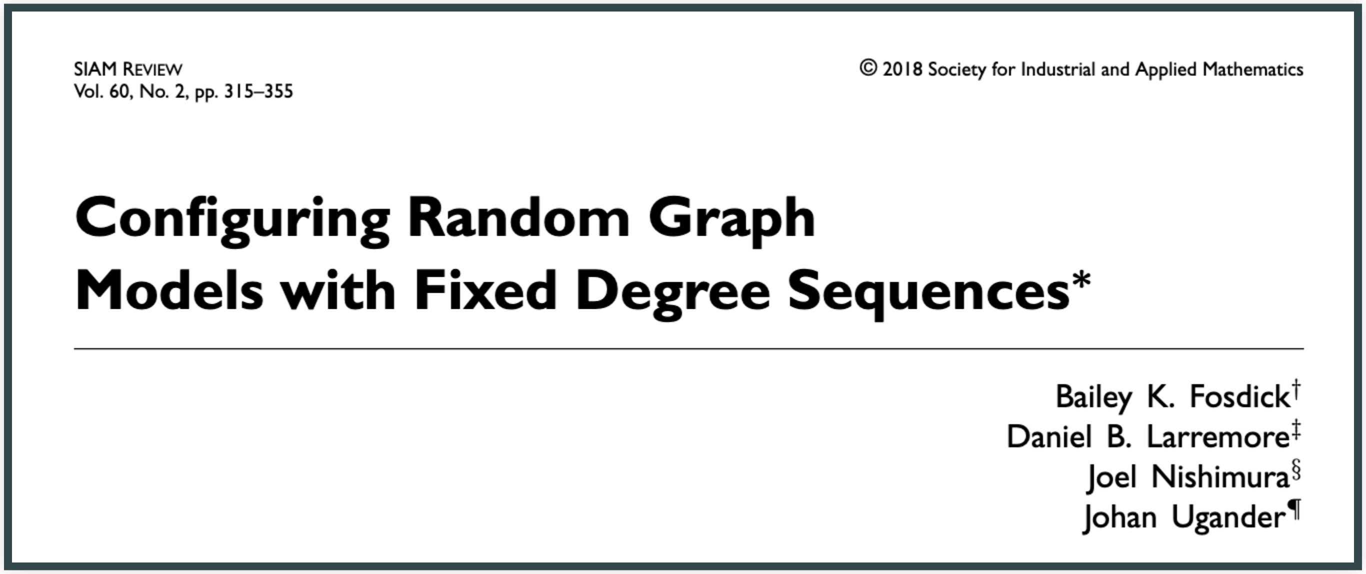
Metropolis-Hastings Sampling

1. Initialize $G_0 \in \mathcal{V}_{\mathbf{d},\mathbf{k}}$.
2. For $i = 0, 1, 2, \dots$

- **Proposal:**
  - Select edges Δ and Γ uniformly at random from E_i.
  - Randomly reshuffle nodes between Δ and Γ, preserving $|\Delta|$ and $|\Gamma|$.
- **Accept** the proposal with probability $a(\Delta, \Gamma) = \frac{1}{m_\Delta m_\Gamma}$, generating G_{i+1}.

3. Return G_i at regular intervals of length t.

Related Work (Read It!)

