Algebraic Statistics of p_1 Network Models: Markov Bases and Their Uses

Stephen E. Fienberg

Department of Statistics and Machine Learning Department
Carnegie Mellon University

Joint work with:
Alessandro Rinaldo, Carnegie Mellon University
and Sonja Petrović, University of Illinois at Chicago

Algebraic Methods in Machine Learning
NIPS Workshop December 2008

December 11, 2008

Context

Context

- Networks are ubiquitous in science

Context

- Networks are ubiquitous in science
- Network graphs $\neq$ graphical model graphs

Context

- Networks are ubiquitous in science
- Network graphs $\neq$ graphical model graphs
- Statistical models for networks:

Context

- Networks are ubiquitous in science
- Network graphs $\neq$ graphical model graphs
- Statistical models for networks:

  New statistical problems (standard asymptotics do not hold—how do we get approximate normality?)

Context

- Networks are ubiquitous in science
- Network graphs $\neq$ graphical model graphs
- Statistical models for networks:

  New statistical problems (standard asymptotics do not hold—how do we get approximate normality?)

- p_1 model, log-linear model, and algebraic statistics:

Context

- Networks are ubiquitous in science
- Network graphs $\neq$ graphical model graphs
- Statistical models for networks:

  New statistical problems (standard asymptotics do not hold—how do we get approximate normality?)

- p_1 model, log-linear model, and algebraic statistics:

  Using algebra-geometry structure to “understand” a family of network models and frame a related set of statistical estimation and inference issues.

General framework for networks evolving over time

- *Network represented by a graph* G_t *at time* t

General framework for networks evolving over time

- *Network represented by a graph* G_t *at time* t
- Nodes and edges can be created and can die

General framework for networks evolving over time

- *Network represented by a graph G_t at time t*
- Nodes and edges can be created and can die
- Can have multiple relationships; attributes for nodes and/or edges

General framework for networks evolving over time

- *Network represented by a graph* G_t *at time* t
- Nodes and edges can be created and can die
- Can have multiple relationships; attributes for nodes and/or edges
- Edges can be directed or undirected

General framework for networks evolving over time

- *Network represented by a graph G_t at time t*
- Nodes and edges can be created and can die
- Can have multiple relationships; attributes for nodes and/or edges
- Edges can be directed or undirected
- Data available to be observed at time $t = t_0$

General framework for networks evolving over time

- *Network represented by a graph* G_t *at time* t
- Nodes and edges can be created and can die
- Can have multiple relationships; attributes for nodes and/or edges
- Edges can be directed or undirected
- Data available to be observed at time $t = t_0$
- *Underlying stochastic process that describes the network structure and evolution.*
  - *Come to my talk on dynamic network modeling this afternoon in Grpahical models Workshop!*

Example: The Framingham "Obesity" Study

- Original Framingham "sample" cohort begun in 1947.
- Offspring cohort of $N_0 = 5,124$ individuals measured beginning in 1971 for $T = 7$ epochs centered at 1971, 1981, 1985, 1989, 1992, 1997, 1999.
- Link information on family members and one close friend.
- Total number of individuals on whom we have obesity measures is N=12,067.
- Details in Christakis and Fowler, *NEJM*, July 2007.

Example: The Framingham "Obesity" Study

- Original Framingham "sample" cohort begun in 1947.
- Offspring cohort of $N_0 = 5,124$ individuals measured beginning in 1971 for $T = 7$ epochs centered at 1971, 1981, 1985, 1989, 1992, 1997, 1999.
- Link information on family members and one close friend.
- Total number of individuals on whom we have obesity measures is N=12,067.
- Details in Christakis and Fowler, *NEJM*, July 2007.

