Applied Math Research Showcase: Statistics, Algebra, & Randomization

Sonja Petrović

Illinois Institute of Technology

April 2017

About me

- PhD in 2008 from University of Kentucky – commutative algebra
- Statistics faculty at Penn State before joining IIT in 2013
- Research with students at IIT:
  - *Dane Wilburne (PhD, networks and algebra)
  - *Denis Bajić (MS Data Science, computation for statistical network models)
  - Martin Dillon (BS summer research, McMorris stipend 2015, and MS 2016, mixing times for Markov chains)
  - Xintong Li (summer research, CoS stipend 2015, computational algebra for graph coloring)
  - Weronika Swiechowicz (BS summer research, CoS stipend 2014, computational algebra)
  - Yuanfang Xiang (summer research, McMorris stipend 2014, maximum likelihood estimation in statistics and multiple roots).

  → Weronika and Yuanfang's joint paper published by the SIAM Undergraduate Research Online (SIURO) journal in 2015.
  - *William Schwartz (advisor: H. Kaul, current PhD, temporal networks)

Getting started: kinds of questions we ask - Part I

Guiding question ('GoF'):

Determine if the observed data fits the proposed statistical model.

Getting started: kinds of questions we ask - Part I

Guiding question ('GoF'):

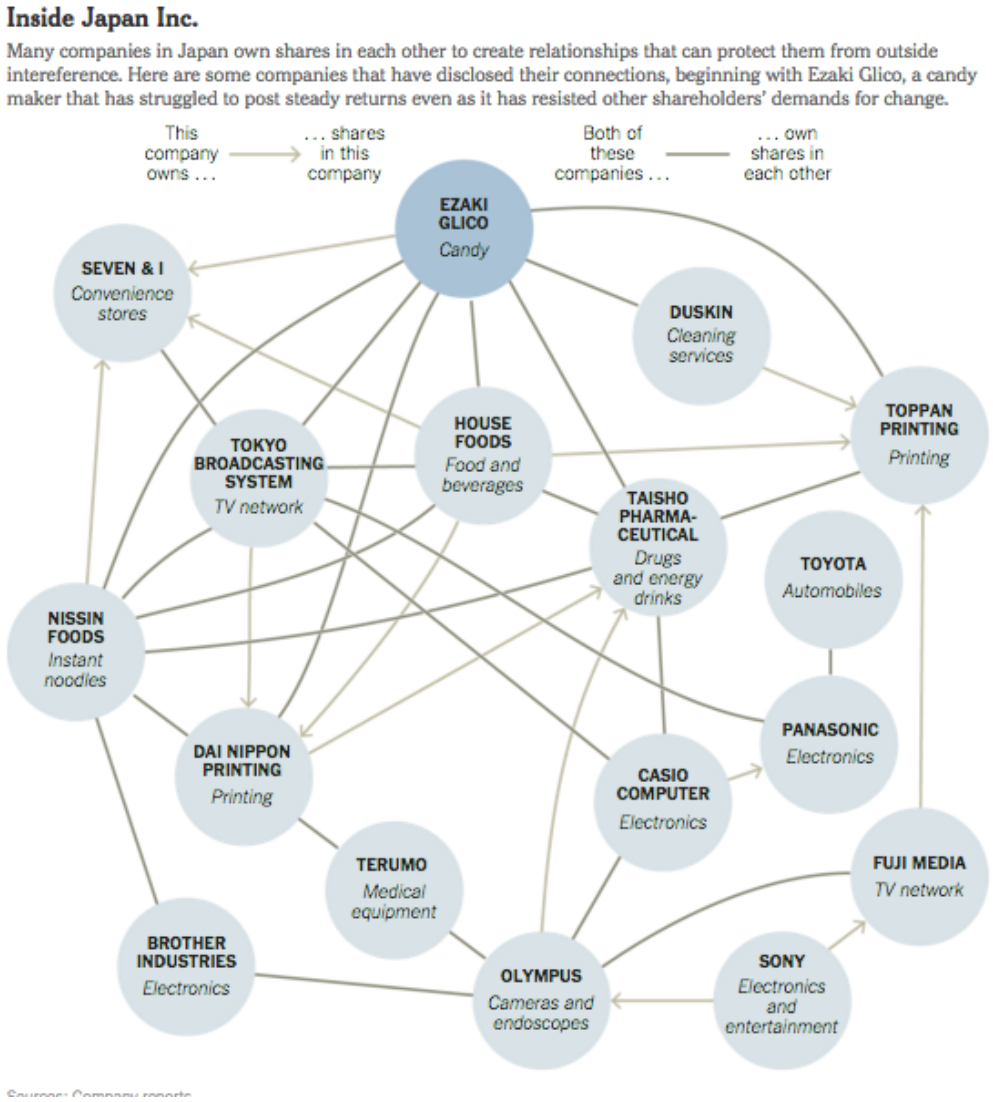
Determine if the observed data fits the proposed statistical model.

