Discrete methods for statistical network analysis:
Exact tests for goodness of fit for network data
Key player: sampling algorithms

Sonja Petrović

Illinois Institute of Technology
Chicago, USA

Joint work with
Despina Stasi (Illinois Institute of Technology)
Elizabeth Gross (San Jose State University)

Annals of the Institute of Statistical Mathematics (AISM) 2016
and
AMS CONM book chapter 2016
(Proceedings of the AMS Special Session on Algebraic and Geometric Methods in Discrete Mathematics)


Two papers that define the p_1 model and discuss its interpretation as a log-linear model on contingency tables:

- An Exponential Family of Probability Distributions for Directed Graphs: Comment Stephen E. Fienberg and Stanley Wasserman JASA 1981 DOI: 10.2307/2287039
- An Exponential Family of Probability Distributions for Directed Graphs JASA 1981 Paul W. Holland and Samuel Leinhardt DOI: 10.2307/2287037

Here is the data set I will be analyzing:

- NYT article (see the figure titled "Inside Japan Inc.").

The Fundamental Theorem of Markov Bases appears here (see also Sullivant book and the MFO Lectures book):

- Algebraic algorithms for sampling from conditional distributions, AOS 1998, Persi Diaconis and Bernd Sturmfels.

References on the toric ideals of graphs:

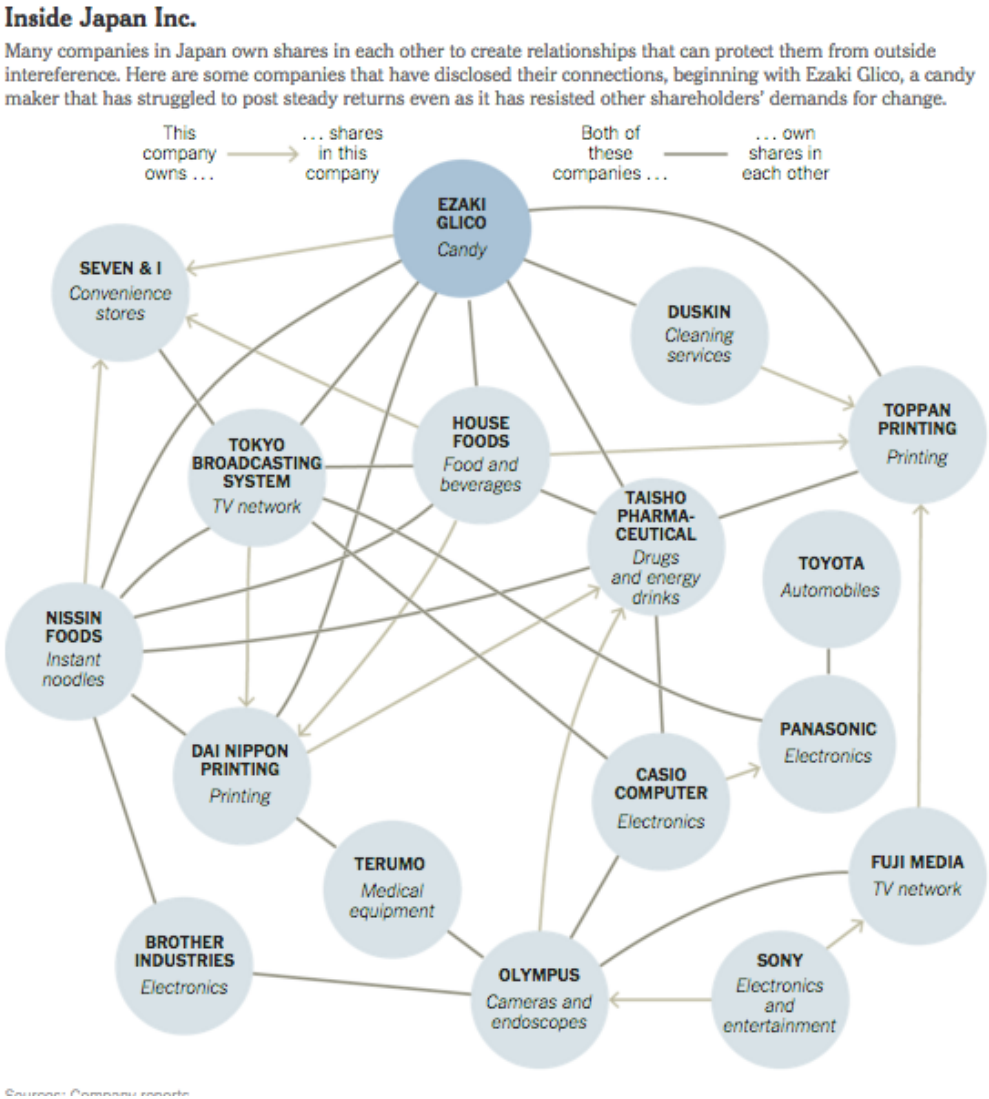
- Toric Ideals Generated by Quadratic Binomials, JAlg 1999, Hidefumi Ohsugi and Takayuki Hibi
- Minimal generators of toric ideals of graphs, Adv. Appl. Math 2010, Enrique Reyes, Christos Tatakis, Apostolos Thoma
- Rees algebras of edge ideals, Comm Alg 1995, Rafael Villarreal