Example: The Framingham "Obesity" Study

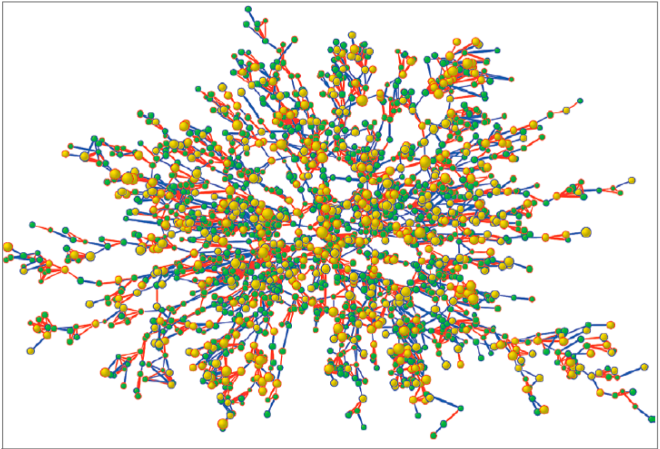
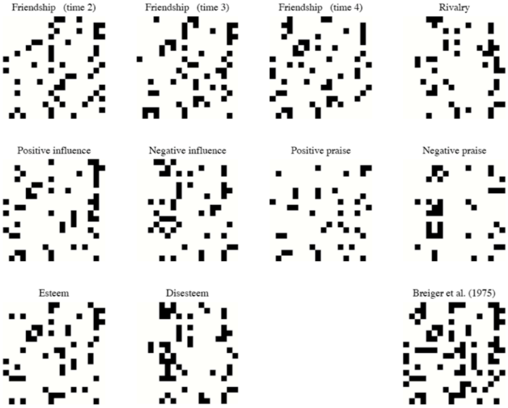


Figure 1. Largest Connected Subcomponent of the Social Network in the Framingham Heart Study in the Year 2000.
Each circle (node) represents one person in the data set. There are 2200 persons in this subcomponent of the social network. Circles with red borders denote women, and circles with blue borders denote men. The size of each circle is proportional to the person's body-mass index. The interior color of the circles indicates the person's obesity status: yellow denotes an obese person (body-mass index, ≥30) and green denotes a nonobese person. The colors of the ties between the nodes indicate the relationship between them: purple denotes a friendship or marital tie and orange denotes a familial tie.

Example -The Collective Dynamics of Smoking in a Large Social Network (J.Fowler)

Example: Monks in a Monastery

- 18 novices observed over two years.
- Network data gather at 4 time points; and on multiple relationships..
- See analyses in Airoldi, et al., (2008)



Holland and Leinhardt's p_1 model

- n nodes, random occurrence of directed edges.

Holland and Leinhardt's p_1 model

- n nodes, random occurrence of directed edges.
- Describe the probability of an edge occurring between nodes i and j:

Holland and Leinhardt's p_1 model

- n nodes, random occurrence of directed edges.
- Describe the probability of an edge occurring between nodes i and j:

$$\log Prob(\text{no edge}) = \log P_{ij}(0,0) = \lambda_{ij}$$

$$\log Prob(\text{from } i \text{ to } j) = \log P_{ij}(1,0) = \lambda_{ij} + \alpha_i + \beta_j + \theta$$

$$\log Prob(\text{from } j \text{ to } i) = \log P_{ij}(0,1) = \lambda_{ij} + \alpha_j + \beta_i + \theta$$

$$\log Prob(\text{bi-directed edge}) = \log P_{ij}(1,1) = \lambda_{ij}+\alpha_i+\beta_j+\alpha_j+\beta_i+2\theta+\rho_{ij}$$

Holland and Leinhardt's p_1 model

- n nodes, random occurrence of directed edges.
- Describe the probability of an edge occurring between nodes i and j:

$$\log Prob(\text{no edge}) = \log P_{ij}(0,0) = \lambda_{ij}$$

$$\log Prob(\text{from } i \text{ to } j) = \log P_{ij}(1,0) = \lambda_{ij} + \alpha_i + \beta_j + \theta$$

$$\log Prob(\text{from } j \text{ to } i) = \log P_{ij}(0,1) = \lambda_{ij} + \alpha_j + \beta_i + \theta$$

$$\log Prob(\text{bi-directed edge}) = \log P_{ij}(1,1) = \lambda_{ij}+\alpha_i+\beta_j+\alpha_j+\beta_i+2\theta+\rho_{ij}$$

- 3 common forms:

  $\rho_{ij} = 0$ (no reciprocal effect)
  $\rho_{ij} = \rho$ (constant reciprocation factor)
  $\rho_{ij} = \rho + \rho_i + \rho_j$ (edge-dependent reciprocation)

Estimation for p_1

- The likelihood function for the p_1 model is clearly of exponential family form.
- For the constant reciprocation version, we have

$$\log p_1(x) \propto x_{++}\theta + \sum_i x_{i+}\alpha_i + \sum_j x_{+}j\beta_j + \sum_{ij} x_{ij}x_{ji}\rho \qquad (1)$$

- Holland-Leinhardt explored goodness of fit of model empirically.
  - They compared $\rho_{ij} = 0$ vs. $\rho_{ij} = \rho$.
  - The problem is that standard asymptotics (normality and chi-sqare tests of fit) are not applicable as the number of parameters increases with the number of nodes.
  - How to test $\rho_{ij} = \rho$ against a more complex model?

Aside—The Normal Distribution was 275 Years Old on November 12, 2008!

Nov. 12. 1733.

APPROXIMATIO AD

SUMMAM TERMINORUM BINOMII

$\overline{a+b}^n$ in Seriem expansi,

Autore *A. D. M.* R. S. S.

QUanquam solutio Problematum ad sortem spectantium non raro exigit ut plures Termini Binomii $\overline{a+b}^n$ in summam colligantur; attamen in potestatibus excelsis res adeo laboriosa videtur, ut perpauci hoc opus aggredi curaverint; *Jacobus* & *Nicolaus Bernoulli* viri Doctissimi primi quod sciam tentarunt quid sua industria in hoc genere præstare posset, in quo etiamsi uterque propositum summa cum laude sit assecutus, aliquid tamen ultra potest requiri, hoc est approximatio ad summam; non enim tam de approximatione videntur fuisse solliciti quam de assignandis certis limitibus quos Summa Terminorum necessario transcenderet. Quam vero viam illi tenuerint, breviter in Miscellaneis meis exposui * quæ consulat Lector si vacat, quod ipsi tamen scripserint melius erit fortasse consulere: Ego quoque in hanc disquisitionem incubui; quod autem eo me primum impulit non profectum fuit ab opinione me cæteros anteiturum, sed ab obsequio in Dignissimum virum qui mihi autor fuerat ut hæc susciperem; Quicquid est, novas cogitationes prioribus subnecto, sed eò ut connexio postremorum cum primis melius appareat, mihi necesse est ut pauca jampridem a me tradita denuo proferam.

I. Duodecim jam sunt anni & amplius cum illud inveneram; si Binomium $1+1$ ad potestatem n permagnam attollatur, ratio quam Terminus Medius habet ad summam Terminorum omnium, hoc est ad 2^n, ad hunc modum poterit exprimi $\frac{2A \times \overline{n-1}^n}{n^n \sqrt{n-1}}$, ubi A eum numerum exponit cujus Logarithmus

A

* Vide Miscellanea Analytica pag. 96.

[2]

rithmus hyperbolicus est $\frac{1}{12} - \frac{1}{360} + \frac{1}{1260} - \frac{1}{1680}$ &c. quam seriem ad libitum mihi continuare licuit, sed quia quantitas $\frac{\overline{n-1}^n}{n^n}$ seu $\overline{1-\frac{1}{n}}^n$ fere data sit, quod consideranti facile patebit, sequitur in Potestate infinita quantitatem illam datum iri, eamque exhibiturum numerum illum cujus Logarithmus hyperbolicus est -1; hinc fiet ut si B designarit numerum illum cujus Logarithmus hyperbolicus est $-1 + \frac{1}{12} - \frac{1}{360} + \frac{1}{1260} - \frac{1}{1680}$ &c. expressio suprascripta evasura sit $\frac{2B}{\sqrt{n-1}}$ seu $\frac{2B}{\sqrt{n}}$, atque adeo si signa seriei mutentur, ponaturque B æqualis numero cujus Logarithmus hyperbolicus est $1 - \frac{1}{12} + \frac{1}{360} - \frac{1}{1260} + \frac{1}{1680}$ &c. expressio illa futura sit $\frac{2}{B\sqrt{n}}$.

