

Exponential-Family Random Graph Models for Social Networks

David Hunter

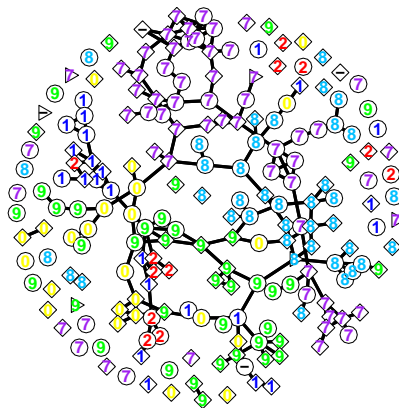
Department of Statistics
Penn State University

Research supported by NIDA Grant DA012831 and NICHD Grant HD041877

Society for Political Methodology, July 18, 2007

Example Network: High School Friendship Data

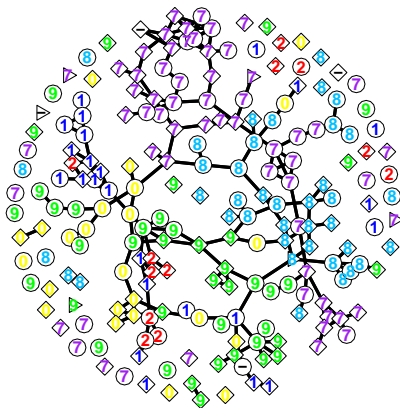
School 10: 205 Students



- An edge indicates a mutual friendship.
- Colored labels give grade level, 7 through 12.
- Circles = female, squares = male, triangles = unknown.

Example Network: High School Friendship Data

School 10: 205 Students



- An edge indicates a mutual friendship.
- Colored labels give grade level, 7 through 12.
- Circles = female, squares = male, triangles = unknown.
- How do covariates influence network formation?
- i.e., how can we do “regression” with a network response?

- 1 The ERG Model Class
- 2 Approximating the MLE
- 3 Maximum Pseudolikelihood
- 4 Assessing Model Goodness-of-Fit

Exponential-Family Random Graph Model (ERGM)

$$P_{\theta}(Y = y) \propto \exp\{\theta^t g(y)\}$$

or

$$P_{\theta}(Y = y) = \frac{\exp\{\theta^t g(y)\}}{\kappa(\theta)},$$

where

- Y is a random network on n nodes (a matrix of 0's and 1's)
- θ is a vector of parameters
- $g(y)$ is a known vector of graph statistics on y
- $\kappa(\theta)$ makes all the probabilities sum to 1

Whence the name ERGM?

Exponential Family

Whenever the density of a random variable may be written

$$f(y) \propto \exp\{\theta^t g(y)\},$$

the family of all such random variables (for all possible θ) is called an **exponential family**.

- Since the random graphs in our model form an exponential family, we call the model an **exponential-family random graph model**.

Whence the name ERGM?

Exponential Family

Whenever the density of a random variable may be written

$$f(y) \propto \exp\{\theta^t g(y)\},$$

the family of all such random variables (for all possible θ) is called an **exponential family**.

- Since the random graphs in our model form an exponential family, we call the model an **exponential-family random graph model**.
- “ERGM” is easier to pronounce than “EFRGM”!

The goal of estimation

Exponential-family Random Graph Model (ERGM)

$$P_{\theta}(Y = y) = \frac{\exp\{\theta^t g(y)\}}{\kappa(\theta)}$$

If θ is not known, the above equation defines a model *class*, not a model.

Goal:

Use observed data (a network y^{obs}) to determine the “best” model from the model class.

In other words, find the “best” θ .

Estimation preliminaries

The model class:

$$P_{\theta}(Y = y) = \frac{\exp\{\theta^t g(y)\}}{\kappa(\theta)}$$

The model class:

$$P_{\theta}(Y = y) = \frac{\exp\{\theta^t g(y)\}}{\kappa(\theta)}$$

- It follows that $\kappa(\theta)$ is a normalizing “constant”:

$$\kappa(\theta) = \sum_{\substack{\text{all possible} \\ \text{graphs } z}} \exp\{\theta^t g(z)\}.$$

The model class:

$$P_{\theta}(Y = y) = \frac{\exp\{\theta^t g(y)\}}{\kappa(\theta)}$$

- It follows that $\kappa(\theta)$ is a normalizing “constant”:

$$\kappa(\theta) = \sum_{\substack{\text{all possible} \\ \text{graphs } z}} \exp\{\theta^t g(z)\}.$$

- Let y^{obs} denote the observed graph, i.e., the data.