WHY search for a well-fitting model?

This is a basic question in statistics, related to hypothesis testing.

The answer for network data comes by using algebraic and graph-theoretic methods.

A toy example: when the GoF question is relevant



Source: New York Times

Key question

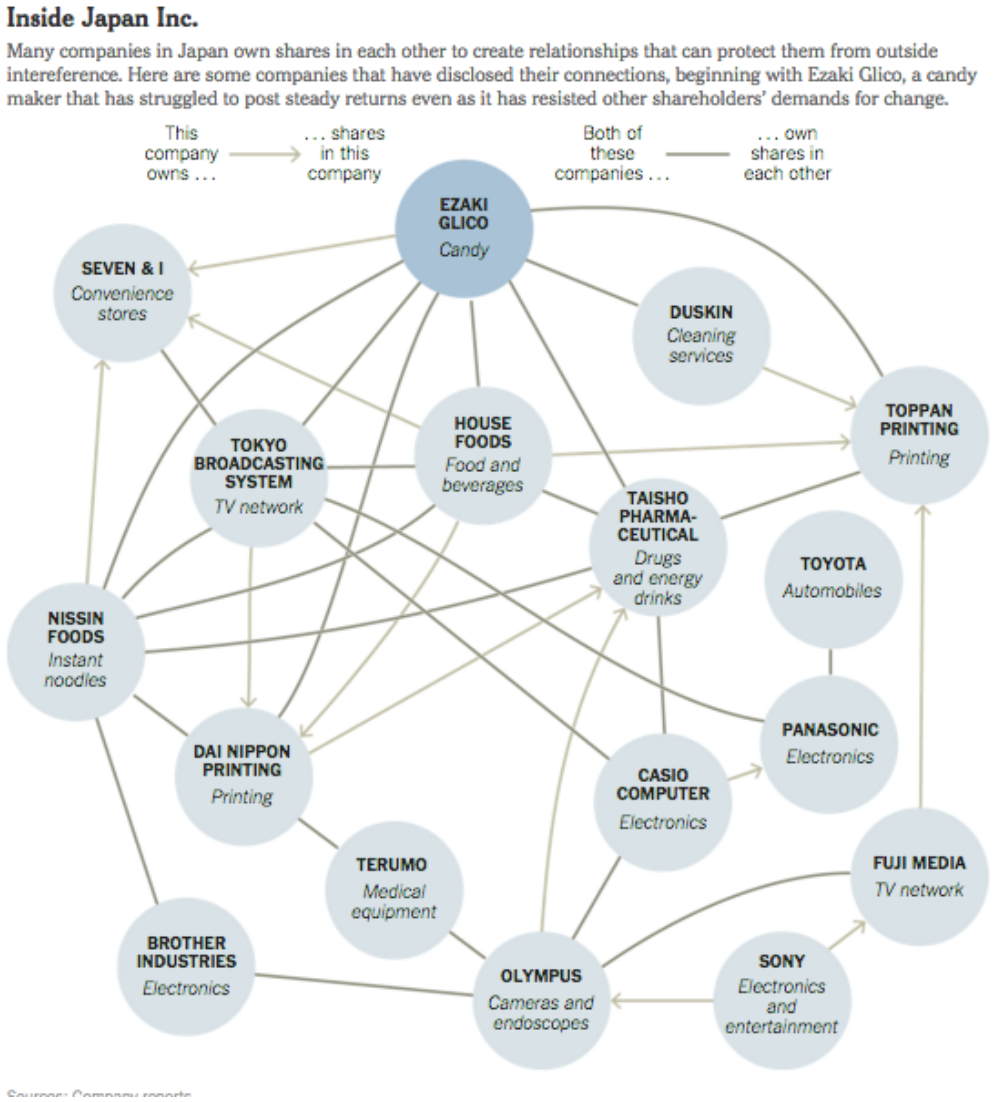
What is the local effect that we would like to capture/measure in the observed directed network?