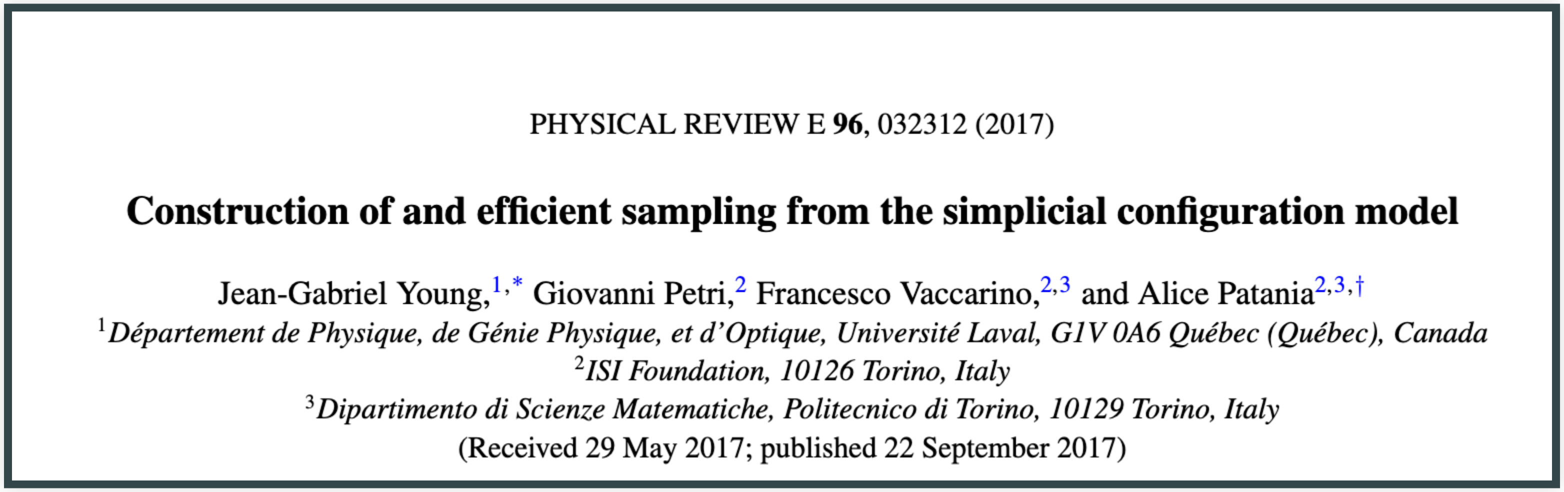


SIAM Review
Vol. 60, No. 2, pp. 315–355

© 2018 Society for Industrial and Applied Mathematics

Configuring Random Graph Models with Fixed Degree Sequences*

Bailey K. Fosdick†
Daniel B. Larremore‡
Joel Nishimura§
Johan Ugander¶



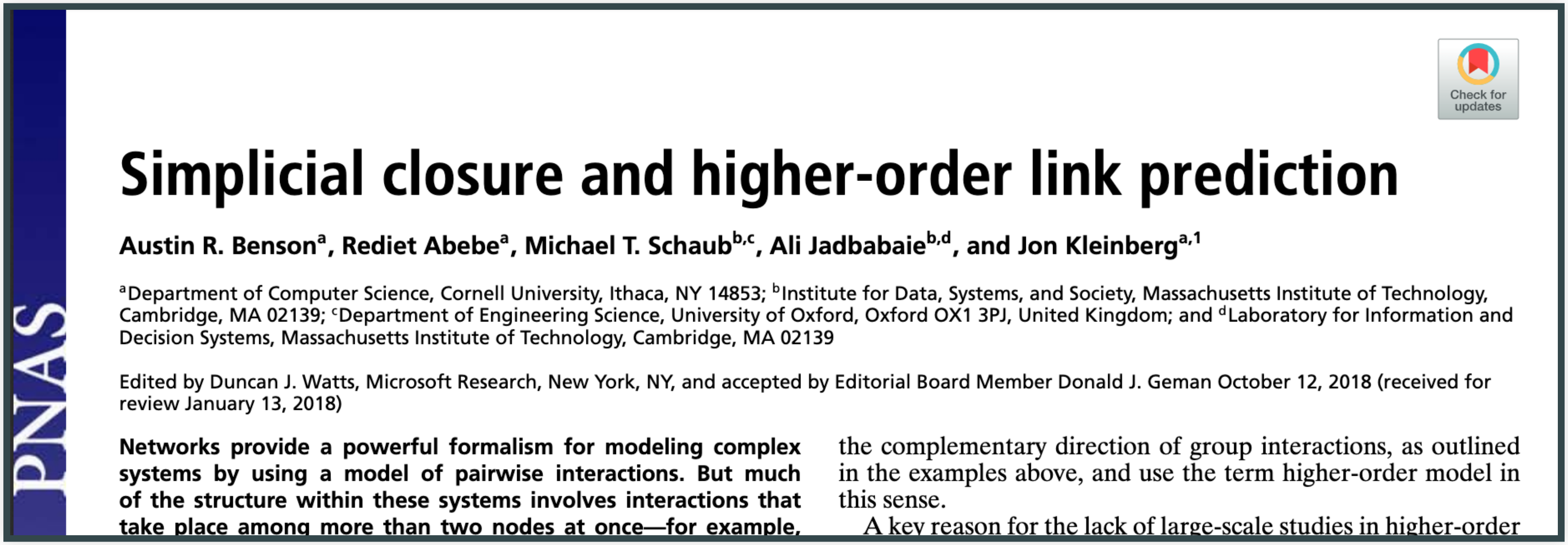
PHYSICAL REVIEW E **96**, 032312 (2017)

Construction of and efficient sampling from the simplicial configuration model

Jean-Gabriel Young,[1,*] Giovanni Petri,[2] Francesco Vaccarino,[2,3] and Alice Patania[2,3,†]
[1]*Département de Physique, de Génie Physique, et d'Optique, Université Laval, G1V 0A6 Québec (Québec), Canada*
[2]*ISI Foundation, 10126 Torino, Italy*
[3]*Dipartimento di Scienze Matematiche, Politecnico di Torino, 10129 Torino, Italy*
(Received 29 May 2017; published 22 September 2017)

Applications

Thanks for the Data!