The model class:

$$P_{\theta}(Y = y) = \frac{\exp\{\theta^t g(y)\}}{\kappa(\theta)}$$

- It follows that $\kappa(\theta)$ is a normalizing “constant”:

$$\kappa(\theta) = \sum_{\substack{\text{all possible} \\ \text{graphs } z}} \exp\{\theta^t g(z)\}.$$

- Let y^{obs} denote the observed graph, i.e., the data.
- Replacing $g(y)$ by $[g(y) - g(y^{\text{obs}})]$ leaves $P_{\theta}(Y = y)$ unchanged

The model class:

$$P_{\theta}(Y = y) = \frac{\exp\{\theta^t g(y)\}}{\kappa(\theta)}, \text{ where } g(y^{\text{obs}}) = 0$$

- It follows that $\kappa(\theta)$ is a normalizing “constant”:

$$\kappa(\theta) = \sum_{\text{all possible graphs } z} \exp\{\theta^t g(z)\}.$$

- Let y^{obs} denote the observed graph, i.e., the data.
- Replacing $g(y)$ by $[g(y) - g(y^{\text{obs}})]$ leaves $P_{\theta}(Y = y)$ unchanged
- Thus, WLOG we may “recenter” $g(y)$ so that $g(y^{\text{obs}}) = 0$

The loglikelihood function

The model class:

$$P_{\theta}(Y = y) = \frac{\exp\{\theta^t g(y)\}}{\kappa(\theta)}, \text{ where } g(y^{\text{obs}}) = 0.$$

- With $g(y)$ recentered so that $g(y^{\text{obs}}) = 0$, the loglikelihood is simply $\ell(\theta) = -\log \kappa(\theta)$.

The loglikelihood function

The model class:

$$P_{\theta}(Y = y) = \frac{\exp\{\theta^t g(y)\}}{\kappa(\theta)}, \text{ where } g(y^{\text{obs}}) = 0.$$

- With $g(y)$ recentered so that $g(y^{\text{obs}}) = 0$, the loglikelihood is simply $\ell(\theta) = -\log \kappa(\theta)$.
- We'd like to find $\hat{\theta}$ to maximize $\ell(\theta) = -\log \kappa(\theta)$.
- Thus, $\hat{\theta}$ denotes the **maximum likelihood estimator**.

Fact:

The method of maximum likelihood is one of the most well-studied topics in the entire field of statistics, and maximum likelihood estimators generally are known to have a lot of nice properties.

Fact:

The method of maximum likelihood is one of the most well-studied topics in the entire field of statistics, and maximum likelihood estimators generally are known to have a lot of nice properties.

Warning: The fact that $P_{\hat{\theta}}(Y = y^{\text{obs}})$ is as large as possible in this model class does NOT mean that y^{obs} is particularly likely relative to other networks!

(The model class itself might be inappropriate.)

- 1 The ERG Model Class
- 2 Approximating the MLE**
- 3 Maximum Pseudolikelihood
- 4 Assessing Model Goodness-of-Fit

It is difficult to find the MLE

The model class:

$$P_{\theta}(Y = y) = \frac{\exp\{\theta^t g(y)\}}{\kappa(\theta)}, \text{ where } g(y^{\text{obs}}) = 0$$

- We wish to maximize the loglikelihood function

$$\ell(\theta) = -\log \kappa(\theta) = -\log \sum_{\substack{\text{all possible} \\ \text{graphs } z}} \exp\{\theta^t g(z)\}.$$

It is difficult to find the MLE

The model class:

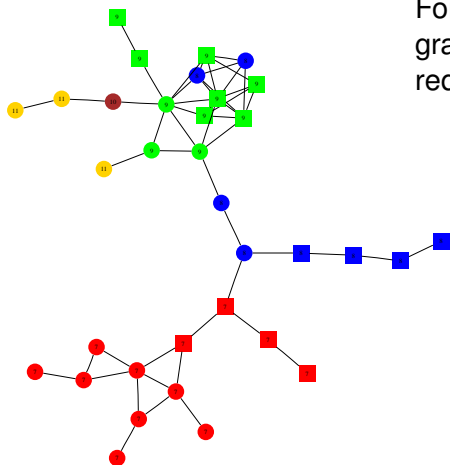
$$P_{\theta}(Y = y) = \frac{\exp\{\theta^t g(y)\}}{\kappa(\theta)}, \text{ where } g(y^{\text{obs}}) = 0$$

