Statistical Inference in Social Network Analysis: Exponential Random Graph Models

jimi adams

Associate Professor

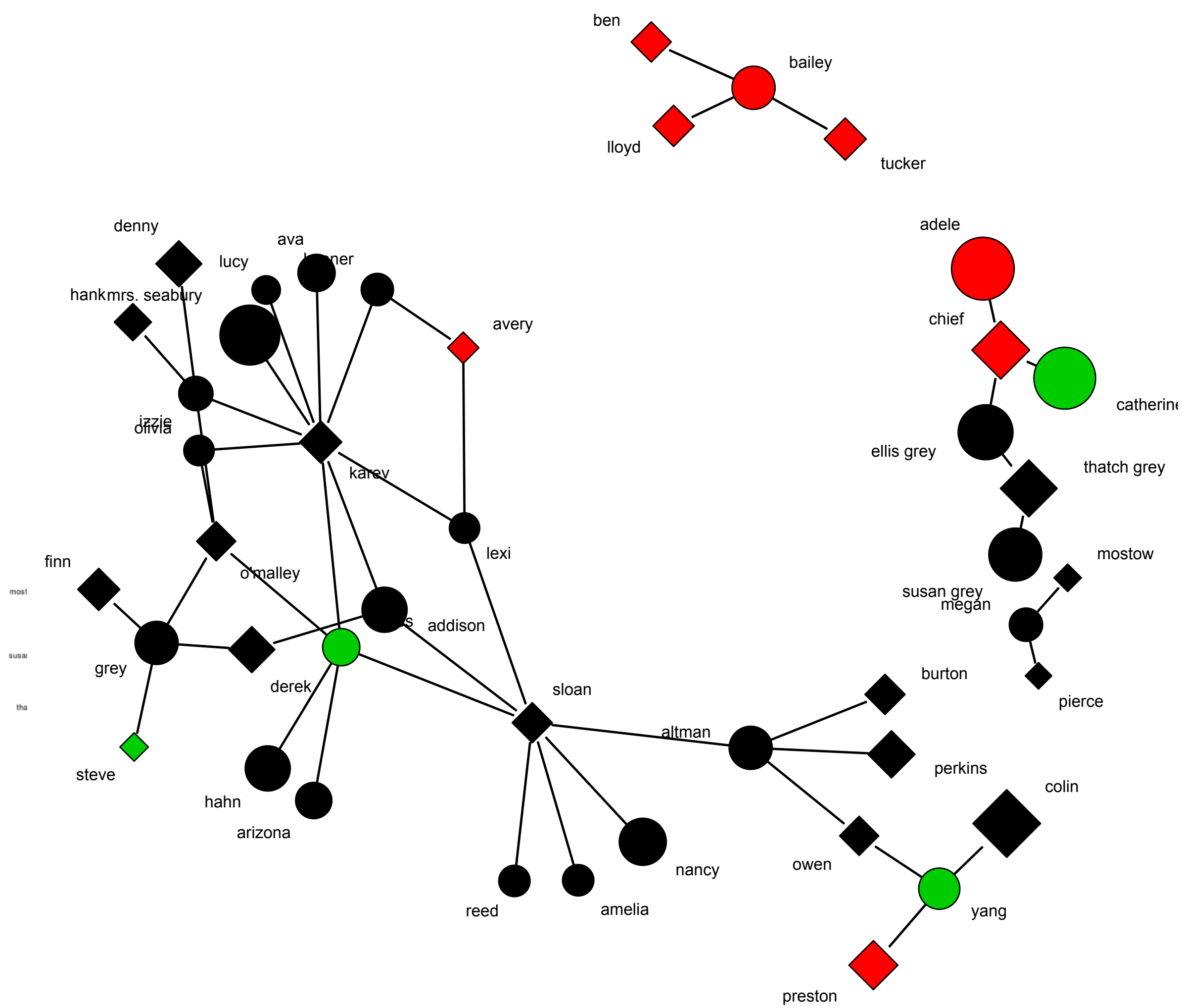
University of Colorado Denver

Department of Health & Behavioral Sciences

Slides & Tutorial files available @ – http://goo.gl/hkELGa

Can we develop a model to account for the structure observed in this network?

What does the model need to account for?



Example adapted from Benjamin Lind:
http://badhessian.org/2012/09/lessons-on-exponential-random-graph-modeling-from-greys-anatomy-hook-ups/

Modeled Interdependencies

<table>
<tr><th></th><th></th><th>None</th><th>w/in Dyad[1]</th><th colspan="2">Dyad +</th></tr>
<tr><td rowspan="2">Outcome</td><td>Attribute</td><td>General Linear Model</td><td>Actor-Partner Independence (APIM)</td><td>Network Auto-Regression</td><td rowspan="2">Stochastic Actor Based Model (SABM, SAOM, Siena)</td></tr>
<tr><td>Network</td><td>{Erdös-Renyi, Bernouli} Random</td><td>MR(QAP)[2], latent space, CUG[3], p1/p2</td><td>ERGM, Relational Event</td></tr>
</table>

[1] Frequently described as “dyad independence” models.

[2] R (sna) = qaptest

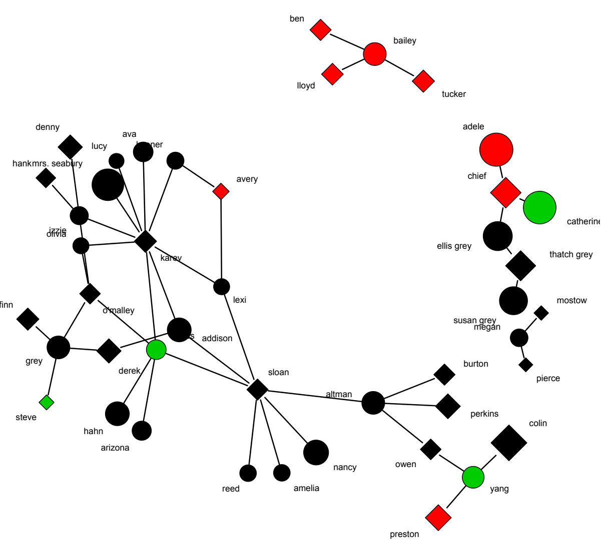
[3] Conditional Uniform Graphs

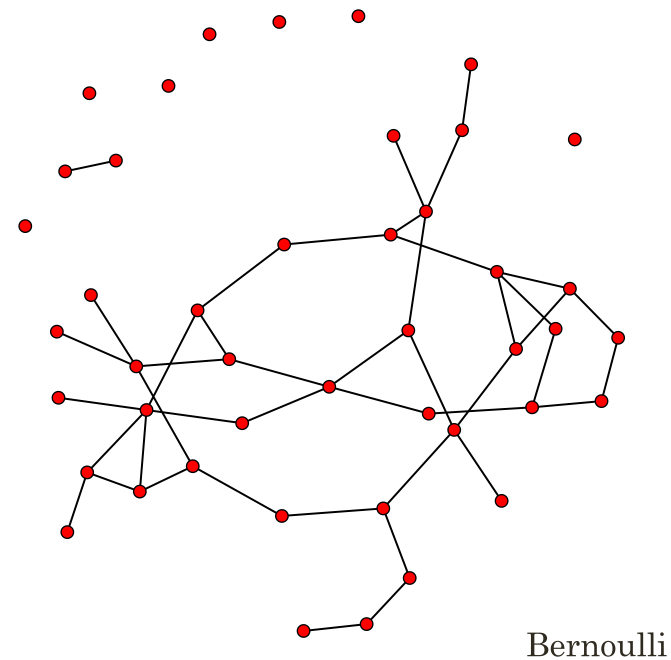
[4] R (sna) = lnam

1. Homogeneous Bernoulli (~Erdős-Rényi):

- $\Pr(X = x) = (1/\kappa) \exp(\theta L(x))$

- Assumes ties form completely at random, or independently from one another – but there are many ways that's a bad assumption.