PNAS

Check for updates

Simplicial closure and higher-order link prediction

Austin R. Benson[a], Rediet Abebe[a], Michael T. Schaub[b,c], Ali Jadbabaie[b,d], and Jon Kleinberg[a,1]

[a]Department of Computer Science, Cornell University, Ithaca, NY 14853; [b]Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA 02139; [c]Department of Engineering Science, University of Oxford, Oxford OX1 3PJ, United Kingdom; and [d]Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139

Edited by Duncan J. Watts, Microsoft Research, New York, NY, and accepted by Editorial Board Member Donald J. Geman October 12, 2018 (received for review January 13, 2018)

Networks provide a powerful formalism for modeling complex systems by using a model of pairwise interactions. But much of the structure within these systems involves interactions that take place among more than two nodes at once—for example,

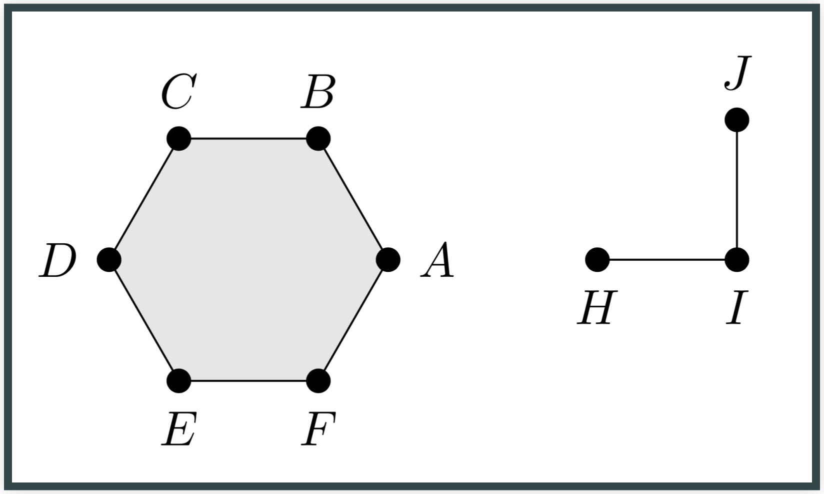
the complementary direction of group interactions, as outlined in the examples above, and use the term higher-order model in this sense.

A key reason for the lack of large-scale studies in higher-order

Time-stamped polyadic data sets: emails; drug co-occurrence; Congressional cosponsorships; academic coauthorships; forum threads...