Not only propensity of each corporation to buy shares, but:

Claim:

Strong reciprocation effect among the corporate directorates.

A toy example: when the GoF question is relevant



Source: New York Times

Key question

What is the local effect that we would like to capture/measure in the observed directed network?

Not only propensity of each corporation to buy shares, but:

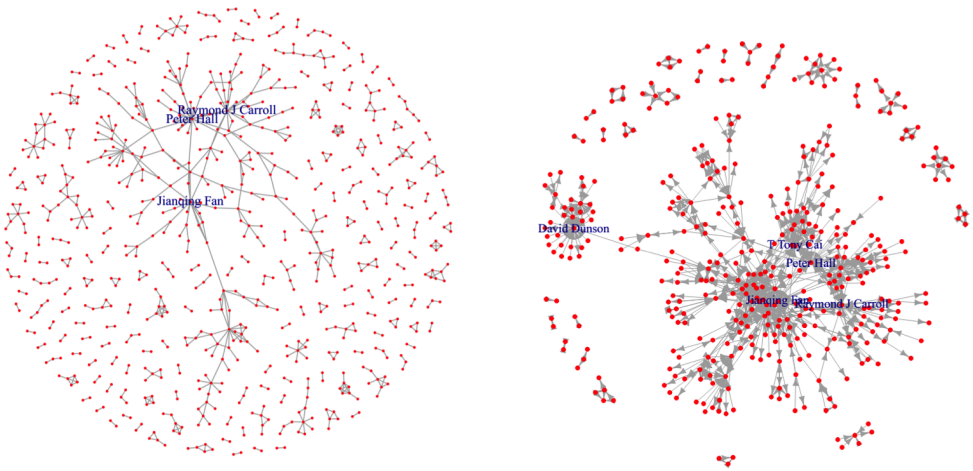
Claim:

Strong reciprocation effect among the corporate directorates.

→A satisfying answer will confirm model/data fit for a statistical model that captures this effect.←

Network data: new frontier for statistical inference

Another recent motivating example:



Ji & Jin, AOAS 2016: Coauthorship and Citation networks of statisticians

Questions:

- Summary statistics/actor identification *(count degrees to identify most collaborative or most cited authors)*: appropriate summaries??
- Model-based inference: other types of models necessary??
- These networks were created by thresholding counts

Interpretable well-fitting models?? ↔ Network(ed) data ↔ Data representation and summaries??

GoF test with dynamic Markov bases

Exact conditional test

Observed g is compared to a reference set of graphs.
$\rightarrow$ Requires sampling. (And a statistical justification.)

GoF test with dynamic Markov bases

Exact conditional test

Observed g is compared to a reference set of graphs.
$\rightarrow$ Requires sampling. (And a statistical justification.)