Cum primum ad hanc disquisitionem animum appuli, eo contentus fui ut præter propter valorem quantitatis B determinarem quod quidem factum fuerat additione paucorum hujus seriei Terminorum, quorum summa spectata fuerat tanquam Logarithmus istius quantitatis, attamen tarditas convergentiæ me deterruerat quominus longius procederem, donec vir Doctissimus mihique amicissimus *Jacobus Stirling* qui post me ad hanc disquisitionem methodo a mea valde diversa se contulit, comperit quantitatem B denotare radicem quadratam circumferentiæ Circuli cujus radius est unitas, ita ut si hæc circumferentia appelletur c, ratio Medii Termini ad Terminos omnes exprimetur per $\frac{2}{\sqrt{nc}}$.

Quanquam vero, si quacunque ratione possit obtineri is numerus qui seriei hyperbolicæ respondet, non multum intersit utrum nec ne ejus relatio ad Circulum perspecta fuerit, attamen libenter fateor hanc patefactionem et labori pepercisse, et elegantiam singularem in solutionem induxisse.

II. Id mihi simul compertum fuit; Logarithmum rationis, quam in Potestate maxima, Medius Terminus habet ad Terminum intervallo l a Medio distantem, posita $n=2m$, sic expressum iri,

$$\overline{m+l-\tfrac{1}{2}} \times \log. \overline{m+l-1} + \overline{m-l+\tfrac{1}{2}} \times \log. \overline{m-l+1} - 2m \log. m + \log. \frac{m+l}{m}.$$

COROL.

[3]

COROLLARIUM I.

Hinc mihi illud colligere licet; si sit m quantitas infinite magna, Logarithmum hujus rationis fore $\frac{ll}{m}$ sive $\frac{2ll}{n}$, ergo Logarithmus rationis quam Terminus a Medio distans Intervallo l habet ad Medium est $-\frac{2ll}{n}$.

Hotelling and the Normal Distribution

Harold Hotelling, the "Normal Statistician", 1946

Display courtesy of Stephen Stigler.

p_1 in Log-linear Form

- Probabilities for four situations between two nodes:
- Define

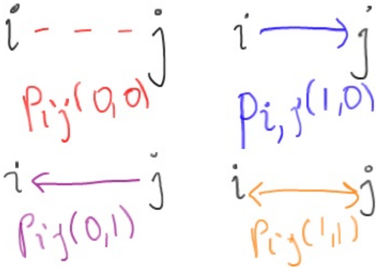
$$y_{ijkl} = \begin{cases} 1 & \text{if} \quad D(x_{ij}, x_{ji}) = (k, l), \\ 0 & \text{otherwise.} \end{cases}$$

- Yields an $n \times n \times 2 \times 2$ 4-way array with "zeros" for $n \times n$ diagonals.
- Fienberg and Wasserman demonstrated that in the 4-way table, the log-linear model of no second-order interaction corresponds to p_1 with constant reciprocation, i.e., [12][13][14][23][24][34], and that the standard iterative proportional fitting algorithm.
- Edge-dependent reciprocation corresponds to log-linear model [12][134][234].

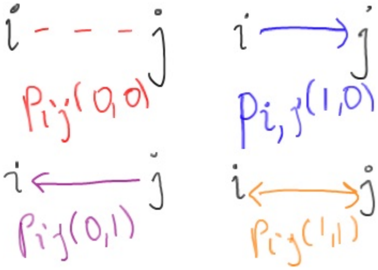
Primer: Algebraic Representation and Computer Algebra Tools

- Log-linear models are represented by polynomial maps and parameter space is toric variety.
- Multinomial likelihood is a monomial.
- Fiber consists of all tables with margins t.
- Markov basis generates fiber and consists of minimal generators of toric ideal.
- Computer algebra tools:
  - *4ti2*
    - We use 4ti2 to generate basis elements (perhaps redundant) for Markov bases for specific values of n. Can use these to compute exact distribution given the MSSs.
  - *Polymake*—Examines fiber for MLE existence.
    - We use Polymake to explore conditions for MLEs to exist, as in example for $\rho_{ij} = 0$. Affects computation and assessment of fit.