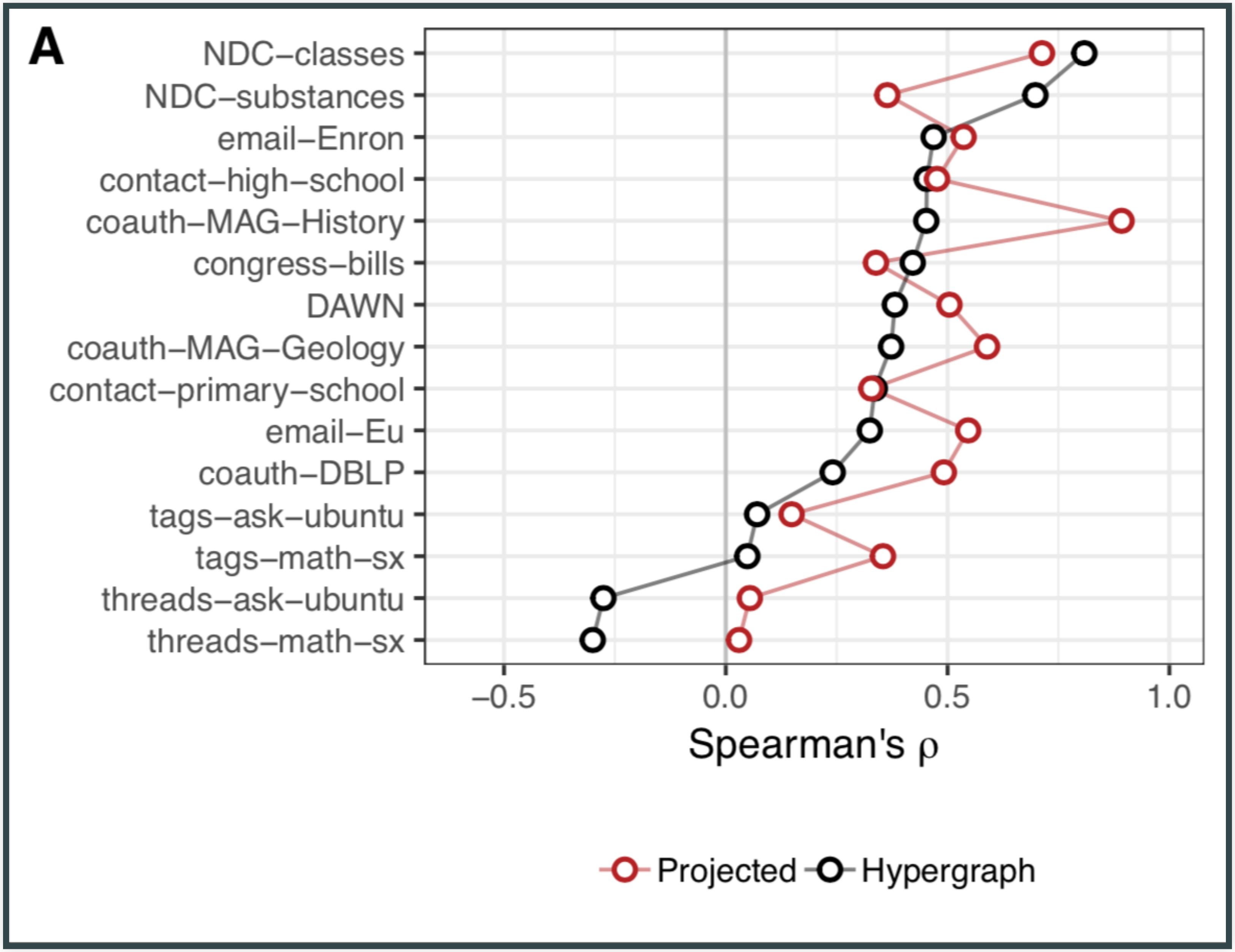
Assortativity in Polyadic Networks

Hypergraph Assortativity

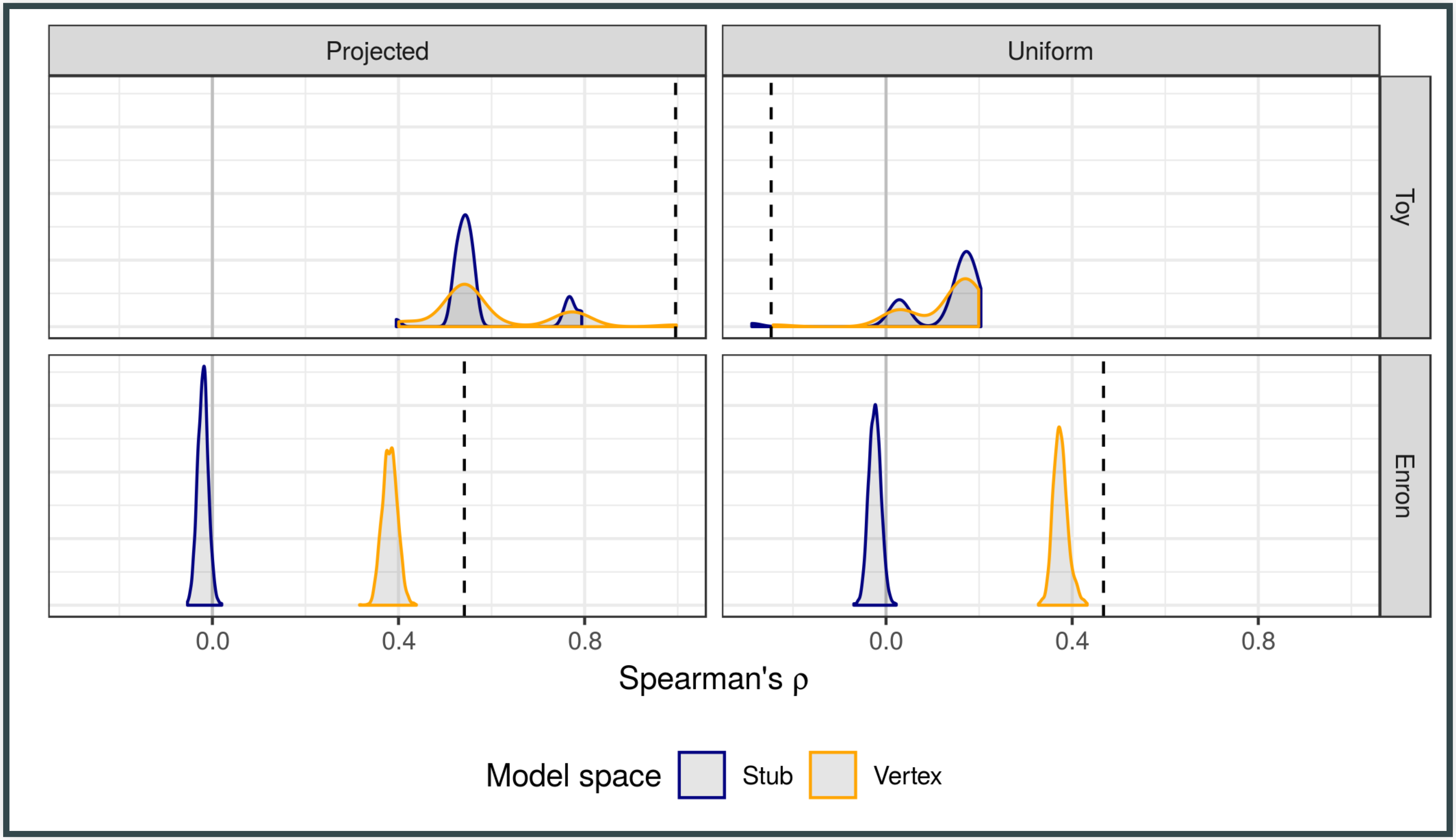


$$\rho_{\text{polyadic}} = \frac{\langle r_u r_v \rangle - \langle r_u \rangle^2}{\langle r_u^2 \rangle - \langle r_u \rangle^2} = -0.25 \ll 0.99 = \rho_{\text{dyadic}}$$

Projected and Hypergraph Assortativities



Significance Tests



Learnings

- Degree-assortativity can be studied **without projections**.
- Doing so **better matches our intuitions** about what assortativity “should mean” in many cases.
- The results can **vary significantly** from projected dyadic approaches.

Revisiting Clustering