[Gross, Petrović, Stasi, Annals of ISM 2016 + 2017$^+$ work]

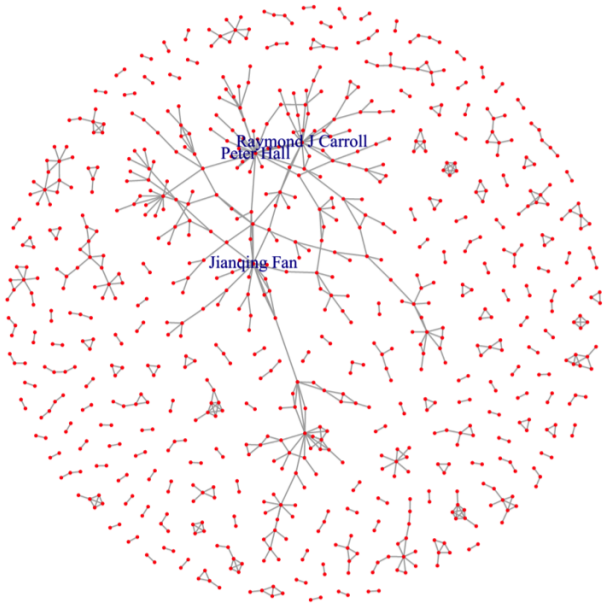
Dynamic sampling algorithm for degree-based network models.

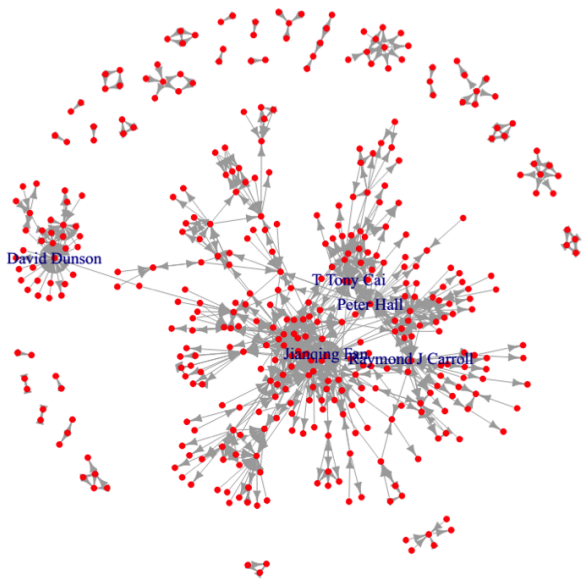
- Data-dependent, small/large moves.

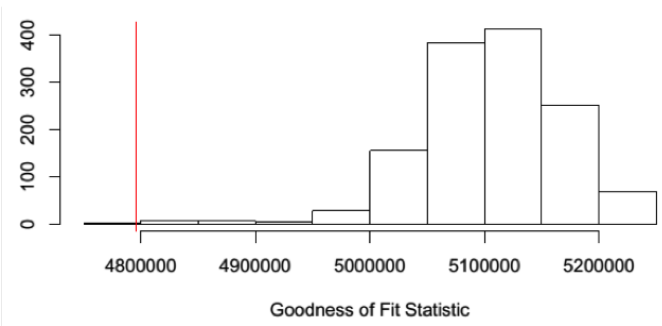
Theorem [Dillon 2016]

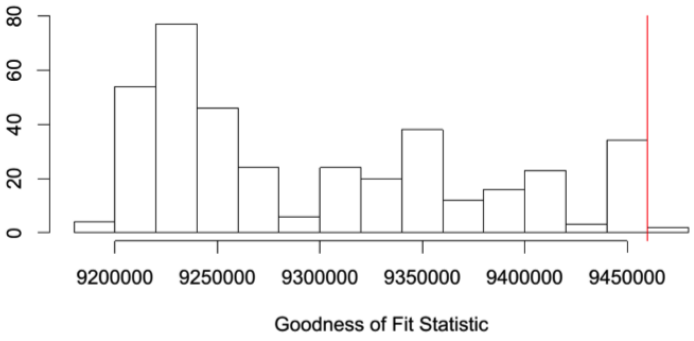
The dynamic Markov bases chain mixes rapidly for all fibers where graph-theoretic simple-switch chain mixes rapidly.

Application: Are degrees a good summary?





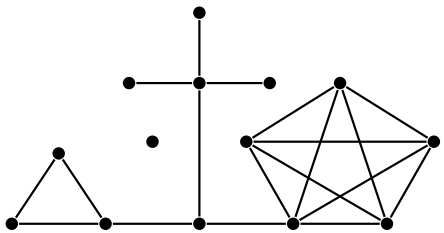




yes! no!

