Markov moves from classical algebraic constructions
(Algebraic statistics and large sparse data sets)

Sonja Petrović

Department of Mathematics, Statistics, and Computer Science
University of Illinois at Chicago

Joint work with Alessandro Rinaldo and Stephen Fienberg, CMU

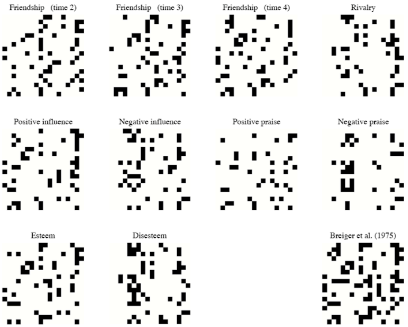
Algebraic Statistics Minisymposium
SIAM meeting, Pittsburgh

July 13, 2010

- Many of the most active areas of statistical research involve large sparse data problems where the number of variables and/or parameters is large, especially relative to the number of independent obser vations.
- Standard statistical theory for estimation and results related to asymptotic behavior often fail in such settings.
- The computational tools associated with algebraic statistics are useful often only for low-dimensional problems, e.g., involving a small number of parameters.
- Upshot: algebraic statistical and the related computational tools can nonetheless provide important insights of value in large sparse contingency table and network settings.

Example: Monks in a monastery

- 18 novices observed over two years.
- Network data gathered at 4 time points; and on multiple relationships.



See analyses in Airoldi, Blei, Fienberg, Xing. (2008) Mixed membership stochastic block models. J. of Machine Learning Research.

Example: The Collective Dynamics of Smoking in a Large Social Network (James Fowler)

Node border= gender (red=female, blue=male). Arrow color = relation (purple=friend, green=spouse). Node color = smoking behavior (white=nonsmoker, gray=smoker); darker shades = more cigarettes consumed per day.

The p_1 random graph model (Holland-Leinhardt)

- n nodes, random occurrence of directed edges.

The p_1 random graph model (Holland-Leinhardt)

- n nodes, random occurrence of directed edges.

- Describe the probability of an edge occurring between nodes i and j:

  $\log Prob(\text{no edge}) = \lambda_{ij}$
  $\log Prob(\text{from } i \text{ to } j) = \lambda_{ij} + \alpha_i + \beta_j + \theta$
  $\log Prob(\text{from } j \text{ to } i) = \lambda_{ij} + \alpha_j + \beta_i + \theta$
  $\log Prob(\text{bi-directed edge}) = \lambda_{ij} + \alpha_i + \beta_j + \alpha_j + \beta_i + 2\theta + \rho_{ij}$

- Parameters:
  - λ_{ij} is a normalizing constant
  - α_i represents node i sending an edge
  - β_i represents node j receiving an edge
  - ρ_{ij} represents the reciprocation effect (3 common forms:

    $\rho_{ij} = 0$,
    $\rho_{ij} = \rho$ constant,
    $\rho_{ij} = \rho + \rho_i + \rho_j$ edge-dependent).

Estimation for p_1

- The likelihood function for the p_1 model is clearly of exponential family form.
- Holland-Leinhardt explored goodness of fit of model empirically by comparing $\rho_{ij} = 0$ vs. $\rho_{ij} = \rho$.
- The problem is that standard asymptotics (normality and chi-squared goodness of fit tests) aren't applicable as the number of parameters increases with the number of nodes.
- Fienberg and Wasserman used the edge-dependent reciprocation model to test $\rho_{ij} = \rho$.
- For a review of these and related models, see: Goldenberg, Zheng, Fienberg, Airoldi. (2010) "A Survey of Statistical Network Models".

The problem:

- Describe a Markov basis for n-node network for large n. (Describe the corresponding toric variety implicitly.)

A classical construction:

- Edge subring of a graph G (or: *toric ring of* G):
- generated by the edges of the G:
- For $G = K_{n,m}$ with vertex sets $\alpha_1, \ldots, \alpha_n$ and $\beta_1, \ldots, \beta_n$, the edge subring is the image of the map

$$p_{ij} \mapsto \alpha_i \beta_j.$$

- The defining ideal is the kernel of this map:
- an example of an element in the ideal is $p_{12}p_{34} - p_{14}p_{32}$.

p_1 model as a toric variety

- To each pair of nodes and edge type we associate a monomial in the model parameters: $p_{12}(1,1) \mapsto \lambda_{12}\alpha_1\beta_2\alpha_2\beta_1\theta^2\rho_{12}$ represents a bi-directed edge between 1 and 2.

p_1 model as a toric variety

- To each pair of nodes and edge type we associate a monomial in the model parameters: $p_{12}(1,1) \mapsto \lambda_{12}\alpha_1\beta_2\alpha_2\beta_1\theta^2\rho_{12}$ represents a bi-directed edge between 1 and 2.