Algebraic Version of p_1 in Log-linear Form

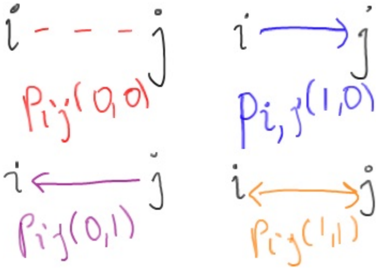


Algebraic Version of p_1 in Log-linear Form



- For each pair of nodes i and j we have a monomial in the model parameters:

Algebraic Version of p_1 in Log-linear Form



- For each pair of nodes i and j we have a monomial in the model parameters:

$$p_{ij}(a,b) \mapsto \lambda_{ij}\alpha_i^a\alpha_j^b\beta_i^b\beta_j^a\theta^{a+b}\rho_{ij}^{min(a,b)} \text{ for all } i<j \in \{1,\dots,n\}$$

p_1 as a toric variety

- The monomial map parametrizes a toric variety:

p_1 as a toric variety

- The monomial map parametrizes a toric variety:
  - the design matrix has $4\binom{n}{2}$ columns (variables)

p_1 as a toric variety

- The monomial map parametrizes a toric variety:
  - the design matrix has $4\binom{n}{2}$ columns (variables)
  - and $3\binom{n}{2}+1$, or $3\binom{n}{2}+2$, or $4\binom{n}{2}+2$ rows (parameters) depending on reciprocation

p_1 as a toric variety

- The monomial map parametrizes a toric variety:
  - the design matrix has $4\binom{n}{2}$ columns (variables)
  - and $3\binom{n}{2} + 1$, or $3\binom{n}{2} + 2$, or $4\binom{n}{2} + 2$ rows (parameters) depending on reciprocation
- For example, $n = 3$ and ρ_{ij} edge-dependent, the variety is a degree-3 hypersurface in $\mathbb{P}^{11}$.

p_1 as a toric variety

- The monomial map parametrizes a toric variety:
  - the design matrix has $4\binom{n}{2}$ columns (variables)
  - and $3\binom{n}{2}+1$, or $3\binom{n}{2}+2$, or $4\binom{n}{2}+2$ rows (parameters) depending on reciprocation
- For example, $n=3$ and ρ_{ij} edge-dependent, the variety is a degree-3 hypersurface in $\mathbb{P}^{11}$.
- Its defining ideal gives Markov basis,

p_1 as a toric variety

- The monomial map parametrizes a toric variety:
  - the design matrix has $4\binom{n}{2}$ columns (variables)
  - and $3\binom{n}{2}+1$, or $3\binom{n}{2}+2$, or $4\binom{n}{2}+2$ rows (parameters) depending on reciprocation
- For example, $n=3$ and ρ_{ij} edge-dependent, the variety is a degree-3 hypersurface in $\mathbb{P}^{11}$.
- Its defining ideal gives Markov basis,
- it will connect all networks with the same sufficient statistics.

Toric ideal of simplification of p_1

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

Toric ideal of simplification of p_1

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

Theorem (FPR)

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 $G_n := K_{n,n} \backslash \{i, i\}$.

Toric ideal of simplification of p_1

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

Theorem (FPR)

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 $G_n := K_{n,n} \backslash \{i, i\}$.

Theorem (FPR)

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.

Main Theorem - toric ideal of p_1

- Upshot: known Graver bases for edge subrings!

Main Theorem - toric ideal of p_1

- Upshot: known Graver bases for edge subrings!
- We now incorporate λ_{ij} into the previous theorems:

Main Theorem - toric ideal of p_1

- Upshot: known Graver bases for edge subrings!
- We now incorporate λ_{ij} into the previous theorems:

Theorem (FPR)

The toric ideal of the p_1 random graph model is the multi-homogenous piece of the toric ideal of a simplified model.