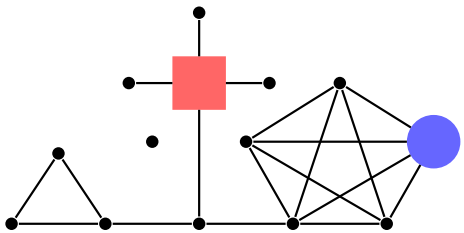
Karwa & Petrović, AOAS 2016: Coauthorship and citation networks of statisticians - comment

What lies beyond node degrees?



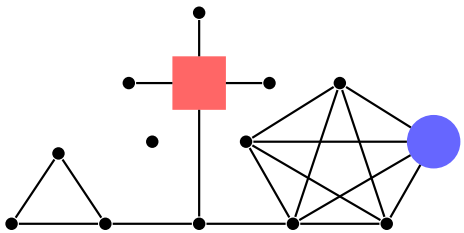
- Are you famous if you have lots of citations?

What lies beyond node degrees? (Interpretability?)



- Are you famous if you have lots of citations?
- Or lots of citations by people who themselves have lots of citations... ?

What lies beyond node degrees? (Interpretability?)



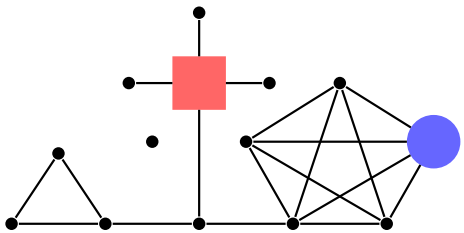
- Are you famous if you have lots of citations?
- Or lots of citations by people who themselves have lots of citations... ?

Core decomposition [Seidman'83]

Descriptive tool to explain properties of observed graphs:

- Core-periphery (rich club) structure
- Importance of a node (robustness)
- Visualization of topology - peel into layers

What lies beyond node degrees? (Interpretability?)



- Are you famous if you have lots of citations?
- Or lots of citations by people who themselves have lots of citations... ?

Core decomposition [Seidman'83]

Descriptive tool to explain properties of observed graphs:

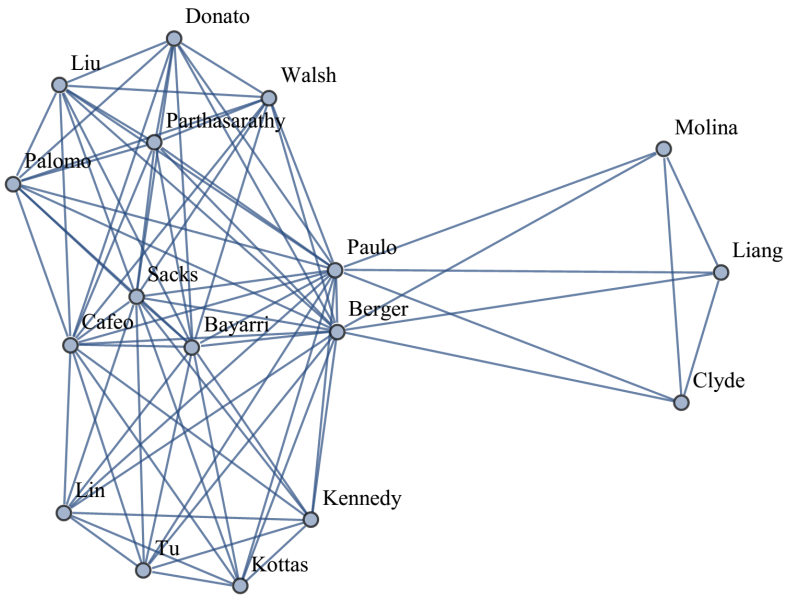
- Core-periphery (rich club) structure
- Importance of a node (robustness)
- Visualization of topology - peel into layers