- The monomial map $C[p_{ij}(a,b)] \to C[\lambda_{ij}, \alpha_i, \beta_i, \theta, \rho_{ij}]$

$$p_{ij}(a,b) \mapsto \lambda_{ij}\alpha_i^a\alpha_j^b\beta_i^b\beta_j^a\theta^{a+b}\rho_{ij}^{min(a,b)}$$

  parametrizes a toric variety, whose design matrix $\mathcal{A}_n$ has:
  - $4\binom{n}{2}$ columns (variables),
  - $\binom{n}{2} + 2n + 1$ rows (parameters) if $\rho_{ij} = 0$;
    $2\binom{n}{2} + 2n + 2$ if $\rho_{ij} = \rho + \rho_i + \rho_j$.

- The kernel of the map (matrix) defines a toric ideal, whose generating set is a Markov basis (Diaconis-Sturmfels '98).

- For $n = 3$ and ρ_{ij} edge-dependent:

\- the design matrix is a rank-11 14×12 matrix

\- the variety is a cubic hypersurface in $\mathbb{P}^{11}$.

3-node network

- Markov bases connect all networks with same sufficient statistics (in- and out- degrees of the nodes).
- For all 3 cases of ρ_{ij}, there is only one Markov move:

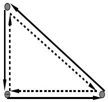


Figure: dashed edges are replaced by full edges.

- remove edges $1 \to 2$, $2 \to 3$ and $3 \to 1$
  replace them by edges $2 \to 1$, $3 \to 2$ and $1 \to 3$.

- This move is represented by the binomial:

$$p_{12}(1,0)p_{23}(1,0)p_{13}(0,1) - p_{12}(0,1)p_{23}(0,1)p_{13}(1,0).$$

Simplification of p_1 and toric ring of a graph

By ignoring normalizing constants λ_{ij} we get a simplified model:

Theorem (P.-Rinaldo-Fienberg)

If $\rho_{ij} = 0$, the ideal of the simplified model equals $I_{G_n} + T_n$ where T_n is generated by $p_{ij}(1,0)p_{ij}(0,1) - p_{ij}(1,1)$ and I_{G_n} is the toric ideal of the edge subring of $G_n := K_{n,n} \setminus \{i,i\}$.

Theorem (P.-Rinaldo-Fienberg)

If $\rho_{ij} = \rho + \rho_i + \rho_j$, the ideal of the simplified model equals $I_{G_n} + Q_n$ where I_{G_n} is as above, and Q_n is the toric ideal of the edge subring of K_n.

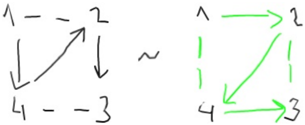
Simplification of p_1 and toric ring of a graph, II

An example with 4 nodes

- What is I_{G_n}? Its generators have a nice description in terms of paths:

Simplification of p_1 and toric ring of a graph, II

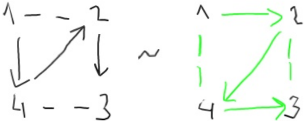
An example with 4 nodes



$$p_{12}(0,0)p_{14}(1,0)p_{23}(1,0)p_{24}(0,1)p_{34}(0,0) -$$
$$p_{12}(1,0)p_{14}(0,0)p_{23}(0,0)p_{24}(1,0)p_{34}(0,1)$$

Simplification of p_1 and toric ring of a graph, II

An example with 4 nodes

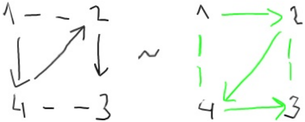


$p_{14}(1,0)p_{23}(1,0)p_{24}(0,1)$ $-$

$p_{12}(1,0)$ $p_{24}(1,0)p_{34}(0,1)$

Simplification of p_1 and toric ring of a graph, II

An example with 4 nodes



$$p_{14}(1,0)p_{23}(1,0)p_{24}(0,1) \quad - \quad p_{12}(1,0) \quad p_{24}(1,0)p_{34}(0,1)$$