Modeled Interdependencies

Outcome

	None	**w/in Dyad**[1]	**Dyad +**	
Attribute	General Linear Model	Actor-Partner Independence (APIM)	Network Auto-Regression[4]	Stochastic Actor Based Model (SABM, SAOM, Siena)
Network	{Erdös-Renyi, Bernouli} Random	MR(QAP)[2], latent space, CUG[3], p1/p2	**ERGM**, Relational Event	

[1] Frequently described as "dyad independence" models.
[2] R (sna) = qaptest
[3] Conditional Uniform Graphs
[4] R (sna) = lnam

1. Homogeneous Bernoulli (~Erdős-Rényi):
   - $Pr(X = x) = (1/\kappa) \exp(\theta L(x))$

2. Bernoulli blockmodel:
   - $Pr(X = x) = (1/\kappa) \exp\{\theta_{11} L_{11}(x)+\theta_{12} L_{12}(x)+\theta_{21} L_{21}(x)+\theta_{22} L_{22}(x)\}$

3. Homogeneous Dyad-independent:
   - $Pr(X = x) = (1/\kappa) \exp\{\theta L(x) + \rho M(x)\}$

4. $$p_1(\mathrm{x}) = K * \exp\left[\rho M + \theta L + \sum \alpha_i X_{i+} + \sum \beta_i X_{+i}\right]$$

   …(detailed on next slide)

- p1 – **Parametric** Dyad Independence model
- Assume independent dyads $D_{ij} = (X_{ij}, X_{ji})$

$p(D_{ij}=(1,1)) = m_{ij},\ i < j$ $\quad$ $p(D_{ij}=(1,0)) = a_{ij},\ i \neq j$

$p(D_{ij}=(0,0)) = n_{ij},\ i < j$ $\quad$ & $m_{ij} + a_{ij} + a_{ji} + n_{ij} = 1$, for all $i < j$

Can estimate the likelihood of observing each dyad *independently.*

And, making an isomorphic homogeneity assumption leads to:

$$p_1(\mathrm{x}) = K * \exp\left[\underbrace{\rho M}_{\text{reciprocity}} + \underbrace{\theta L}_{\text{density}} + \sum \underbrace{\alpha_i X_{i+}}_{\text{activity}} + \sum \underbrace{\beta_i X_{+i}}_{\text{popularity}}\right]$$

- p2 – A random effects model where nodal in-/out- degree are regressed on nodal/dyadic covariates

Holland P, Leinhardt J. 1981. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*;76:33-65.

Modeled Interdependencies

Outcome		None	w/in Dyad[1]	Dyad +	
	Attribute	General Linear Model	Actor-Partner Independence (APIM)	Network Auto-Regression	Stochastic Actor Based Model (SABM, SAOM, Siena)
	Network	{Erdös-Renyi, Bernouli} Random	MR(QAP)[2], latent space, CUG[3], p1/p2	**ERGM** (p*), Relational Event	

[1] Frequently described as "dyad independence" models.

[2] R (sna) = qaptest

[3] Conditional Uniform Graphs

[4] R (sna) = lnam

We'll approach the model in the following steps:

- The general form of the model

Estimation

1. Available Terms

2. Model Diagnostics

Goodness of Fit

3. Simulation & Visualization
4. GoF Tests

ERGMs take the form of a probability distribution of graphs:

$$p(Y = y) = \frac{1}{\kappa(\theta)} \exp\left(\theta' z(y)\right)$$

- Y = [Yij] i.e., a system of tie variables
- y = realization of Y, i.e., the observed network
- $z(y)$ is a vector of network statistics (Qs)
- θ is a parameter vector corresponding to z(y)
- $\kappa(\theta)$ is a normalizing factor making p sum to 1
  - $\kappa(\theta)$ = Σexp{θ'z(y)} for all possible network configurations

- What is $\kappa(\theta)$?
  - y is undirected

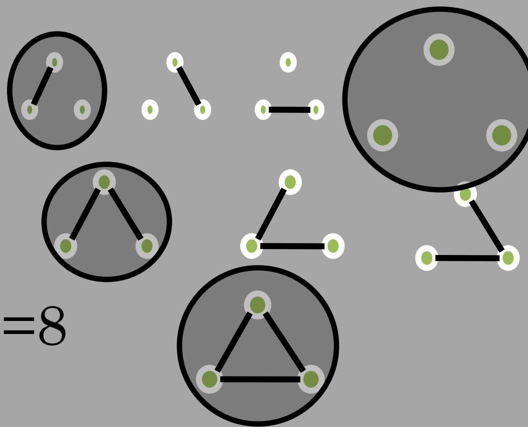
$$2^{\frac{n(n-1)}{2}}$$

  - y is directed

$$2^{n(n-1)}$$

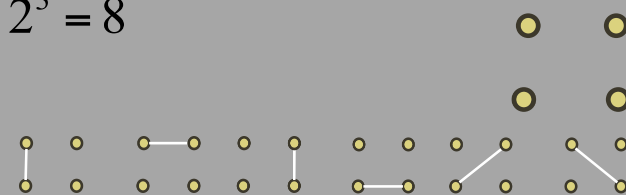
Homogeneity Assumption

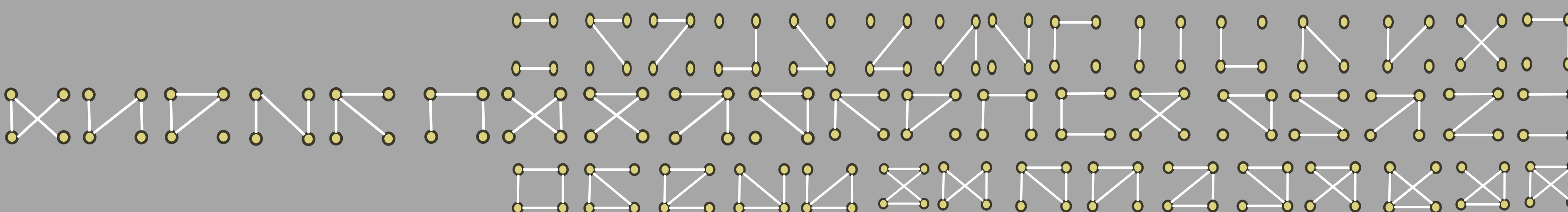
3=4, 4=11

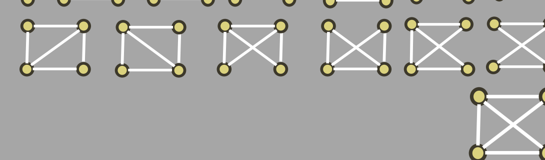


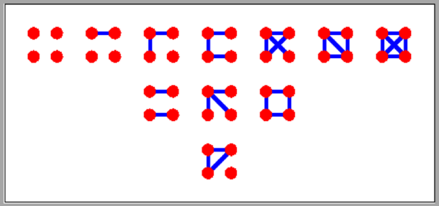
n= 3, $\kappa(\theta)$=8

$$2^{\frac{3(3-1)}{2}} = 2^3 = 8$$

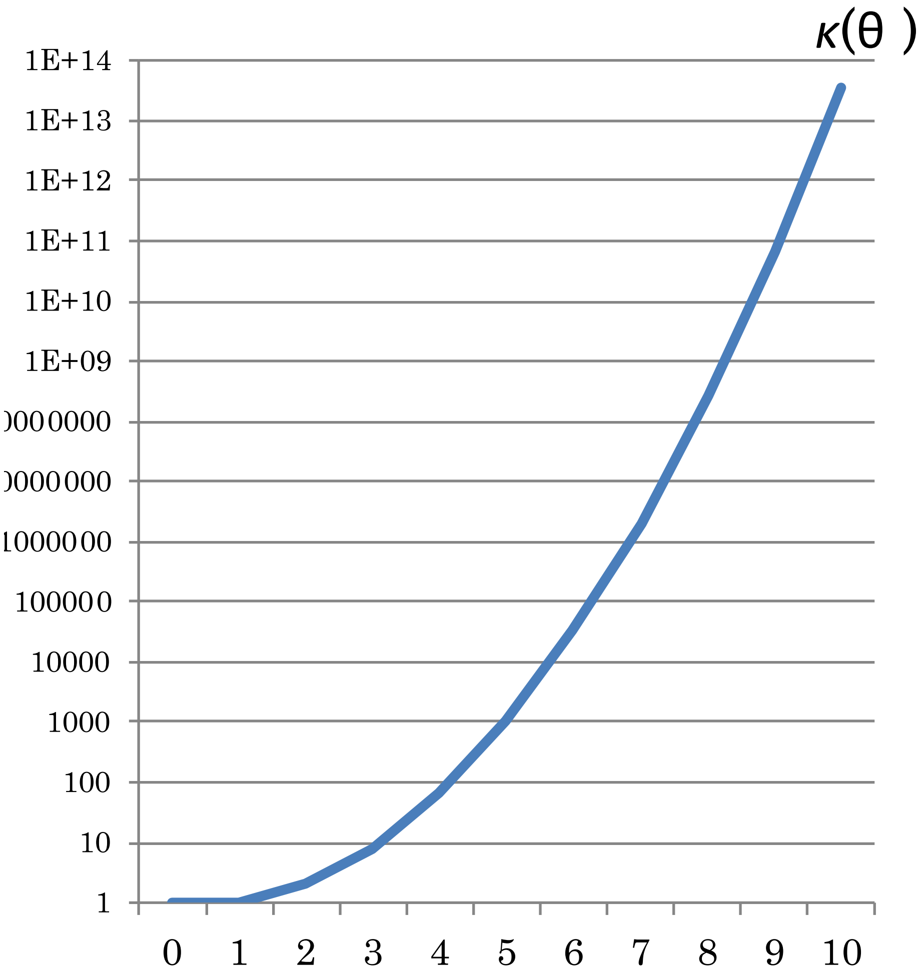








$$2^{\frac{4(4-1)}{2}} = 2^6 = 64$$



Source: David Hunter

Or if you prefer, for 34 nodes:

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
configurations

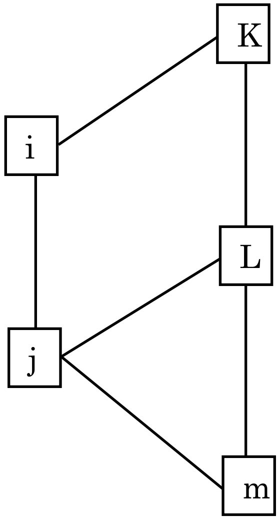
i.e., $\kappa(\theta)$ is virtually impossible to specify for graphs of any considerable size. That’s a problem.

Regression model assumption #1 (or 2) is independence of observations. So,

- For ego networks from random population samples, this approach typically isn't violated, and is thus appropriate.
- For any connected networks however, those *network variables* are *by definition* dependent, and thus violate the assumptions of the model, leaving us needing an alternative.

Exponential Random Graph Model
A Quick Diversion – Why not just a Logit?

		nodematch	kstar(2)	triangle	tie
i	j	1	3	0	1
i	K	0	2	0	1
i	L	0	5	2	0
i	m	1	4	1	0
j	K	0	5	2	0
j	L	0	4	1	1
j	m	1	3	1	1
K	L	1	3	0	1
K	m	0	4	1	0
L	m	0	3	1	1



nodematch(capitalized) + kstar(2) + triangle

```
> # compare the output from the two approaches
> cbind(ex1a$coef,ex2a$coef)  # nodematch terms are identical
                 glm          ergm
(Intercept) 0.000000 -3.700720e-16
nm          1.098612  1.098612e+00
> cbind(ex1b$coef,ex2b$coef)  # kstar(2) terms are quite close (probably a miscount)
                  glm      ergm
(Intercept)  85.08265  77.59528
ks          -21.44395 -19.51945
> cbind(ex1c$coef,ex2c$coef)  # triangle terms NOT close (degeneracy in ERGM)
                  glm       ergm
(Intercept)  19.94404 -0.3771544
tr          -19.53857 -2.3943960
> cbind(ex1d$coef,ex2d$coef)  # nodematch and kstar(2) terms don't work well
together
                  glm        ergm
(Intercept) 165.06443 150.5996473
nm          -20.15293   0.8303108
ks          -41.26611 -37.8408808
```

Source: David Schaefer

- Can we estimate the ERGM without computing $\kappa(\theta)$?

$$Y_{i,j} = 1 \text{ or } 0\text{, whether the ij edge is present/absent}$$

$$Y_{i,j}^{+} = \text{ Y with i,j element forced to 1}$$

$$Y_{i,j}^{-} = \text{Y with ij element forced to 0}$$

$$Y_{i,j}^{c} = \text{ Y excluding } \mathrm{y}_{\mathrm{i,j}}$$

- The ERGM we specified estimates the total probability of *Y*:

$$p(Y=y)=\frac{1}{\kappa(\theta)}\exp\left(\theta' z(y)\right)$$

- The terms in the previous slide, we can state the conditional probability of an edge as:
  - Existing: $p\left(Y_{ij}=1 \middle| Y_{ij}^{c}\right)$
  - Absent: $p\left(Y_{ij}=0 \middle| Y_{ij}^{c}\right)$,which also can be noted as:

$$1-p\left(Y_{ij}=1 \middle| Y_{ij}^{c}\right)$$

  - Which allows us to specify the logit p* to model the log odds ratio of Y_{ij}=1 as:

$$\frac{p\left(Y_{ij}=1 \middle| Y_{ij}^{c}\right)}{p\left(Y_{ij}=0 \middle| Y_{ij}^{c}\right)}$$

- So if instead of our ERGM: $$p(Y=y)=\frac{1}{\kappa(\theta)}\exp\left(\theta' z(y)\right)$$

- We estimate the conditional probability of edges:

$$p(Y_{ij}=1\,|\,Y_{ij}^{c})=\frac{1}{\kappa(\theta)}\exp\left(\theta' z\left(Y_{i,j}^{+}\right)\right)$$

- Usefully: $$\log\frac{p\left(Y_{ij}=1\,\middle|\,Y_{ij}^{c}\right)}{p\left(Y_{ij}=0\,\middle|\,Y_{ij}^{c}\right)}=\frac{\cancel{\frac{1}{\kappa(\theta)}}\exp\left(\theta' z\left(Y_{ij}^{+}\right)\right)}{\cancel{\frac{1}{\kappa(\theta)}}\exp\left(\theta' z\left(Y_{ij}^{-}\right)\right)}=\theta' z\left(Y_{i,j}^{+}-Y_{i,j}^{-}\right)$$

- Thus: $$\log\left(\frac{p\left(Y_{ij}=1\,\middle|\,Y_{ij}^{c}\right)}{p\left(Y_{ij}=0\,\middle|\,Y_{ij}^{c}\right)}\right)=\theta'\Delta z(y_{ij})$$ no κ(θ)!

We'll approach the model in the following steps:

- The general form of the model

Estimation

1. Available Terms

2. Model Diagnostics

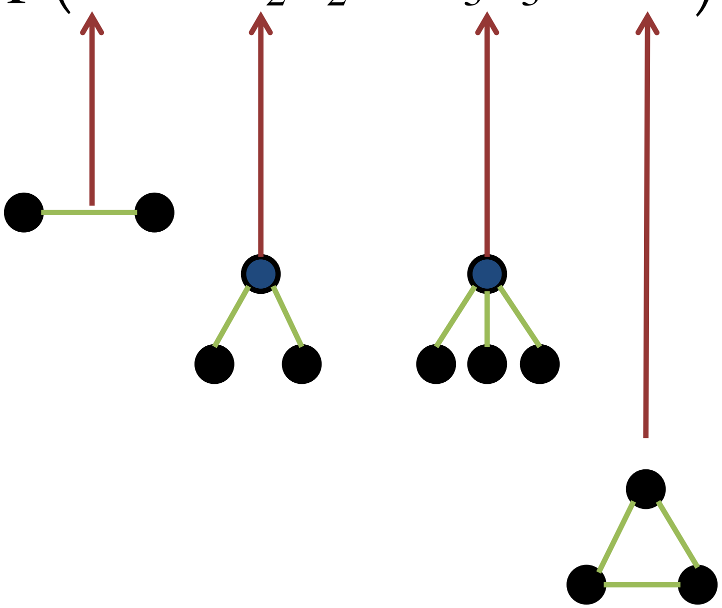
Goodness of Fit

3. Simulation & Visualization
4. GoF Tests

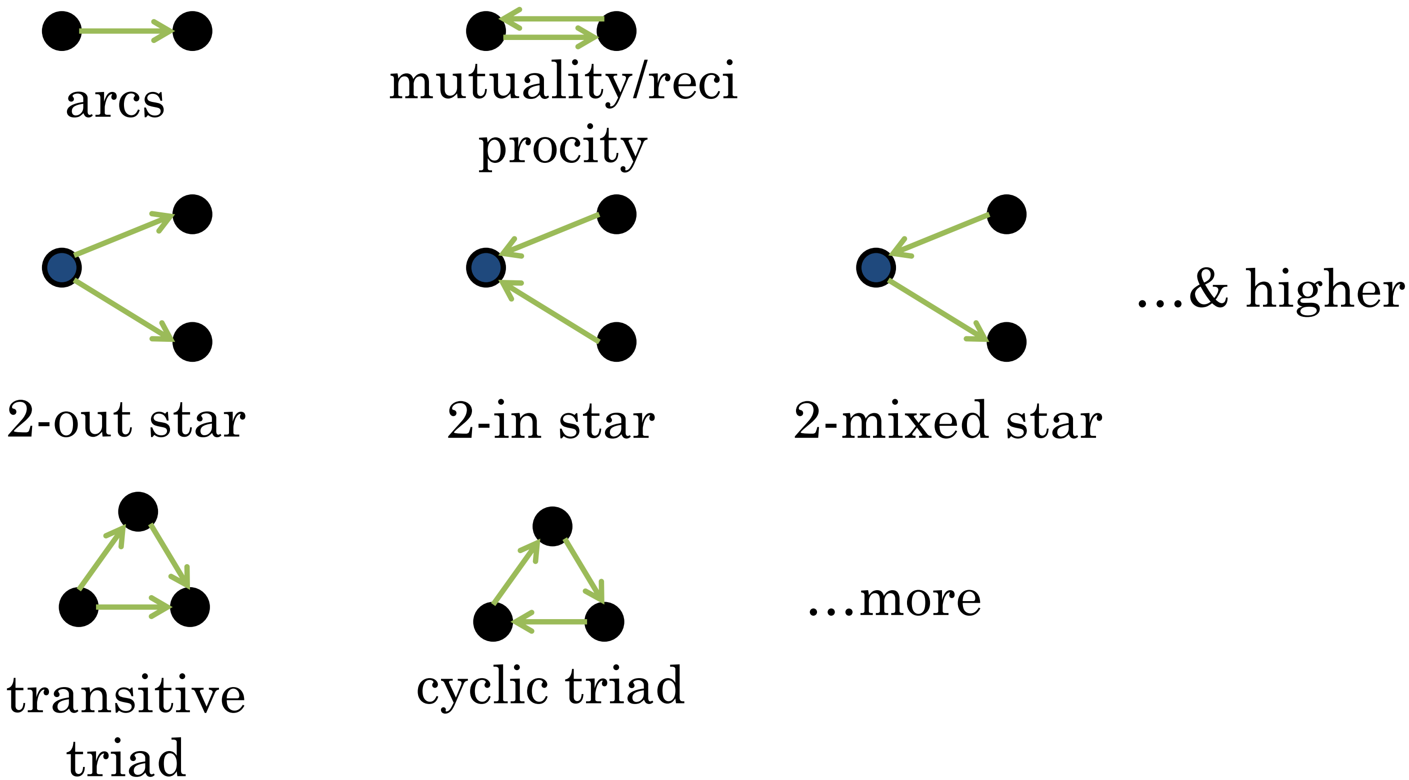
- A Markovian assumption:
  - Edges conditional iff they share a node
  - This model is then comprised of ties (edges), stars & triangles

$$p(Y=y)=\frac{1}{\kappa(\theta)}\exp\left(\theta L+\sigma_2 S_2+\sigma_3 S_3+\tau T\right)$$

- We estimate parameters:
  - Edge (θ) - propensities for the graph to have L edges
  - Star (σ_k) – propensity for individuals to have connections to k multiple partners
  - Triangles (τ) – propensity for local clustering

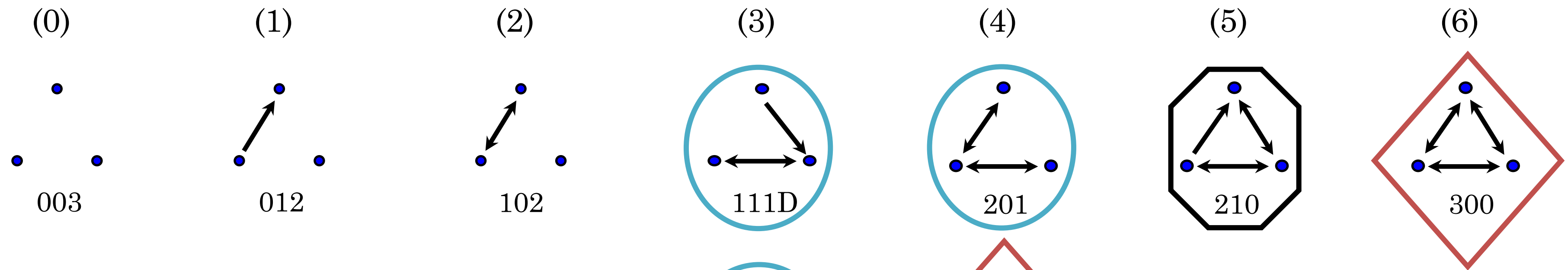


Adapting the Markov model for directed graphs.