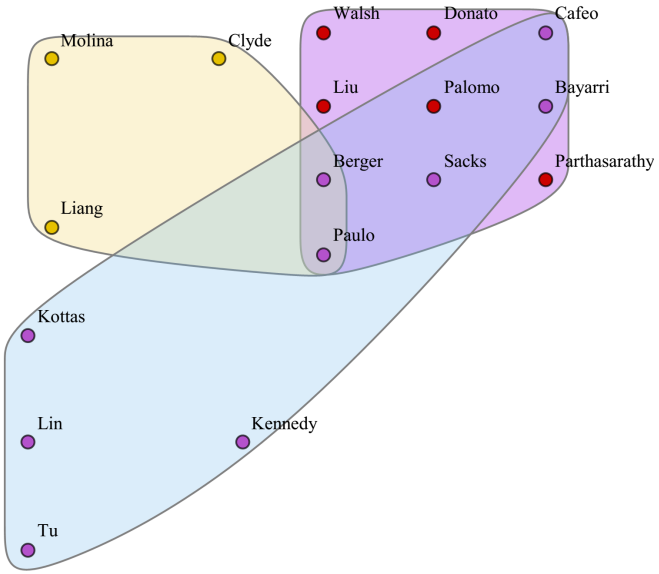
[Karwa, Pelsmajer, Petrović, Stasi, Wilburne: EJS 2016]

Statistical model: k-cores ERGM for undirected networks.

Computational issues: some addressed, some work in progress.

Graphs vs. hypergraphs





Karwa & Petrović, AOAS 2016: Coauthorship and Citation networks of statisticians - comment

[Most collaborative authors by hypergraph degree $\neq$ by graph degree.]

- A hyperedge of size k $\leftrightarrow$ a paper by k authors.
  Heterogeneity of number of coauthors.
- Hypergraph degree of i = number of hyperedges containing i.
- Cf. [Stasi, Sadeghi, Rinaldo, Petrović, Feinberg 2014]: Beta models for random hypergraphs with a given degree sequence.

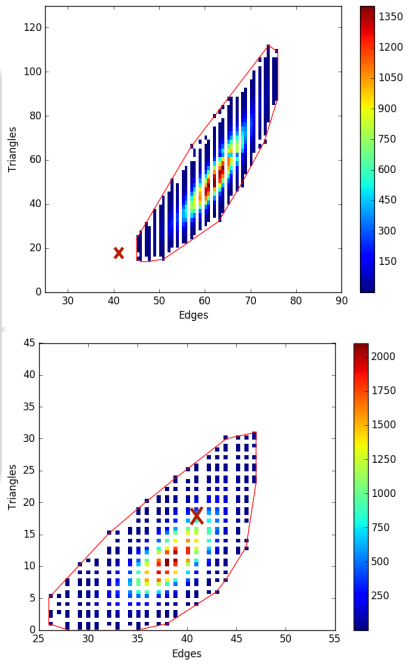
Most real-world networks are 'sparse' → adjust the models!

- Consider ERGMs: large flexible family of network models, natural to model networks through their summary statistics
- Large & growing literature and applications
- But some challenges remain...

Karwa, Petrović, Bajić 2017$^+$

Degeneracy-restricted ERGMs:

- Fix the degenerate behavior of ERGMs.
- Solve the computational intractability: polynomial-time algorithm for sampling.



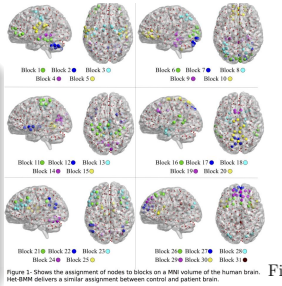
→ Theoretical development was guided by simulations.

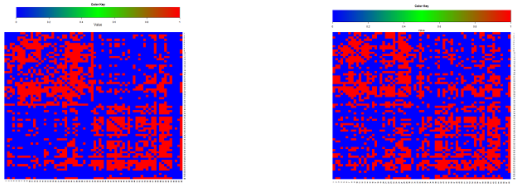
→ New algorithm was the key.

Stochastic blockmodels: community-based modeling

What if...

nodes in the network are naturally grouped?





(a) Subject 1 (b) Subject 8

Fig 8: Heatmap of the adjacency matrices; blue pixels indicate 0 and red pixels indicate 1

Connections in the brain modeled as a network: regions are nodes, edges exist if regions correlate.