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

Main Theorem - toric ideal of p_1

- Upshot: known Graver bases for edge subrings!
- We now incorporate λ_{ij} into the previous theorems:

Theorem (FPR)

The toric ideal of the p_1 random graph model is the multi-homogenous piece of the toric ideal of a simplified model.

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

- In progress: Markov moves for all three cases of ρ_{ij}.

Example with 4 nodes

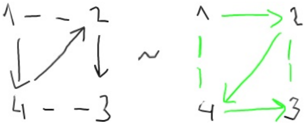


Figure: A degree-5 binomial

Example with 4 nodes

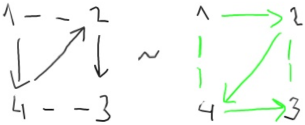


Figure: A degree-5 binomial

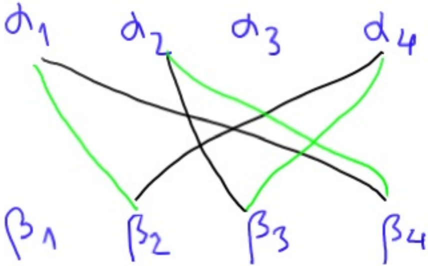


Figure: the corresponding path in $K_{4,4} \backslash \{i, i\}$

3-node network

For all 3 cases of ρ_{ij}, there is only one Markov move:

3-node network

For all 3 cases of ρ_{ij}, there is only one Markov move:

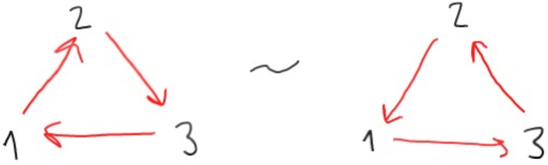


Figure: A binomial of degree 3

3-node network

For all 3 cases of ρ_{ij}, there is only one Markov move:

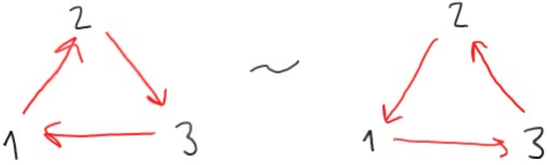


Figure: A binomial of degree 3

representing the binomial:

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

3-node network

For all 3 cases of ρ_{ij}, there is only one Markov move:

Figure: A binomial of degree 3

representing the binomial:

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

The corresponding Markov move is to remove edges $1 \rightarrow 2$, $2 \rightarrow 3$ and $3 \rightarrow 1$, and replace them by the same edges oriented in the opposite direction.

4-node network, ρ_{ij} constant

- Reminder: $\rho_{ij} = \rho$, including $\rho_{ij} = 0$.

4-node network, ρ_{ij} constant

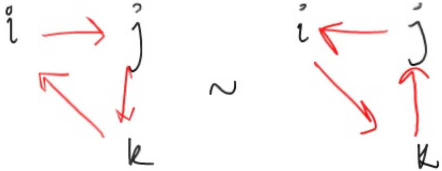


Figure: A binomial of degree 3

4-node network, ρ_{ij} constant

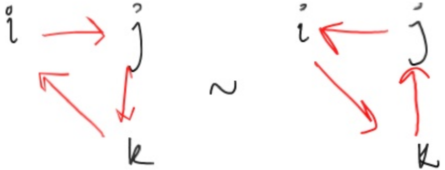


Figure: A binomial of degree 3

Figure: A binomial of degree 4

4-node network, ρ_{ij} Constant

Figure: A more complicated binomial of degree 4

4-node network, ρ_{ij} Constant

Figure: A more complicated binomial of degree 4

represents

$$p_{ij}(0,0)p_{jk}(1,1)p_{kl}(0,1)p_{il}(1,0) - p_{ij}(1,0)p_{jk}(1,0)p_{kl}(1,1)p_{il}(0,0)$$

4-node network, ρ_{ij} Constant

Figure: A more complicated binomial of degree 4

represents

$$p_{ij}(0,0)p_{jk}(1,1)p_{kl}(0,1)p_{il}(1,0) - p_{ij}(1,0)p_{jk}(1,0)p_{kl}(1,1)p_{il}(0,0)$$

- The moves come in degrees $3, 4, 5, 6$.