References for further results on Markov bases and sampling constraints:

- Distance-reducing Markov bases, Bernoulli 2005, Akimichi Takemura and Satoshi Aoki.
- Chp6 of Markov bases in algebraic statistics, Springer textbook, Satoshi Aoki, Hisayuki Hara, Akimichi Takemura.

$\rightarrow$ For testing the p_1 model using this methodology, see arXiv:1401.4896.
$\rightarrow$ A survey of discrete methods in (algebraic) statistics for networks is here: arXiv:1510.02838.

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 send/receive links..

Claim:

Strong reciprocation effect among the corporate directorates.

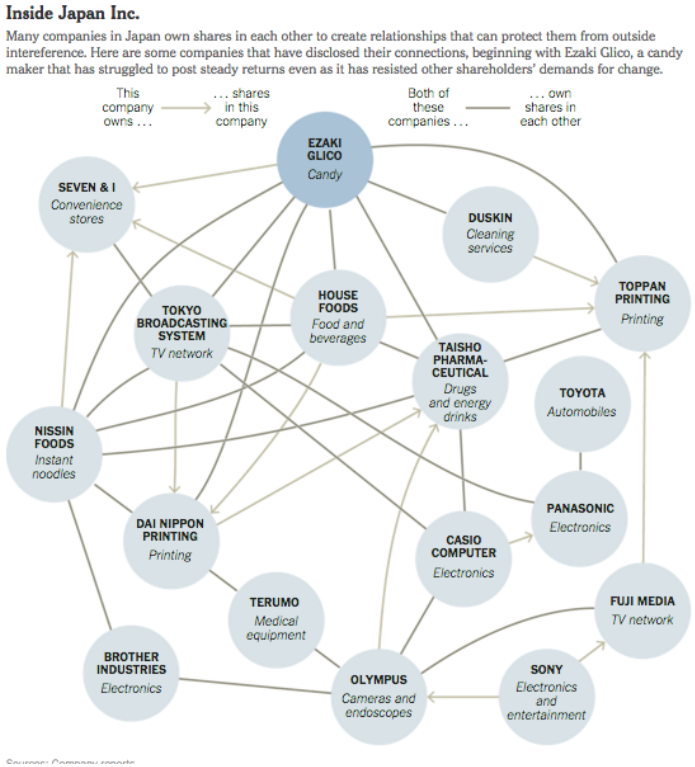
→A satisfying answer will confirm model/data fit for a statistical model that captures this effect.←

Network data: new frontier for statistical inference

Challenges:

1. Defining good network models?
2. Principled statistical inference?
3. Sample size 1;
   number of parameters increasing?
4. Scaling inference and model fitting
   questions to large networks?