- We wish to maximize the loglikelihood function

$$\ell(\theta) = -\log \kappa(\theta) = -\log \sum_{\substack{\text{all possible} \\ \text{graphs } z}} \exp\{\theta^t g(z)\}.$$

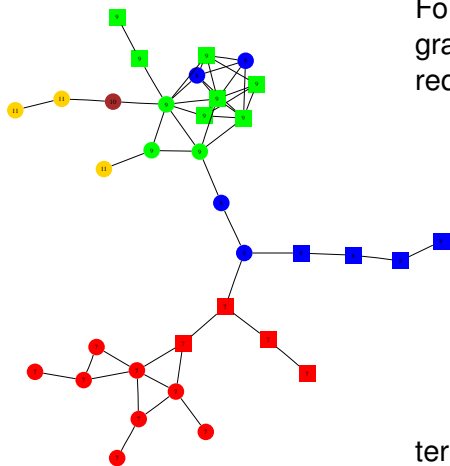
- Alas, $\kappa(\theta)$ is a nasty beast: Merely evaluating (let alone maximizing) $\ell(\theta)$ is somewhat computationally burdensome...

How burdensome?



For this undirected, 34-node graph, computing $\ell(\theta)$ directly requires summation of

How burdensome?



For this undirected, 34-node graph, computing $\ell(\theta)$ directly requires summation of

7,547,924,849,643,082,704,483,
109,161,976,537,781,833,842,
440,832,880,856,752,412,600,
491,248,324,784,297,704,172,
253,450,355,317,535,082,936,
750,061,527,689,799,541,169,
259,849,585,265,122,868,502,
865,392,087,298,790,653,952

terms.

A nifty fact regarding the MLE $\hat{\theta}$

- Because we're dealing with an exponential family of models,

$$E_{\hat{\theta}} g(Y) = g(y^{\text{obs}})$$

and no other value of θ has this property.

- In words:

The MLE gives the unique model in the model class under which the mean value of the vector of statistics equals its observed value.

A nifty fact regarding the MLE $\hat{\theta}$

- Because we're dealing with an exponential family of models,

$$E_{\hat{\theta}} g(Y) = g(y^{\text{obs}})$$

and no other value of θ has this property.

- In words:

The MLE gives the unique model in the model class under which the mean value of the vector of statistics equals its observed value.

- This fact may even be exploited to approximate $\hat{\theta}$.
(See Snijders 2002, *J. of Social Structure*)

A different approach

- Suppose we fix θ_0 . A bit of algebra shows that

$$-\log E_{\theta_0} [\exp \{(\theta - \theta_0)^t g(Y)\}] = \ell(\theta) - \ell(\theta_0).$$

A different approach

- Suppose we fix θ_0 . A bit of algebra shows that

$$-\log E_{\theta_0} [\exp \{(\theta - \theta_0)^t g(Y)\}] = \ell(\theta) - \ell(\theta_0).$$

- More to the point,

$$-\log E_{\theta_0} [\textit{blah blah } Y \textit{ blah}] = \ell(\theta) - \ell(\theta_0).$$

A different approach

- Suppose we fix θ_0 . A bit of algebra shows that

$$-\log E_{\theta_0} [\exp \{(\theta - \theta_0)^t g(Y)\}] = \ell(\theta) - \ell(\theta_0).$$

- More to the point,

$$-\log E_{\theta_0} [\textit{blah blah } Y \textit{ blah}] = \ell(\theta) - \ell(\theta_0).$$

- Thus, $\ell(\theta) - \ell(\theta_0)$ involves an expected value (a mean).

Law of Large Numbers to the Rescue!

The LOLN suggests that we approximate an unknown population mean by a sample mean.

Law of Large Numbers to the Rescue!

The LOLN suggests that we approximate an unknown population mean by a sample mean.

Thus,

$$\begin{aligned}\ell(\theta) - \ell(\theta_0) &= -\log \mathbb{E}_{\theta_0} \left(\exp \{ (\theta - \theta_0)^t g(Y) \} \right) \\ &\approx -\log \frac{1}{m} \sum_{i=1}^m \exp \{ (\theta - \theta_0)^t g(Y_i) \},\end{aligned}$$

where Y_1, Y_2, \dots, Y_m is a random sample of networks from the distribution defined by the ERGM with parameter θ_0 .