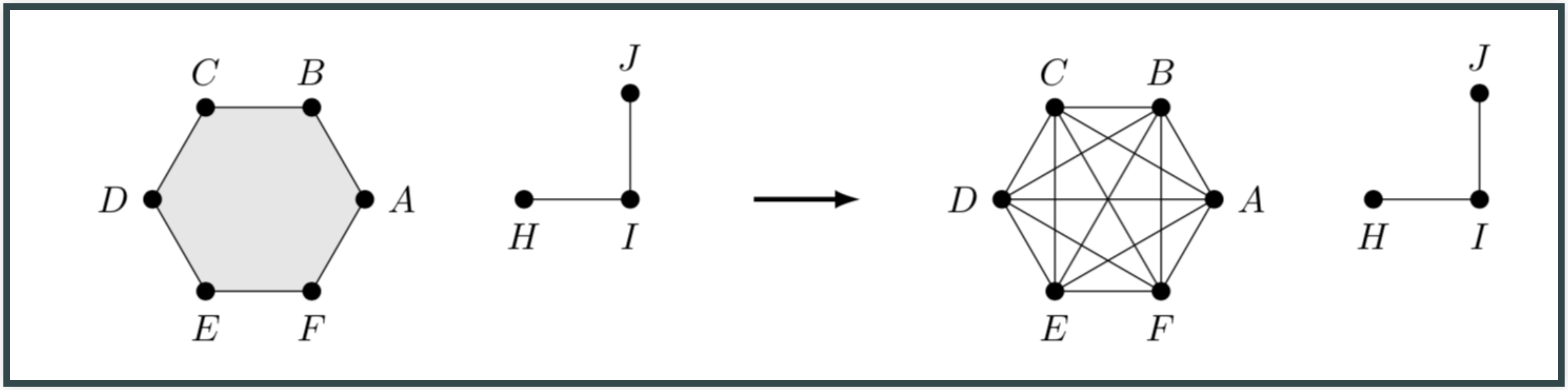
The Global Clustering Coefficient

“Your friends tend to know each other...”

$$C = \frac{3 \times \text{Number of triangles}}{\text{Number of 2-stars}}$$

Recall (Strogatz and Watts 1998): Social networks display clustering; configuration models don’t; we need a small world model...

Triadic Closure in Polyadic Networks



$$C = 3 \times \frac{\binom{6}{3}}{3 \times \binom{6}{3} + 1} \approx 0.98$$

- **“Trivial”** triadic closure from within edges.
- **“Nontrivial”** triadic closure from correlations between edges (Newman 2001).
- How to distinguish? Compare to a null distribution...