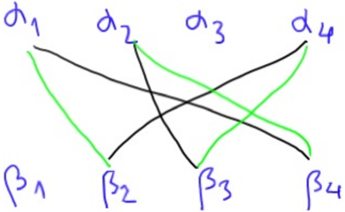


Figure: the corresponding path in $K_{4,4} \backslash \{i,i\}$

Toric ideal of the p_1 model

- Incorporate λ_{ij} into the previous theorems:

Theorem (P.-Rinaldo-Fienberg)

The toric ideal of the p1 random graph model is the multi-homogenous piece of the toric ideal of the simplified model.

By multi-homogeneous, we mean with respect to each pair i, j.

Toric ideal of the p_1 model

- Incorporate λ_{ij} into the previous theorems:

Theorem (P.-Rinaldo-Fienberg)

The toric ideal of the p1 random graph model is the multi-homogenous piece of the toric ideal of the simplified model.

By multi-homogeneous, we mean with respect to each pair i, j.

- We claim that homogenizing simple moves appropriately produces the whole Markov basis for the model:

Conjecture

Minimal Markov (Gröbner) bases for the p_1 models can be obtained from Markov (Gröbner) bases of the simplified model by repeated lifting and overlapping of the binomials in the minimal Markov bases of various $(n-1)$-node subnetworks.

- N-fold structure of the design matrices

Network model challenges

- How to use algebraic statistics results for (1) existence of MLEs and (2) to assess fit of p_1 to large-scale network settings?
- Linking algebraic statistics for loglinear models to results for p_1 .
- Extending results from p_1 to Exponential Random Graph Models.
- Algebraic statistics for mixed-membership stochastic block models.

Why reparametrize?, redundancy, symmetry

- Fienberg-Wasserman: p_1 model is a $n^2 \times 2 \times 2$ contingency table (n^2 dyads, 2×2 configurations)
- Highly redundant! Undesirable for finding Markov bases: $4n^2$ indeterminates instead of $2n(n-1)$. (OK for MLE.)
- Number of generators explodes combinatorially: for the case of constant reciprocation, $\rho_{ij} = \rho$, the ideal of the network on $n = 3$ nodes has 107 minimal generators, and the one of the 4-node network has $80,610$.
- Non-applicable Markov basis elements; symmetries
- We were able to analyze the $n = 4$ case and reduce all of the $80,610$ moves to the ones we get using our design matrices, but the effort was nontrivial.
- Therefore, at least from the point of view of studying Markov bases, the parametrization we are using is preferable.

Revealing "simple" moves

The following degree-five binomial appears as a minimal generator of the ideal of a 4-node network:

$p_{li}(1,0)p_{ij}(1,0)p_{jk}(1,0)p_{lj}(0,0)p_{ik}(0,0)-$
$p_{li}(0,0)p_{ij}(0,1)p_{jk}(0,0)p_{lj}(1,0)p_{ik}(1,0)$

This move can be obtained by the following sequence of simple moves:

replace the edges (l,i) and (j,k) by the edges (l,k) and (j,i)

followed by

replace the edges (i,j) and (l,k) by the edges (i,k) and (l,j)..

Figure: A sequence of two moves on 4 nodes: dashed edges are replaced by full edges.

Simple moves

- In fact, for $n = 3, 4, 5$, we can get all Markov moves in our list as decompositions of these simpler moves!

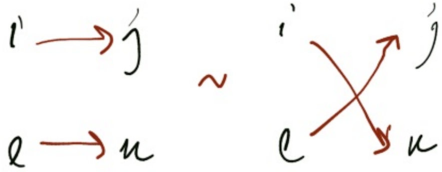


Figure: An essential, simple move

- Bidirected edges appear in this same pattern in all Markov moves. These are generators of Q_n from Theorem 2.

Next?

- Prove the "homogenization" claim in terms of generators!
- Prove the decomposition to simple moves, even though they are not in the toric ideal as defined.
- This decomposition would identify precisely the Markov moves in this setting with moves of Holland-Leinhardt in some cases.

Thank you for your attention!

Reference: Petrović, Rinaldo, Fienberg.
" Algebraic statistics for a directed random graph model with reciprocation. " AMS CONM Series volume on Algebraic Methods in Statistics and Probability.
arXiv:0909.0073v2