4-node network, ρ_{ij} edge-dependent

Reminder: $\rho_{ij} = \rho + \rho_i + \rho_j$.

4-node network, ρ_{ij} edge-dependent

Reminder: $\rho_{ij} = \rho + \rho_i + \rho_j$.

What follows is a list of all possible moves on a 4-node network, with respect to symmetry of course.

4-node network, ρ_{ij} edge-dependent

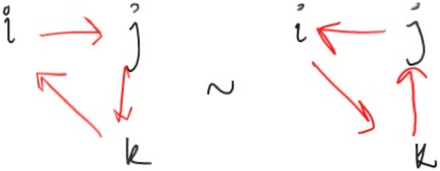


Figure: A binomial of degree 3

4-node network, ρ_{ij} edge-dependent

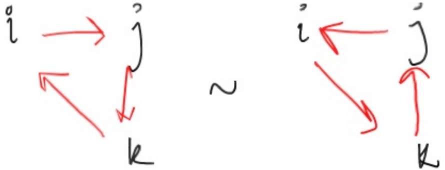


Figure: A binomial of degree 3

Figure: A binomial of degree 4

4-node Network, ρ_{ij} edge-dependent

Figure: A binomial of degree 4

4-node Network, ρ_{ij} edge-dependent

Figure: A binomial of degree 4

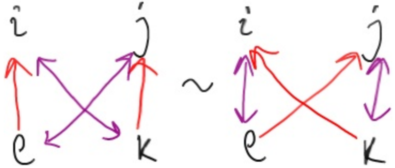


Figure: A binomial of degree 4

4-node network, ρ_{ij} edge-dependent

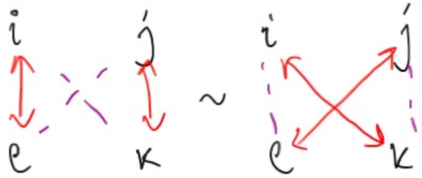


Figure: A binomial of degree 4

4-node network, ρ_{ij} edge-dependent

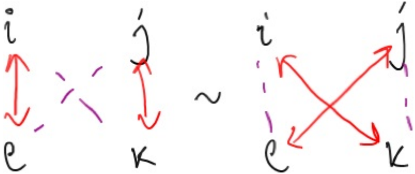


Figure: A binomial of degree 4

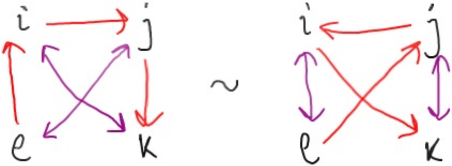


Figure: A binomial of degree 5

4-node network, ρ_{ij} edge-dependent

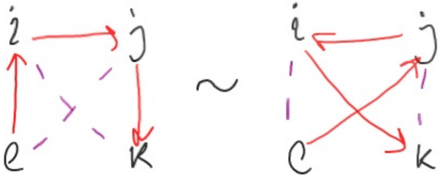


Figure: A binomial of degree 5

4-node network, ρ_{ij} edge-dependent

Figure: A binomial of degree 5

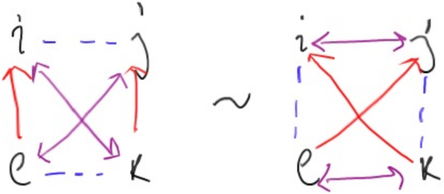


Figure: A binomial of degree 6

Summary

- Graphs and network models.
- p_1 and its algebraic representation:
    - Reviewed of p_1 and log-linear models.
    - Developed Markov bases.
- Markov bases and proposed uses:
    - Existence of MLE.
    - Generate exact distribution for p_1 and use for assessing goodness-of-fit.
- Future work includes:
    - Completing algebraic characterization of p_1 models and putting them to use.
    - Generalizations to exponential random graph models (ERGM) models [also known as p^* models]:
        - Many complex statistical issues including existence of MLEs (non-existence = degeneracies involving zero estimates) and near-degeneracies.
        - Scaling tools up to be of use for analysis large networks with efficient computation.

... The End ...