Null Significance Tests

	C	$\langle C\rangle_h$	$\langle C\rangle_{\mathrm{dyadic}}$
congress-bills	0.61	0.60 (0.00)	0.45 (0.00)
coauth-MAG-Geology	0.82	0.82 (0.00)	0.00 (0.00)
email-Enron	*0.66*	0.82 (0.01)	0.64 (0.01)
email-Eu	*0.54*	0.60 (0.01)	0.40 (0.00)
tags-ask-ubuntu	*0.57*	0.61 (0.01)	0.19 (0.00)
threads-math-sx	*0.29*	0.40 (0.01)	0.04 (0.00)

Hypothesis: When multi-way interactions are "cheap," potential triangles get absorbed by polyadic edges. When coordination is required (e.g. collaborations), triangles may be more prevalent.

Edge Intersection Profiles

The Simplicial Clustering Hypothesis

Hypothesis: Edges have larger intersections than would be expected by random chance.

Related to *simplicial closure* hypothesis (Patania, Petri, and Vaccarino 2017; Benson et al. 2018).

Intersection Profile

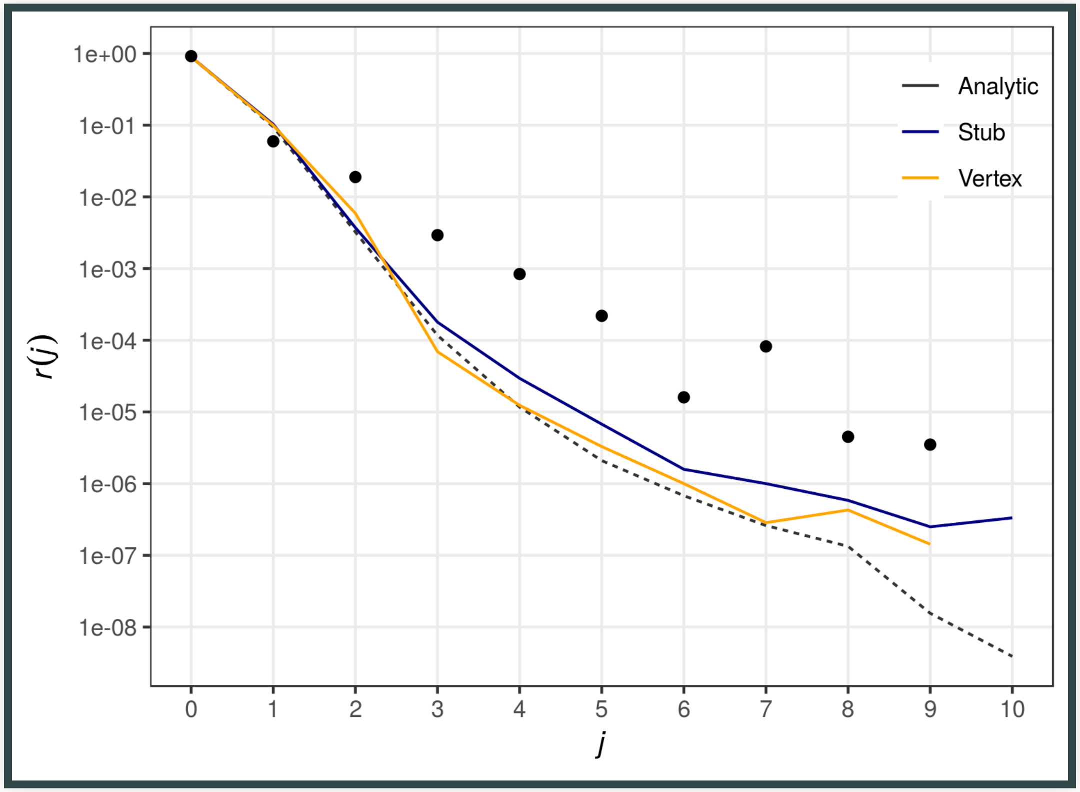
Definition: The *edge intersection profile* of a hypergraph is the empirical distribution

$$r(j) = \mathbb{P}(|\Delta \cap \Gamma| = j) \ ,$$

where Δ and Γ are uniformly random edges.

Large $r(j) \implies$ intersections of size j are common.

Intersection Profile of Enron Emails



Large intersections much more common than expected under any null.

Null Profiles for Large Networks

What if a network is too big for Monte Carlo sampling?