More to the point,

$$\begin{aligned}\ell(\theta) - \ell(\theta_0) &\approx -\log \frac{1}{m} \sum_{i=1}^m \exp \{(\theta - \theta_0)^t \mathbf{s}(Y_i)\} \\ &= -\log \frac{1}{m} \sum_{i=1}^m (\text{blah } \theta \text{ blah } Y_i \text{ blah}).\end{aligned}$$

More to the point,

$$\begin{aligned}\ell(\theta) - \ell(\theta_0) &\approx -\log \frac{1}{m} \sum_{i=1}^m \exp \{(\theta - \theta_0)^t s(Y_i)\} \\ &= -\log \frac{1}{m} \sum_{i=1}^m (\text{blah } \theta \text{ blah } Y_i \text{ blah}).\end{aligned}$$

- Given a random sample of networks from P_{θ_0} , we can thus approximate (and subsequently maximize) the loglikelihood shifted by a constant.

More to the point,

$$\begin{aligned}\ell(\theta) - \ell(\theta_0) &\approx -\log \frac{1}{m} \sum_{i=1}^m \exp \{(\theta - \theta_0)^t \mathbf{s}(Y_i)\} \\ &= -\log \frac{1}{m} \sum_{i=1}^m (\text{blah } \theta \text{ blah } Y_i \text{ blah}).\end{aligned}$$

- Given a random sample of networks from P_{θ_0} , we can thus approximate (and subsequently maximize) the loglikelihood shifted by a constant.
- We can approximate the MLE if we can simulate random networks!

How should θ_0 be chosen?

- Theoretically, the estimated value of $\ell(\theta) - \ell(\theta_0)$ converges to the true value as the size of the MCMC sample increases, regardless of the value of θ_0 .

How should θ_0 be chosen?

- Theoretically, the estimated value of $\ell(\theta) - \ell(\theta_0)$ converges to the true value as the size of the MCMC sample increases, regardless of the value of θ_0 .
- However, in practice this convergence can be agonizingly slow, especially if θ_0 is not chosen close to the maximizer of the likelihood.

How should θ_0 be chosen?

- Theoretically, the estimated value of $\ell(\theta) - \ell(\theta_0)$ converges to the true value as the size of the MCMC sample increases, regardless of the value of θ_0 .
- However, in practice this convergence can be agonizingly slow, especially if θ_0 is not chosen close to the maximizer of the likelihood.
- A choice that sometimes works is the MPLE (maximum pseudolikelihood estimate)

- 1 The ERG Model Class
- 2 Approximating the MLE
- 3 Maximum Pseudolikelihood**
- 4 Assessing Model Goodness-of-Fit

Conditional log-odds of an edge

Notation: For a network y and a pair (i, j) of nodes,

- $y_{ij} = 0$ or 1 , depending on whether there is an edge
- y_{ij}^c denotes the status of all pairs in y other than (i, j)
- y_{ij}^+ denotes the same network as y but with $y_{ij} = 1$
- y_{ij}^- denotes the same network as y but with $y_{ij} = 0$

Conditional log-odds of an edge

Notation: For a network y and a pair (i, j) of nodes,

- $y_{ij} = 0$ or 1 , depending on whether there is an edge
- y_{ij}^c denotes the status of all pairs in y other than (i, j)
- y_{ij}^+ denotes the same network as y but with $y_{ij} = 1$
- y_{ij}^- denotes the same network as y but with $y_{ij} = 0$

Conditional on $Y_{ij}^c = y_{ij}^c$, Y has only two possible states, depending on whether $Y_{ij} = 0$ or $Y_{ij} = 1$.

Conditional log-odds of an edge

Notation: For a network y and a pair (i, j) of nodes,

- $y_{ij} = 0$ or 1 , depending on whether there is an edge
- y_{ij}^c denotes the status of all pairs in y other than (i, j)
- y_{ij}^+ denotes the same network as y but with $y_{ij} = 1$
- y_{ij}^- denotes the same network as y but with $y_{ij} = 0$

Conditional on $Y_{ij}^c = y_{ij}^c$, Y has only two possible states, depending on whether $Y_{ij} = 0$ or $Y_{ij} = 1$.

Let's calculate the ratio of the two respective probabilities.