One recent motivating example:



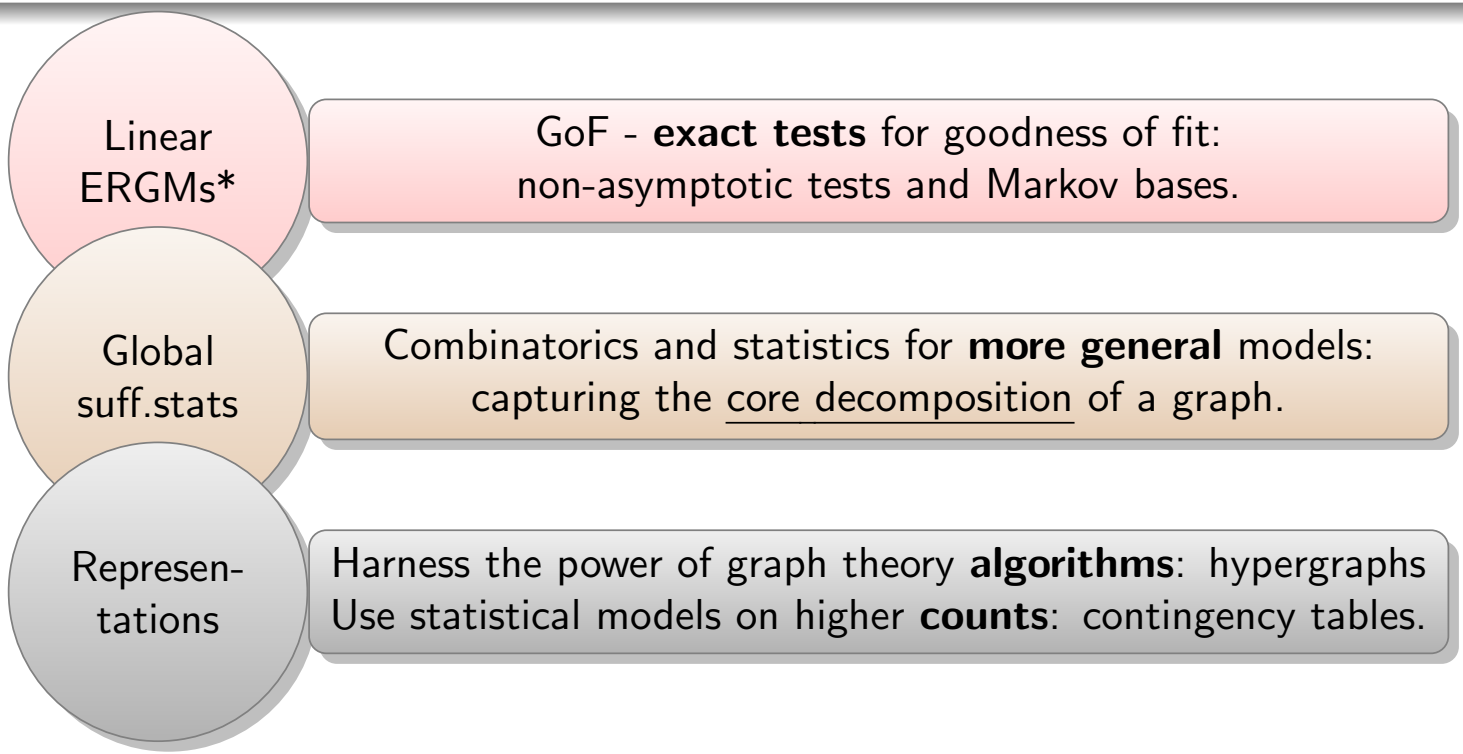
New York Times 2014

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

State of the art: significant challenges for inference

Goodness-of-fit tests - heuristic, model diagnostics
(clustering coeff., triangle count,... Comparison: observed $\leftrightarrow$ simulated)

Hunter, Goodreau, Handcock '08,
Goldenberg, Zheng, Fienberg, Airoldi '09 [FTML].