Source: Pip Pattison

Exponential Random Graph Model
2. Parameterization



In theory you could use all triads

- typically modeled w/
  - transitive AND
  - intransitive

  or
  - presence/absence of a few – e.g., cycles & transitive triples

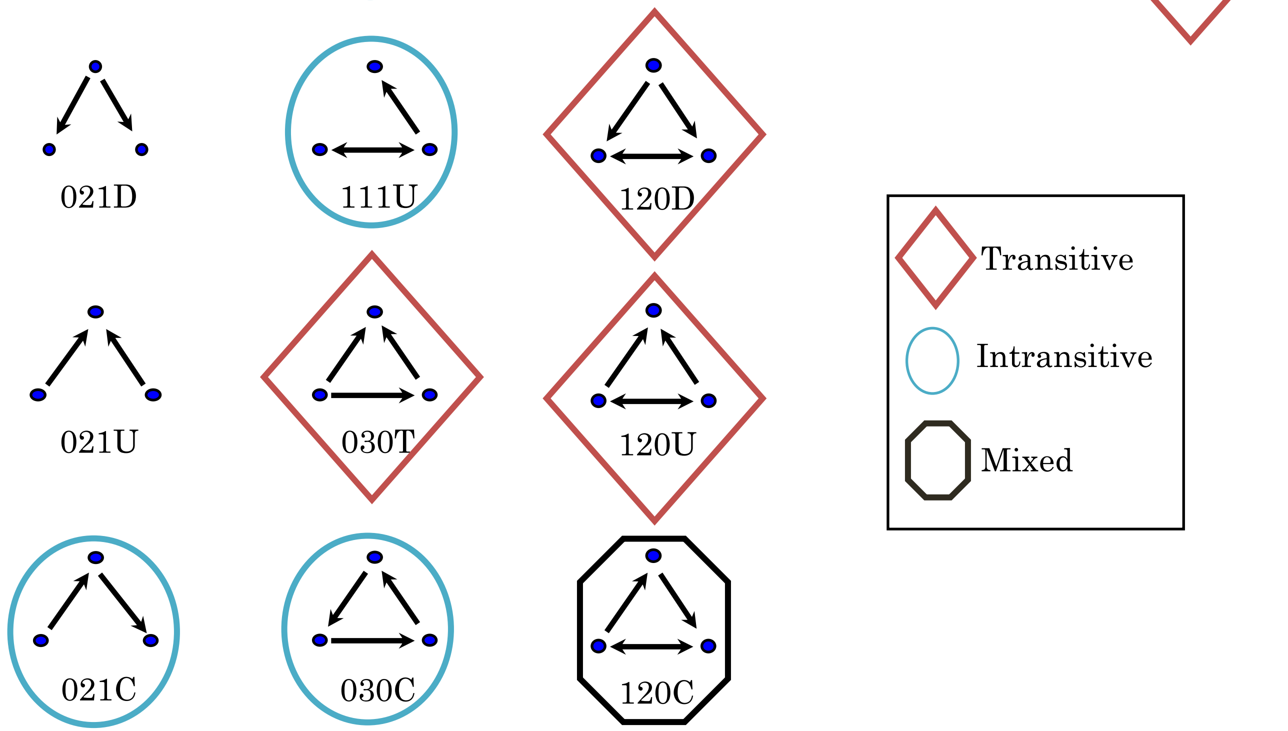


Image Source: Jim Moody

Using attribute data

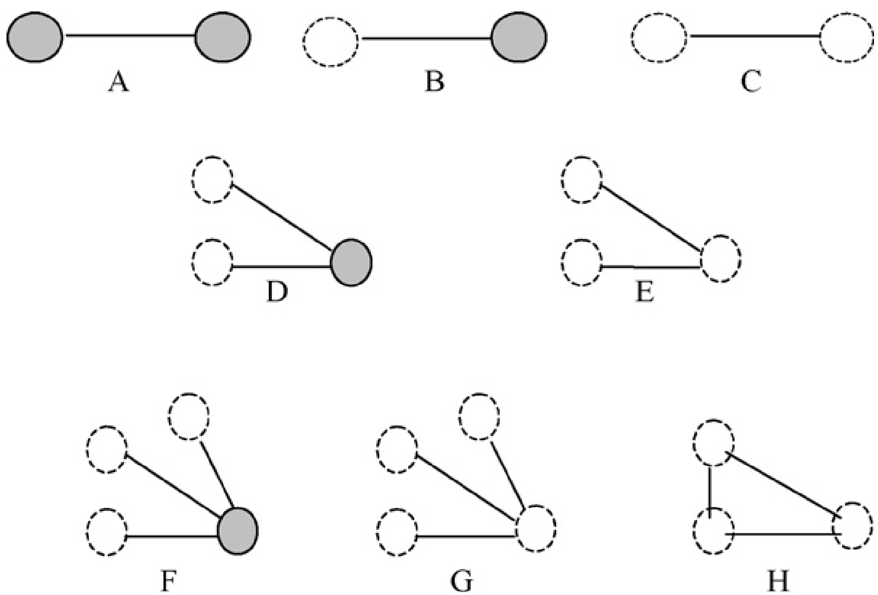
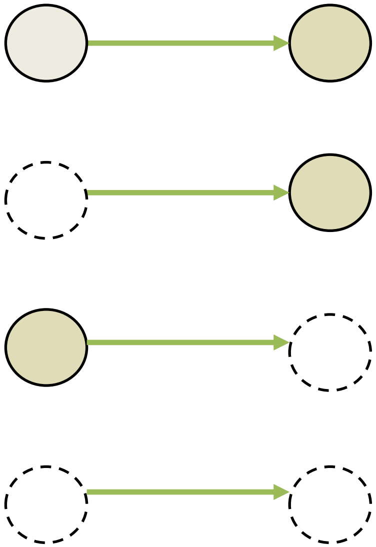


Fig. 2. (A–H) Configurations for a Markov attribute—Markov graph social selection model.



Robins G, Pattison P, Kalish Y, Lusher D. An Introduction to Exponential Random Graph (p*) Models for Social Networks. *Social Networks* 2007;29(2):173-191.

We'll approach the model in the following steps:

- The general form of the model

Estimation

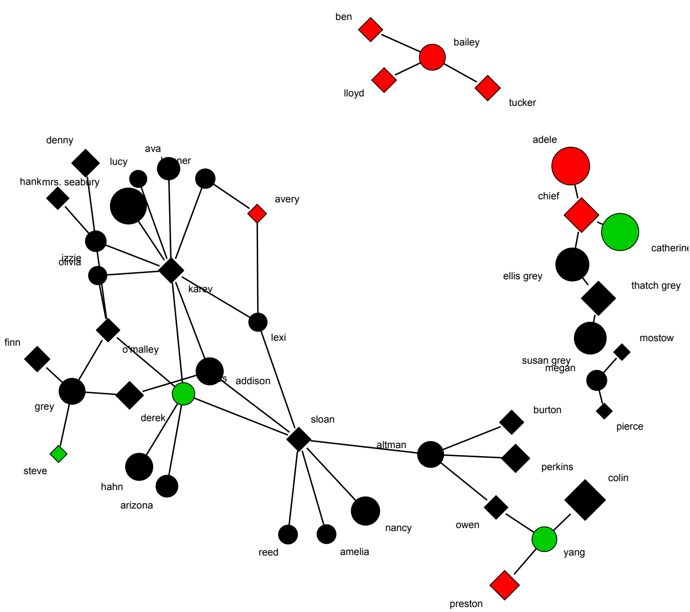
1. Available Terms
2. Model Diagnostics

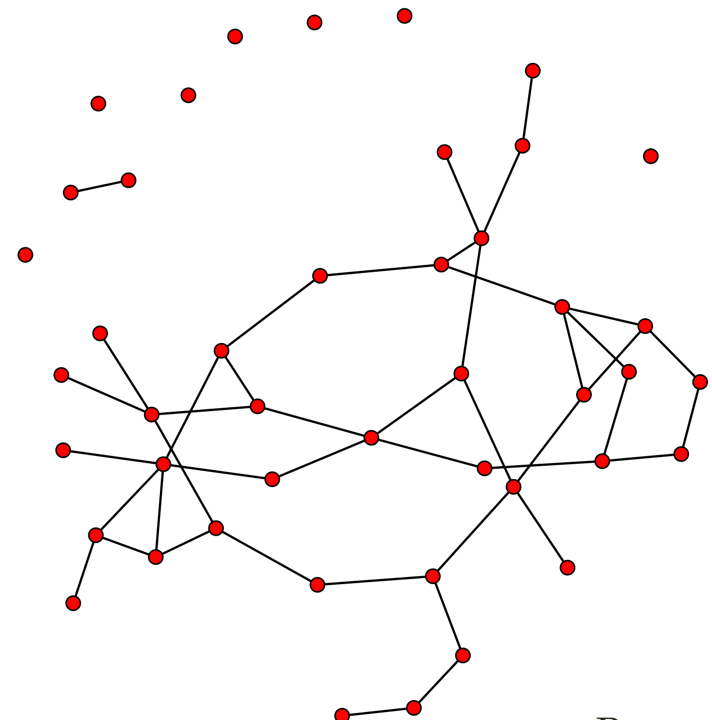
Goodness of Fit

3. Simulation & Visualization
4. GoF Tests

- So, you've fit a model and can interpret the coefficients, but is it a ***useful*** model?
  - Terms included
    - Theoretically derived – remember dependencies (but don't include complete combinations)
    - Can define your own if necessary
  - Examining "fit"
    - If the model fits the observed data well, networks sampled from the distribution it defines should resemble the observed network.
    - So, you can simulate networks from the model and plot those as a first "ocular test" of fit.

Grey's Anatomy Example

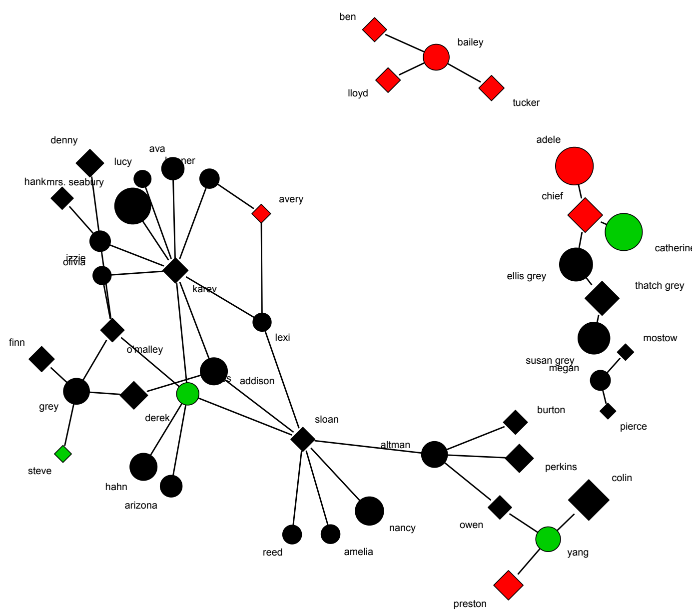


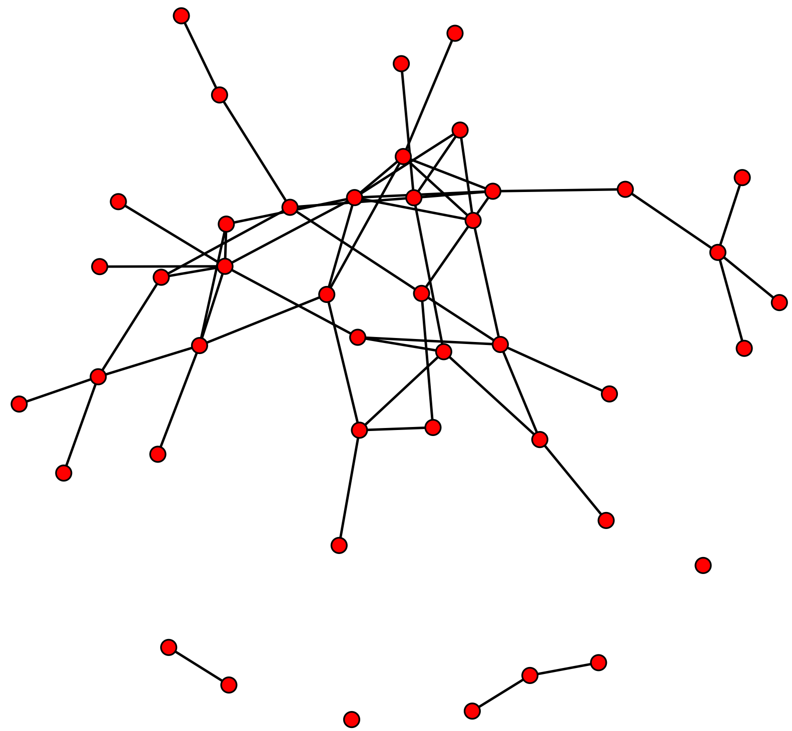


Bernoulli

ergm(greys~edges)