[We'll use $P_\theta(Y = y) = \exp\{\theta^t g(y)\} / \kappa(\theta)$.]

Conditional log-odds of an edge

Notation: For a network y and a pair (i, j) of nodes,

- $y_{ij} = 0$ or 1 , depending on whether there is an edge
- y_{ij}^c denotes the status of all pairs in y other than (i, j)
- y_{ij}^+ denotes the same network as y but with $y_{ij} = 1$
- y_{ij}^- denotes the same network as y but with $y_{ij} = 0$

$$\frac{P(Y_{ij} = 1 | Y_{ij}^c = y_{ij}^c)}{P(Y_{ij} = 0 | Y_{ij}^c = y_{ij}^c)} = \frac{\exp\{\theta^t g(y_{ij}^+)\}}{\exp\{\theta^t g(y_{ij}^-)\}}$$

A lot of cancellation happened on the right hand side!

Conditional log-odds of an edge

Notation: For a network y and a pair (i, j) of nodes,

- $y_{ij} = 0$ or 1 , depending on whether there is an edge
- y_{ij}^c denotes the status of all pairs in y other than (i, j)
- y_{ij}^+ denotes the same network as y but with $y_{ij} = 1$
- y_{ij}^- denotes the same network as y but with $y_{ij} = 0$

$$\frac{P(Y_{ij} = 1 | Y_{ij}^c = y_{ij}^c)}{P(Y_{ij} = 0 | Y_{ij}^c = y_{ij}^c)} = \exp\{\theta^t [g(y_{ij}^+) - g(y_{ij}^-)]\}$$

A lot of cancellation happened on the right hand side!

Conditional log-odds of an edge

Notation: For a network y and a pair (i, j) of nodes,

- $y_{ij} = 0$ or 1 , depending on whether there is an edge
- y_{ij}^c denotes the status of all pairs in y other than (i, j)
- y_{ij}^+ denotes the same network as y but with $y_{ij} = 1$
- y_{ij}^- denotes the same network as y but with $y_{ij} = 0$

$$\log \frac{P(Y_{ij} = 1 | Y_{ij}^c = y_{ij}^c)}{P(Y_{ij} = 0 | Y_{ij}^c = y_{ij}^c)} = \theta^t [g(y_{ij}^+) - g(y_{ij}^-)]$$

Conditional log-odds of an edge

Notation: For a network y and a pair (i, j) of nodes,

- $\delta(y_{ij}^c)$ denotes the vector of change statistics,

$$\delta(y_{ij}^c) = g(y_{ij}^+) - g(y_{ij}^-).$$

So $\delta(y_{ij}^c)$ is the conditional log-odds of edge (i, j) .

$$\log \frac{P(Y_{ij} = 1 | Y_{ij}^c = y_{ij}^c)}{P(Y_{ij} = 0 | Y_{ij}^c = y_{ij}^c)} = \theta^t \delta(y_{ij}^c)$$

Maximum Pseudolikelihood: Intuition

- What if we assume that there is no dependence (or very weak dependence) among the Y_{ij} ?

Maximum Pseudolikelihood: Intuition

- What if we assume that there is no dependence (or very weak dependence) among the Y_{ij} ?
- In other words, what if we approximate the marginal $P(Y_{ij} = 1)$ by the conditional $P(Y_{ij} = 1 | Y_{ij}^c = y_{ij}^c)$?

Maximum Pseudolikelihood: Intuition

- What if we assume that there is no dependence (or very weak dependence) among the Y_{ij} ?
- In other words, what if we approximate the marginal $P(Y_{ij} = 1)$ by the conditional $P(Y_{ij} = 1 | Y_{ij}^c = y_{ij}^c)$?
- Then the Y_{ij} are independent with

$$\log \frac{P(Y_{ij} = 1)}{P(Y_{ij} = 0)} = \theta^t \delta[(y^{\text{obs}})_{ij}^c],$$

so we obtain an estimate of θ using straightforward logistic regression.

Maximum Pseudolikelihood: Intuition

- What if we assume that there is no dependence (or very weak dependence) among the Y_{ij} ?
- In other words, what if we approximate the marginal $P(Y_{ij} = 1)$ by the conditional $P(Y_{ij} = 1 | Y_{ij}^c = y_{ij}^c)$?
- Then the Y_{ij} are independent with

$$\log \frac{P(Y_{ij} = 1)}{P(Y_{ij} = 0)} = \theta^t \delta[(y^{\text{obs}})_{ij}^c],$$

so we obtain an estimate of θ using straightforward logistic regression.