Stochastic blockmodels: community-based modeling

What if...

nodes in the network are naturally grouped?

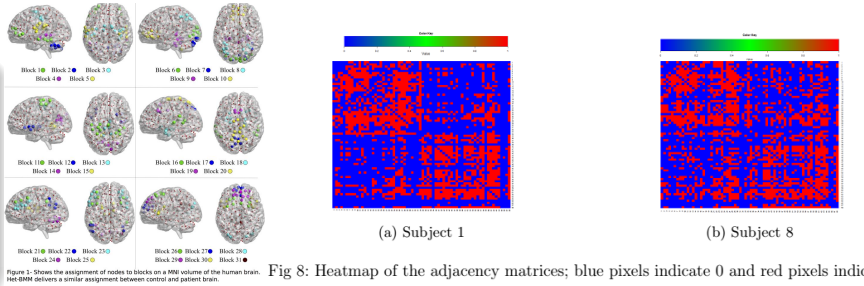


Fig 8: Heatmap of the adjacency matrices; blue pixels indicate 0 and red pixels indicate 1

Connections in the brain modeled as a network: regions are nodes, edges exist if regions correlate.

← MRC 2016 network models working group derived:

The first exact (non-asymptotic) goodness-of-fit test for network models where nodes in groups/blocks/communities that are unknown.

Simulations on connectome data show that the blockmodel based on degrees fits, while the one based on edge counts does not.

[*Exact tests for stochastic blockmodels*, arXiv:1612.06040.]

Getting started: kinds of questions we ask - Part II

Guiding question:

Can we use randomization to study algebraic structures?

A computational algebraist's interest in randomness is motivated by a search for improved average-time algorithms for solving systems of polynomial equations with special structure.

A commutative algebraist's interest in randomness is motivated by a search for 'typical' or 'average' examples of algebraic structures, in contrast to 'extreme' examples.

Random sampling in computational algebra

Framework 1 [De Loera, Petrović, Stasi: JSC '16]

Randomized algorithms for computing with polynomials.

- Systems of polynomial equations are ubiquitous in optimization, statistics, biology, and other fields in science and engineering.
- Solving them is a cornerstone of computational algebra today, but it is well-known that many algorithms have high worst-case complexity.
- We construct new randomized algorithms for polynomial ideals (e.g. solving polynomial systems, computing small/minimal generating sets) that has expected runtime linear in the number of input polynomials.

Random sampling in computational algebra

Framework 1 [De Loera, Petrović, Stasi: JSC '16]

Randomized algorithms for computing with polynomials.

- Systems of polynomial equations are ubiquitous in optimization, statistics, biology, and other fields in science and engineering.
- Solving them is a cornerstone of computational algebra today, but it is well-known that many algorithms have high worst-case complexity.
- We construct new randomized algorithms for polynomial ideals (e.g. solving polynomial systems, computing small/minimal generating sets) that has expected runtime linear in the number of input polynomials.

Random monomial ideals

Framework 2 [De Loera, Petrović, Silverstein, Stasi, Wilburne: 2017$^+$]

A framework for studying random algebraic structures: monomial ideals.

- The Erdös-Renýi-type model for random monomial ideals generalizes the basic models for random graphs and random simplicial complexes.
- Allows for the study of average-case behavior: we can get a handle on both the average and the extreme behavior of these random ideals and how various ranges of the probability parameter control those properties.
- This provides a toolbox to search for interesting examples and formulate conjectures. A `Macaulay2` package is in the works in 2017.

Getting started, revisited:

Basic courses - background:

Math 476 / Math 563 (statistics) $\leftarrow$ Part I
Math 431 / Math 530 (computational algebra) $\leftarrow$ Part II

Seminars and other background:

Algebraic statistics seminar (approx. weekly, except current semester)
Graph theory courses
Computational statistics and using/programming in `R`, `Python`, ...
Discrete mathematics seminar

As for any research project, papers/preprints and a couple of books for background will always be recommended reading.