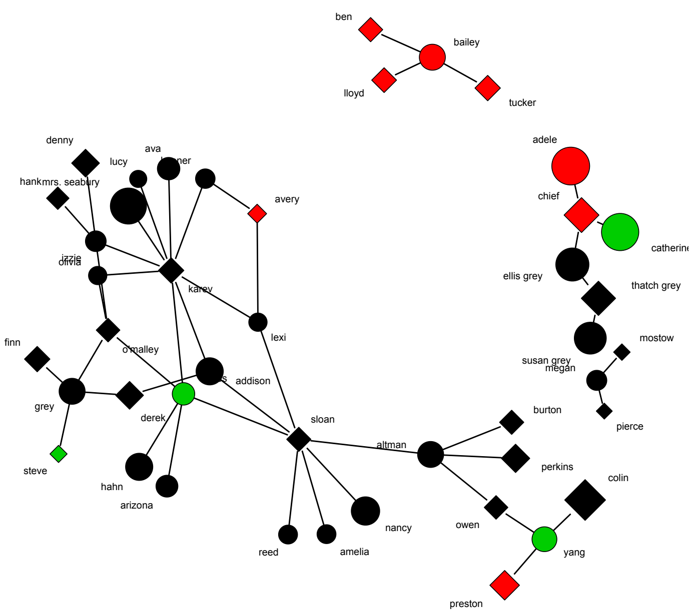
Grey's Anatomy Example

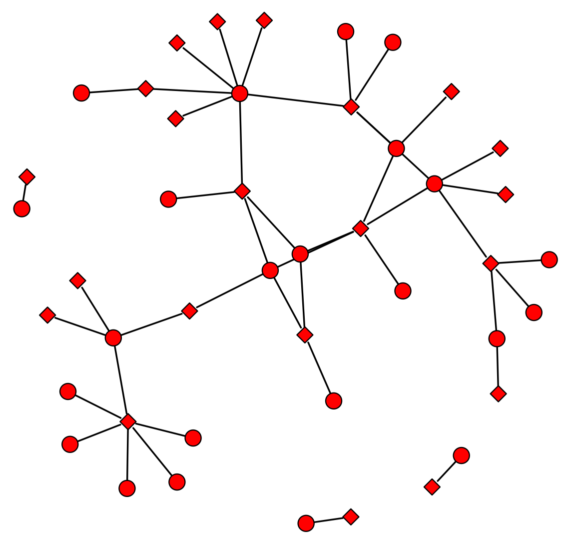




ergm(greys~edges+degree(1))

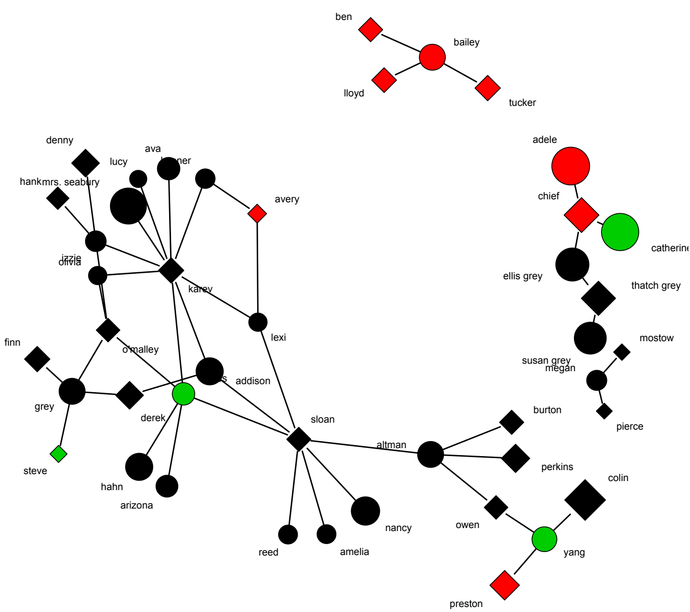
Grey's Anatomy Example

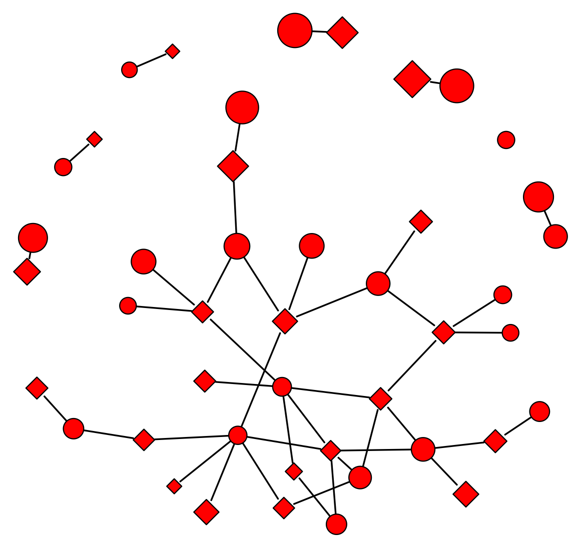




ergm(greys~edges+degree(1)
+nodematch("sex"))

Grey's Anatomy Example





ergm(greys~edges+degree(1)
+nodematch("sex")+absdiff("birthYear"))

We'll approach the model in the following steps:

- The general form of the model

Estimation

1. Available Terms
2. Model Diagnostics

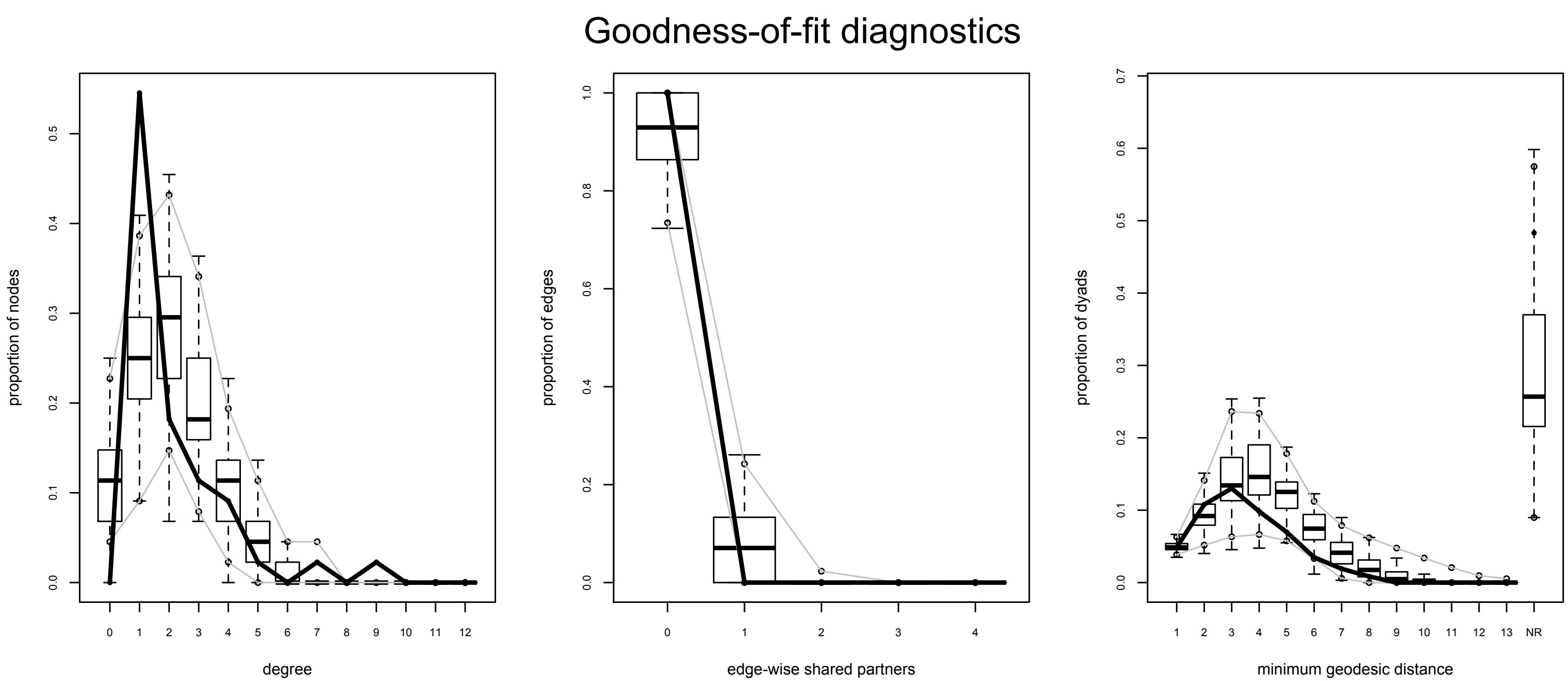
Goodness of Fit

3. Simulation & Visualization
4. GoF Tests

- Obviously, while the "ocular test" is a good first step, we'd like to more formally evaluate the fit of the model.
- First, examine AIC/BIC in same ways you're used to for logit models generally for those parameters specified in the model.
- Second, we typically check a few statistics that are *NOT* explicitly included in the model.
  - The defaults (if unspecified) in R's ergm include:
    - the degree distribution
    - edgewise shared partners
    - geodesic distance