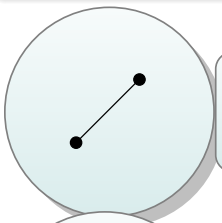
*Can be embedded in a community-based modeling framework, e.g., used within/between blocks in a SBM. [Ongoing work & Karwa+2016+]

ERGMs: model family test-bed

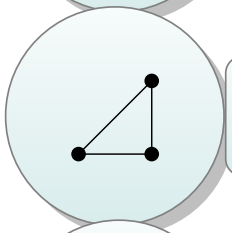
Specify informative network statistics on $\mathcal{G}_n$ (capture key features)

$$t : \mathcal{G}_n \to \mathbb{R}^d, \quad g \mapsto t(g) = (t_1(g), \ldots, t_d(g)) \in \mathbb{R}^d,$$

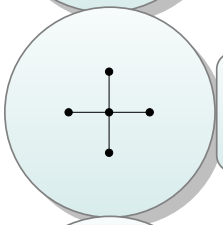
such that $P_\theta(G = g) = \exp\{\theta^T t(g) - \psi(\theta)\}$.



Erdős-Renyi-Gilbert model



Total number of triangles



k-stars

degen +edge

Degeneracy + number of edges [Kim et al. '16]

Deg. seq.

β model [Rinaldo, P., Fienberg 2011; Chatterjee, Diaconis '11] + hypergraphs[SSPFR'15]

Bi-deg. seq.

p_1 model [Holland, Leinhardt '81]

Trian-gle+star

Markov graph model [Frank, Strauss '86]

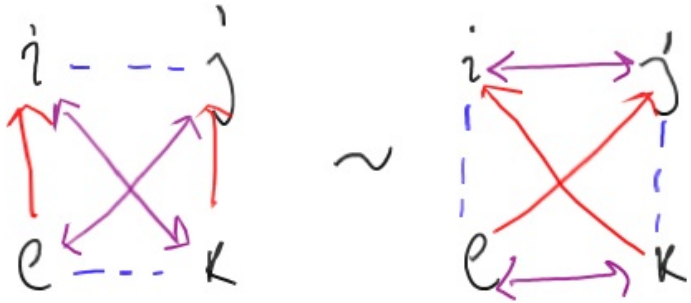
Shell disrib.

k-core ERGM [Karwa, Pelsmajer, P., Stasi, Wilburne '16]

GoF test with dynamic Markov bases

Exact conditional test

Observed g is compared to a fiber $\mathcal{F}_{t(g)}$: set of all possible g' with $t(g') = t(g)$.

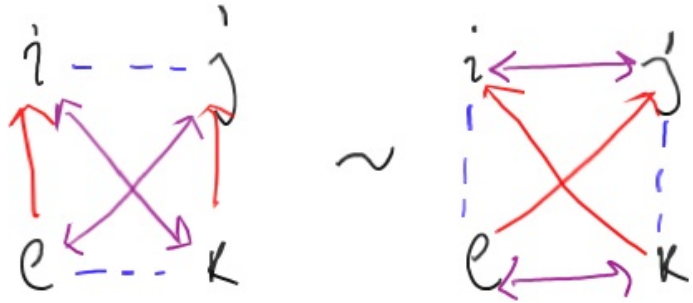


$\uparrow\uparrow$ Calculate a (valid) GoF statistic for each g' (e.g. chi-square statistic)

GoF test with dynamic Markov bases

Exact conditional test

Observed g is compared to a fiber $\mathcal{F}_{t(g)}$: set of all possible g' with $t(g') = t(g)$.