- Result: The **maximum pseudolikelihood estimate**.

Maximum Pseudolikelihood: Intuition

- What if we assume that there is no dependence (or very weak dependence) among the Y_{ij} ?
- In other words, what if we approximate the marginal $P(Y_{ij} = 1)$ by the conditional $P(Y_{ij} = 1 | Y_{ij}^c = y_{ij}^c)$?
- Then the Y_{ij} are independent with

$$\log \frac{P(Y_{ij} = 1)}{P(Y_{ij} = 0)} = \theta^t \delta[(y^{\text{obs}})_{ij}^c],$$

so we obtain an estimate of θ using straightforward logistic regression.

- Result: The **maximum pseudolikelihood estimate**.
- For independence models, MPLE = MLE!

Warnings about MPLE

Unfortunately, little is known about the quality of MPL estimates in general, but we do know some ways in which they can be misleading.

- If the model is bad, you'll get MPLE results quite easily (unlike MLE results), masking the problem.
- If the model is good, in many cases the MPLE looks “close” to the MLE; however, “close” can be deceiving, since small changes in θ can sometimes lead to large differences in the behavior of randomly generated networks.

- 1 The ERG Model Class
- 2 Approximating the MLE
- 3 Maximum Pseudolikelihood
- 4 Assessing Model Goodness-of-Fit**

ERGM
class
 $\exp\{\theta^t g(y)\}$

Goodness of fit intuition

ERGM
class

$$\exp\{\theta^t g(y)\}$$



↑
 y^{obs}



Goodness of fit intuition

ERGM
class
 $\exp\{\theta^t g(y)\}$

→

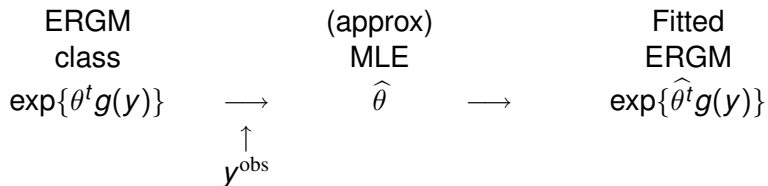
(approx)
MLE

$\hat{\theta}$

↑
 y^{obs}



Goodness of fit intuition



Goodness of fit intuition

ERGM
class
 $\exp\{\theta^t g(y)\}$

→

↑
 y^{obs}



(approx)
MLE

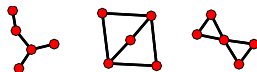
$\hat{\theta}$

→

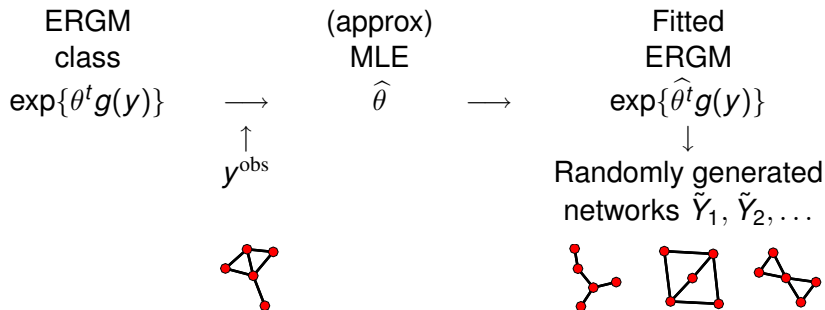
Fitted
ERGM
 $\exp\{\hat{\theta}^t g(y)\}$

↓

Randomly generated
networks $\tilde{Y}_1, \tilde{Y}_2, \dots$



Goodness of fit intuition

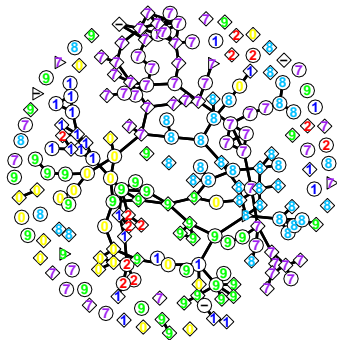


- Question: How does y^{obs} “look” as a representative of the sample $\tilde{Y}_1, \tilde{Y}_2, \dots$?