Exponential Random Graph Model
4. Examine the Model w/ GoF Stats

ergm(greys~edges)

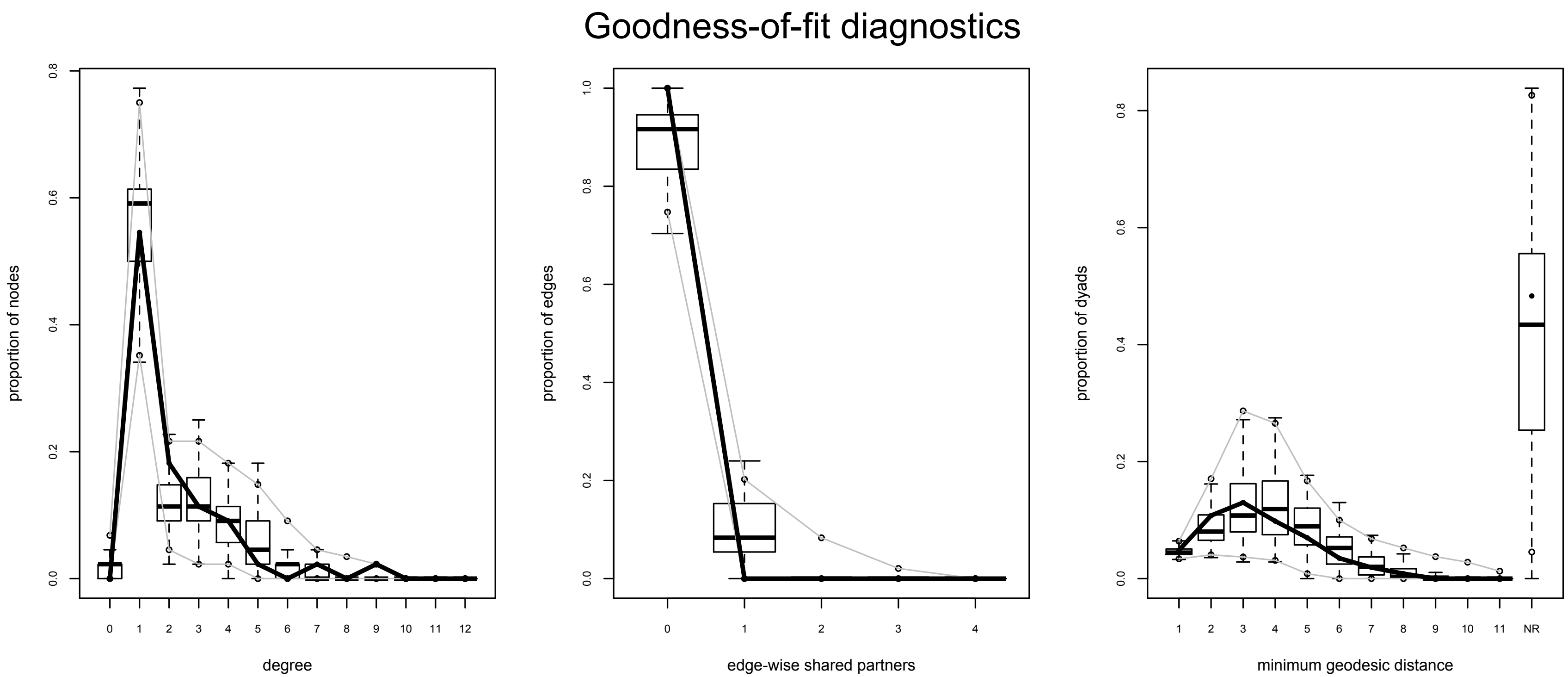


	obs	min	mean	max	MC p-value
0	0	0	4.91	12	0.02
1	24	3	10.71	18	0.00

	obs	min	mean	max	MC p-value
1	46	30	46.87	63	1.00
Inf	457	85	276.19	640	0.18

Exponential Random Graph Model
4. Examine the Model w/ GoF Stats

ergm(greys~edges+degree(1))

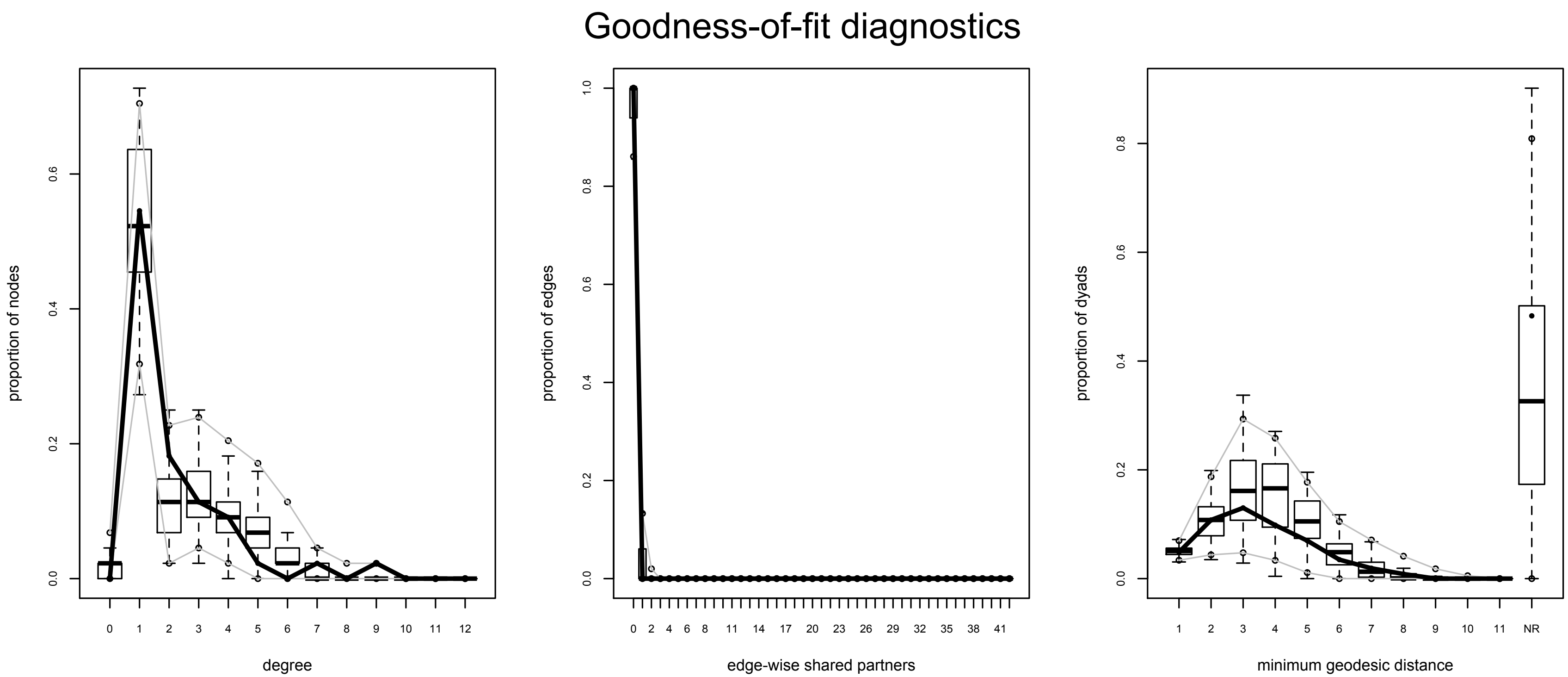


	obs	min	mean	max	MC p-value
0	0	0	0.88	3	0.88
1	24	13	24.97	34	0.80
2	8	1	5.16	12	0.26

	obs	min	mean	max	MC p-value
1	46	31	44.17	75	0.74
2	102	34	84.93	224	0.52
Inf	457	0	401.74	793	0.68