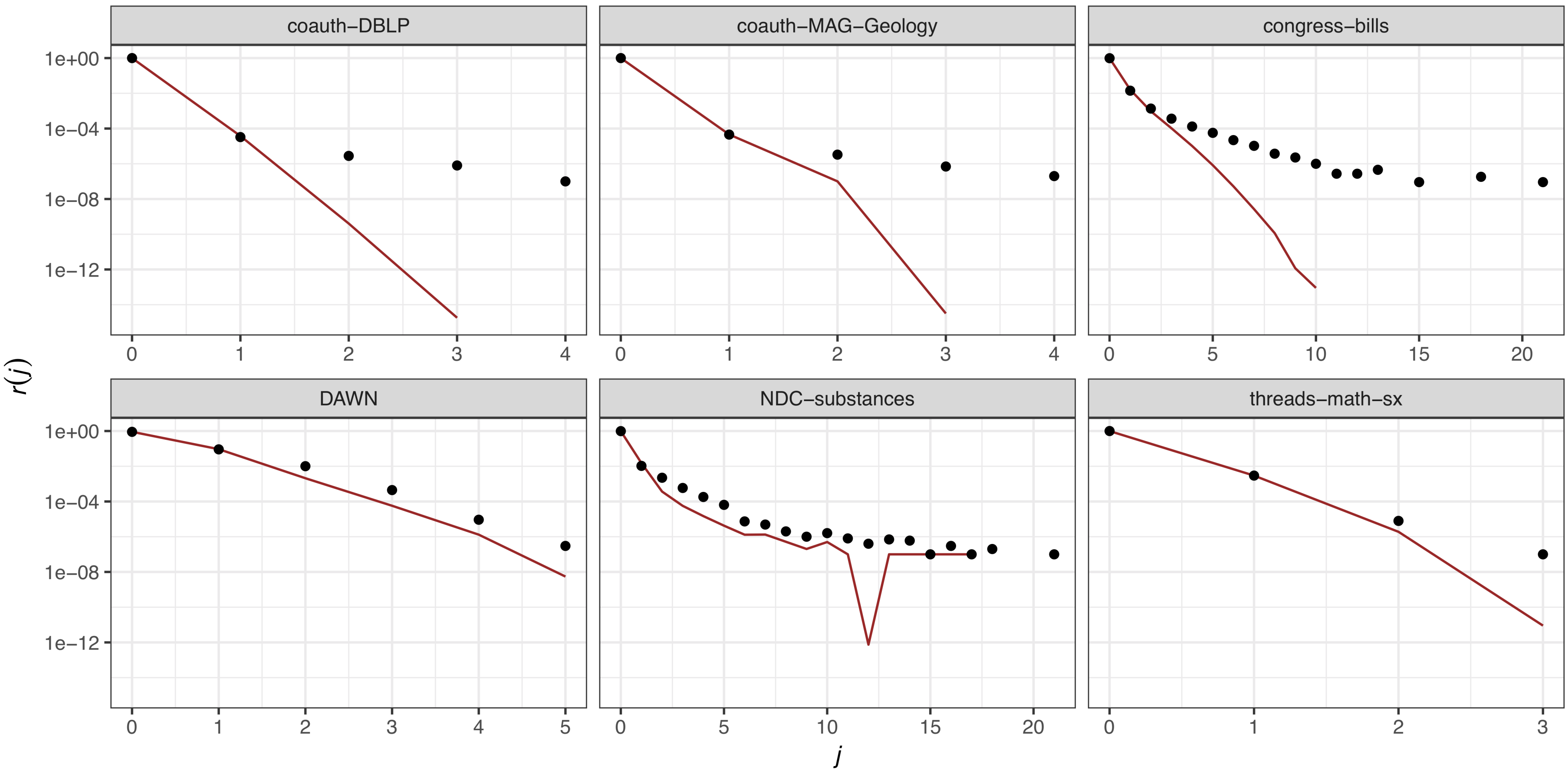
Let $r_{k\ell}(j) = \eta\left(|\Delta \cap \Gamma| = j \,|\, |\Delta| = k, |\Gamma| = \ell\right)$.

Theorem: When $\langle d^2 \rangle < \infty$,

$$r_{k\ell}(j) = (1 + O(n^{-1}))j!\binom{k}{j}\binom{\ell}{j}\left(\frac{1}{n}\frac{\langle d^2 \rangle - \langle d \rangle}{\langle d \rangle^2}\right)^j$$

with high probability as $n \to \infty$.

Intersection Profiles for Large Networks



Learnings

- Intersection profiles are **natively polyadic, scalable** measures of edge correlations.
- Some empirical networks are significantly clustered in this measure; others less so.

Wrapping Up

Takeaways

- Dyadic graph techniques can produce **unreliable results** on polyadic data sets.
- We should instead use natively polyadic **metrics** and **models**.
- Random hypergraph null models suggest **new perspectives** on some conventional wisdom in network science.