The eyeball test

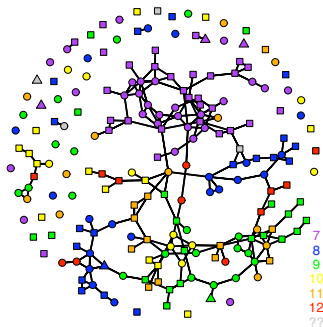
The data:

School 10: 205 Students



Simulated network,
model A:

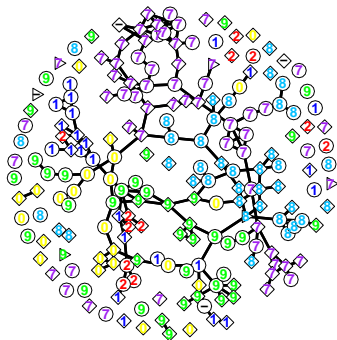
Simulated graph: By grade



The eyeball test (cont'd)

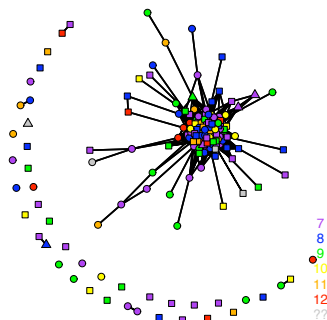
The data:

School 10: 205 Students



Simulated network,
model B:

Simulated graph: By grade



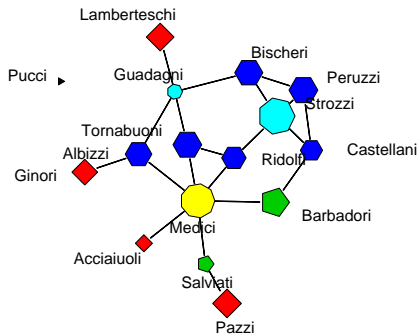
(Yikes!)

The models

- Model A: $g(y)$ contains terms for
 - # of edges
 - Homophily effects of grade, sex, and race factors
 - Main effects of grade, sex, and race factors
 - $\sum_i (.632)^i EP_i$, where $EP_i = \#$ edges with i shared partners
- Model B: $g(y)$ contains terms for
 - # of edges
 - # of neighbors of the same sex (homophily effect)
 - # of 2-stars
 - # of triangles

(Note: It was necessary to use MPLE to fit Model B)

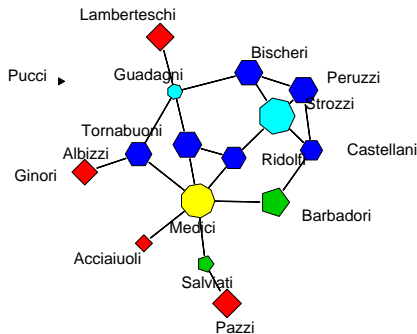
A well-known example:



Florentine marriage data

- Edge indicates marriage tie between families
- Sides=degree + 3
- Color=degree
- Size=log(wealth)

A well-known example:



Florentine marriage data

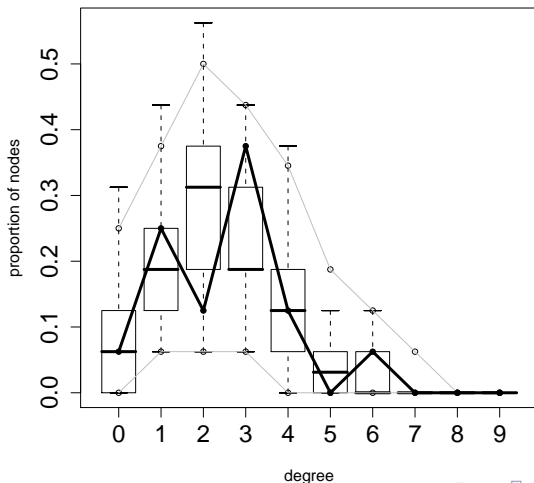
- Edge indicates marriage tie between families
- Sides=degree + 3
- Color=degree
- Size=log(wealth)

```
modell <- ergm(flomarriage ~ edges + kstar(2))
```

Graphical GOF check: degree distribution

```
modell <- ergm(flomarriage ~ edges + kstar(2))
```