$\uparrow\uparrow$ Calculate a (valid) GoF statistic for each g' (e.g. chi-square statistic)

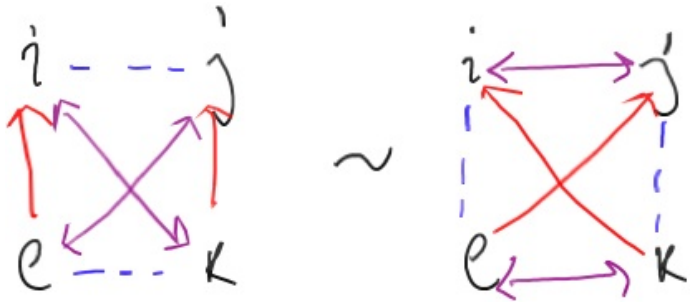
Challenges:

1. Computing a Markov basis
   - Pre-computing all (!) moves (Diaconis-Sturmfels '98) - conditional distributions on contingency tables
   - Structural results for many families (many, many authors...)
2. Applicability
   - MB are data independent! [Dobra et al.]; suggests generating only applicable moves, one at a time.
   - ... Is it true?: algebra+mixing=too many problems to solve?

GoF test with dynamic Markov bases

Exact conditional test

Observed g is compared to a fiber $\mathcal{F}_{t(g)}$: set of all possible g' with $t(g') = t(g)$.



↑↑ Calculate a (valid) GoF statistic for each g' (e.g. chi-square statistic)

For the three variants of the p_1 random graph model, $t(g)$ is:

1) $\rho_i + \rho_j$	in-degrees and out-degrees of nodes + reciprocated degrees
2) $\rho_{ij} = \rho$	in-degrees and out-degrees of nodes + number of reciprocated edges
3) $\rho_{ij} = 0$	in-degrees and out-degrees of nodes; where reciprocated edges are not 'special'.

We can draw upon graph-theory literature for more results on this chain:

GoF test with dynamic Markov bases

[Gross-P.-Stasi, Ann. ISM 2016; 2017^+ work]

Dynamic generation of all applicable Markov moves for p_1 models. def. p_1

- Data-dependent, small/large moves, complete Markov chain graph.
- Translated: (multiplicities OK.)
  Network (or any table) $g \leftrightarrow$ set $e(g)$ of edges of **model hypergraph**
  Sufficient statistics $t(g) \leftrightarrow$ **degree sequence** of $e(g)$.

$\downarrow$gain$\downarrow$

Theorem [Dillon 2016]

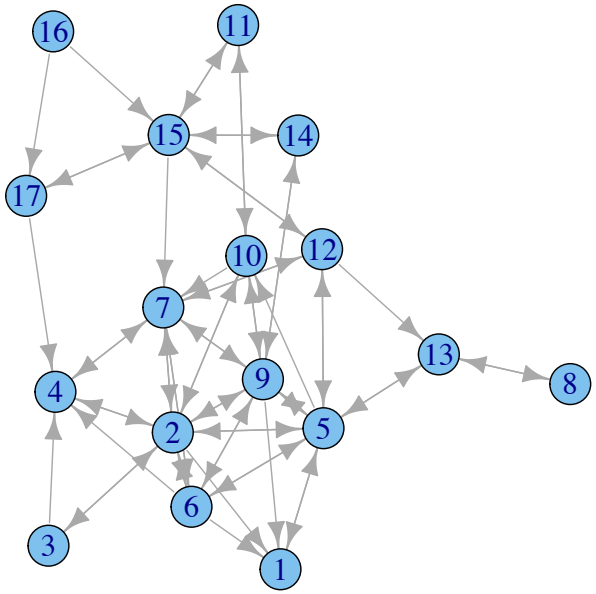
The dynamic Markov bases chain mixes rapidly for all fibers where graph-theoretic simple-switch chain mixes rapidly.