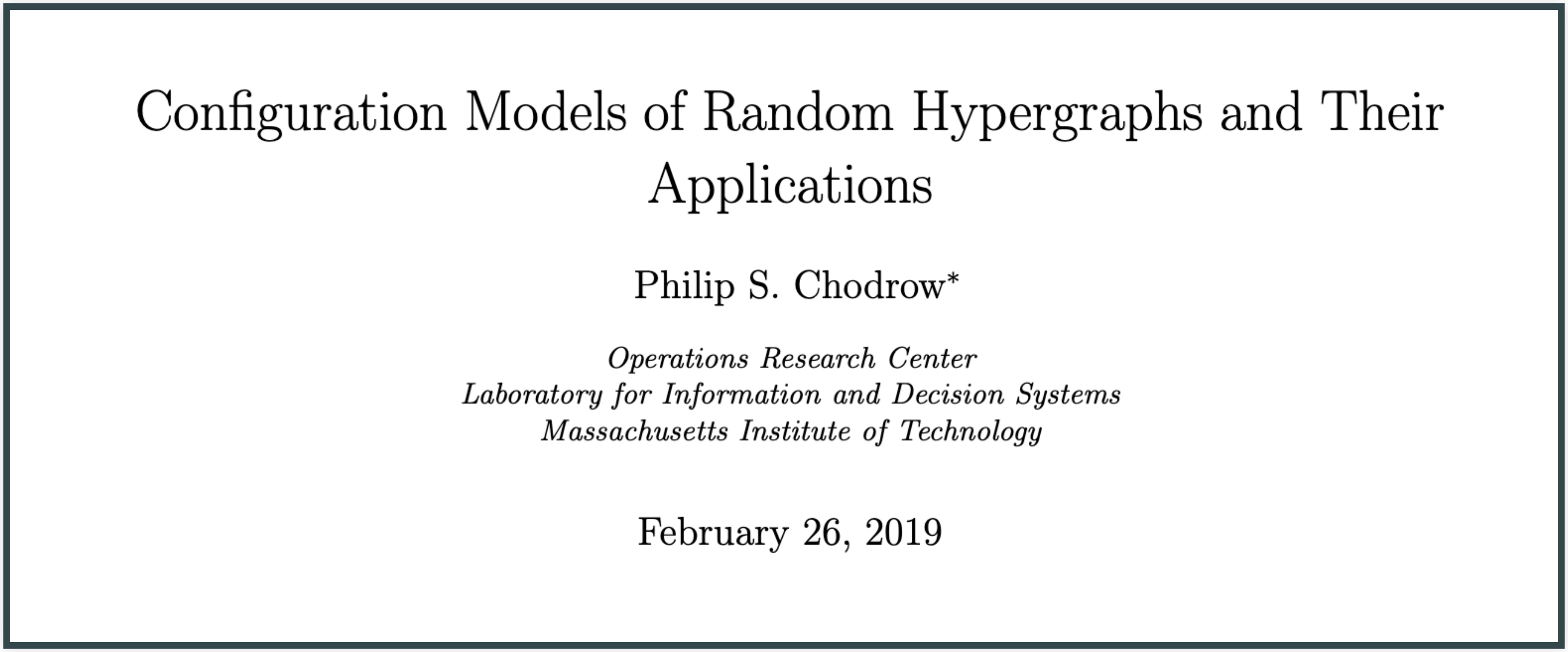
Potential Future Directions

- Community detection in polyadic data sets
- Optimal projections (when you must...)
- Mean-field theory for hypergraph dynamics

Thanks to...

- ...**Patrick Jaillet** (MIT), for helpful discussions.
- ...The **National Science Foundation**, for support under GRFP Grant 1122374.
- ...**The organizers**, for the opportunity to present today.
- ...To **you**, for your time and attention!

Learn More



Configuration Models of Random Hypergraphs and Their Applications

Philip S. Chodrow*

Operations Research Center
Laboratory for Information and Decision Systems
Massachusetts Institute of Technology

February 26, 2019

- **arXiv:** 1902.09302
- **GitHub:** PhilChodrow/hypergraph
- philchodrow.com
- @PhilChodrow

References

Benson, Austin R., Rediet Abebe, Michael T. Schaub, Ali Jadbabaie, and Jon Kleinberg. 2018. “Simplicial Closure and Higher-order Link Prediction.” *Proceedings of the National Academy of Sciences* 115 (48): 11221–30. http://arxiv.org/abs/1802.06916.

Newman, Mark E. J. 2001. “Scientific collaboration networks. I. Network construction and fundamental results.” *Physical Review E* 64 (1): 8. https://doi.org/10.1103/PhysRevE.64.016131.

Patania, Alice, Giovanni Petri, and Francesco Vaccarino. 2017. “The shape of collaborations.” *EPJ Data Science* 6 (1): 1–16. https://doi.org/10.1140/epjds/s13688-017-0114-8.

Strogatz, S. H., and D. J. Watts. 1998. “Collective dynamics of ’small-world’ networks.” *Nature* 393 (June): 440–42.