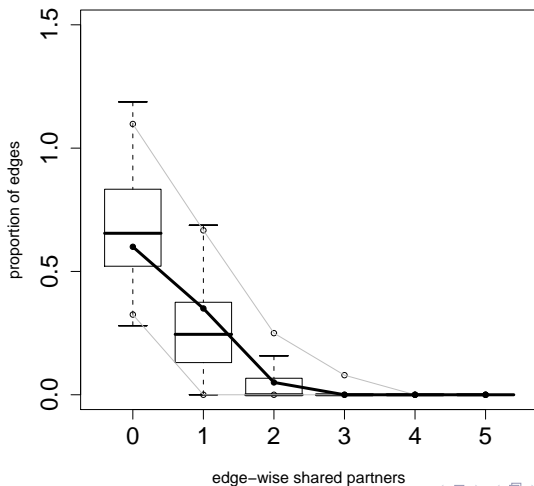
Goodness-of-fit diagnostics



Graphical GOF: edgewise shared partner distribution

```
modell1 <- ergm(flomarriage ~ edges + kstar(2))
```

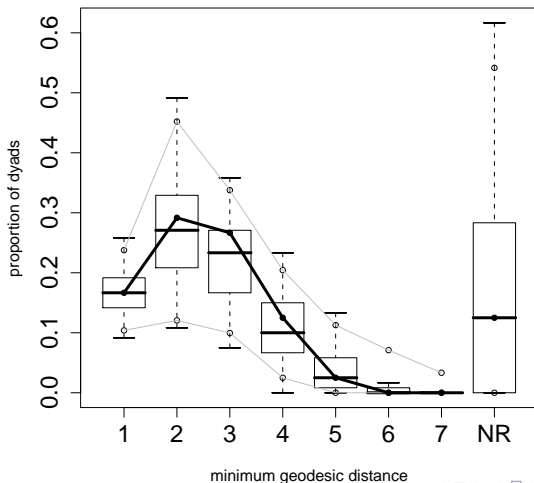
Goodness-of-fit diagnostics



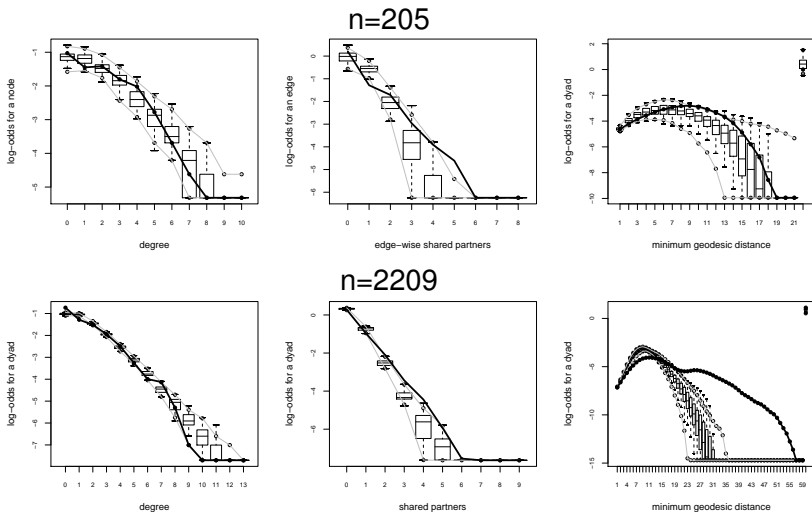
Graphical GOF check: geodesic distance distribution

```
modell1 <- ergm(flomarriage ~ edges + kstar(2))
```

Goodness-of-fit diagnostics



GOF check: Less trivial examples



Hunter, Goodreau, Handcock (2007), JASA, to appear.

Some Useful References

- Frank, O. and D. Strauss (1986), Markov graphs, *JASA*
- Geyer, C. J. and E. Thompson (1992), Constrained Monte Carlo maximum likelihood for dependent data, *J. Roy. Stat. Soc. B*
- Handcock, M. S. (2003) Assessing degeneracy in statistical models of social networks,
<http://www.csss.washington.edu/Papers>
- Holland, P. W. and S. Leinhardt (1981), An exponential family of probability distributions for directed graphs, *JASA*
- Snijders, Tom A. B. (2002), Markov chain Monte Carlo estimation of exponential random graph models, *J. Soc. Struct.*
- Strauss, D. and M. Ikeda (1990), Pseudolikelihood estimation for social networks, *JASA*
- Wasserman, S. and P. Pattison (1996), Logit models and logistic regression for social networks: I. An introduction to Markov graphs and p^* , *Psychometrika*