Rapid mixing:

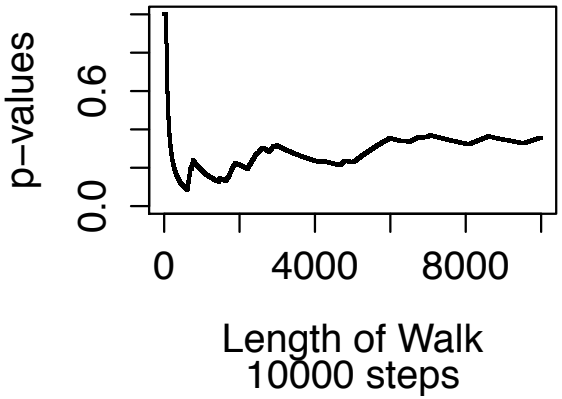
Minimum probability of escaping, any subset of the fiber graph is bounded below by $1/poly(n)$.

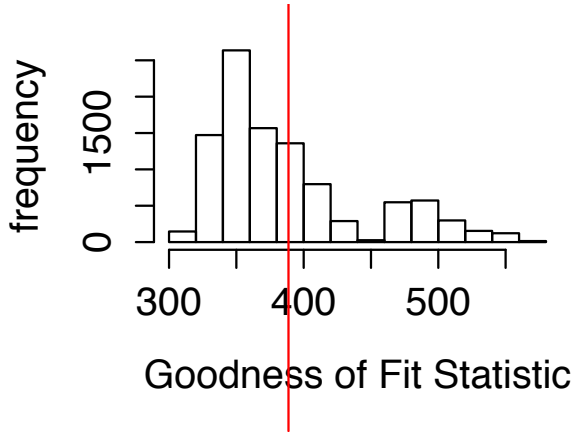
Fast mixing for large classes of deg.seq.: Erdös, Miklós, Toroczkai, 2016.

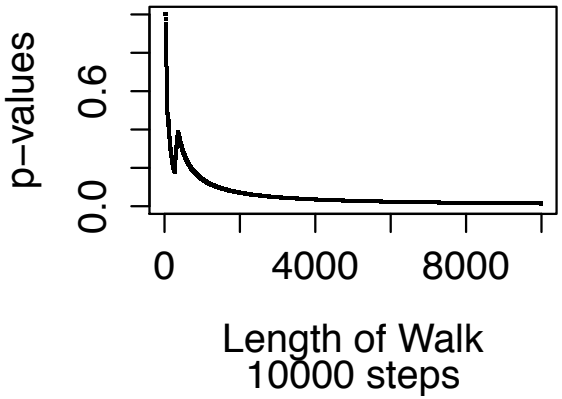
Application 1: Testing for reciprocation effect

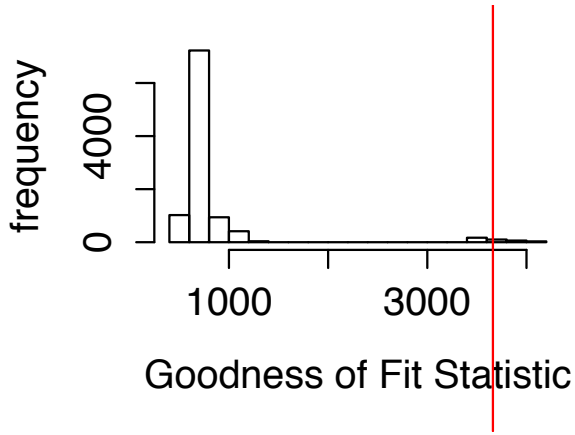


1- Seven 2- Ezaki 3- Duskin 4- Toppan 5- Tokyo Broadcasting 6- House foods 7- Taisho pharma. 8- Toyota 9- Nissin 10- Dai Nippon Print 11- Terumo 12- Casio 13- Panasonic 14- Brother 15- Olympus 16- Sony 17- Fuji.









Reject the zero reciprocation p_1 model:

Statistical evidence in support of Prime Minister Shinzo Abe's claim. (!)

ADVERTISEMENTS

This is YOUR journal

JOURNAL OF ALGEBRAIC STATISTICS
http://jalgstat.library.iit.edu/

AS2020 conference

Algebraic Statistics 2020
University of Hawai'i at Mãnoa, Honolulu HI
stay tuned for more information