Exponential Random Graph Model
4. Examine the Model w/ GoF Stats

ergm(greys~edges+degree(1)+nodematch("sex"))

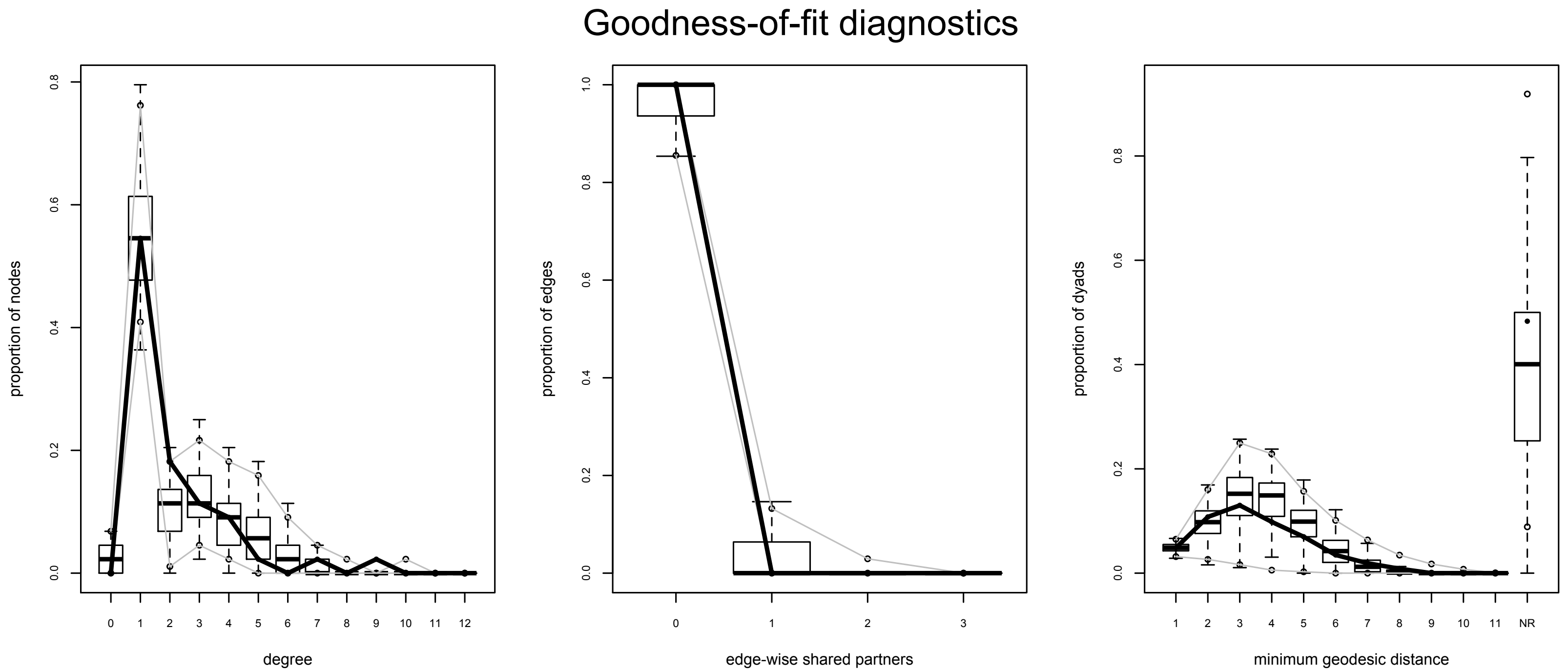


	obs	min	mean	max	MC p-value
0	0	0	0.81	6	0.98
1	24	12	23.26	32	0.98
2	8	1	5.15	13	0.40

	obs	min	mean	max	MC p-value
1	46	29	47.86	76	0.88
2	102	33	101.68	188	1.00
Inf	457	0	324.97	853	0.54

Exponential Random Graph Model
4. Examine the Model w/ GoF Stats

ergm(greys~edges+degree(1)+nodematch(“sex”)+absdiff(“birthyear”))



	obs	min	mean	max	MC p-value
0	0	0	1.07	3	0.72
1	24	16	24.38	35	1.00
2	8	0	4.69	9	0.18
3	5	1	5.28	11	1.00

	obs	min	mean	max	MC p-value
1	46	27	45.89	63	1.00
2	102	15	92.51	160	0.80
3	123	10	135.88	243	0.74
Inf	457	0	390.81	886	0.54

We'll approach the model in the following steps:

- The general form of the model

Estimation

1. Available Terms
2. Model Diagnostics

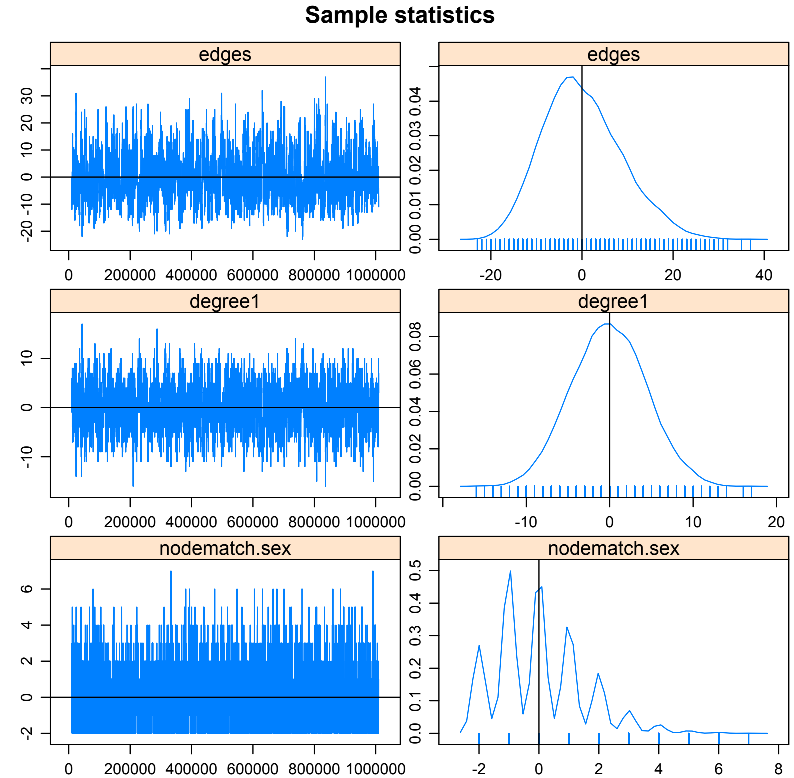
Goodness of Fit

3. Simulation & Visualization
4. GoF Tests

- Given the proposed set of z(y) statistics, the desire is to estimate θ with MLE
  - but recall, the normalizing constant needed is virtually impossible to calculate for most networks.
  - For dyadic independence models, the model can be approximated with logistic regression, where the MPLE is approximately equivalent to the MLE.
  - For dyadic dependence models, this doesn't work.
    - So we instead use Markov Chain Monte Carlo MLE.
      - MCMC approaches rely on iteratively sampling the probability space and comparing to observed graph
      - Iterations should lead to convergence (parameter estimates stabilize)

- Looking for MCMC convergence.
  - the tutorial provided guides you through a degenerate example

- For details on what to look for here, see Goodreau et al (2008) & Morris et al (2008) *JoSS* papers.




```
Are sample statistics significantly different from observed?
                edges     degree1 nodematch.sex Overall (Chi^2)
diff.      0.2116000 -0.00920000     0.0075000              NA
test stat. 0.6196627 -0.05348105     0.2929747       1.5434278
P-val.     0.5354799  0.95734863     0.7695415       0.6722841
```

- Model degeneracy is when the MCMC fails to converge
  - i.e., it's wandering the probability space and can't find parameters that approach the observed network's
- This can happen for a few reasons:
  - You've specified a model that doesn't fit the data.
  - It's been found MCMC estimation of Markov random graphs for certain parameter values lead to models that place almost all probability on empty or complete graphs.
    - **Especially likely for models with high clustering**
    - Can result from:
      - Markov assumption being too restrictive
      - The means of including transitivity in the model is too simplistic

Exponential Random Graph Model
Examine the Model Estimation

196 *G. Robins et al. / Social Networks 29 (2007) 192–215*

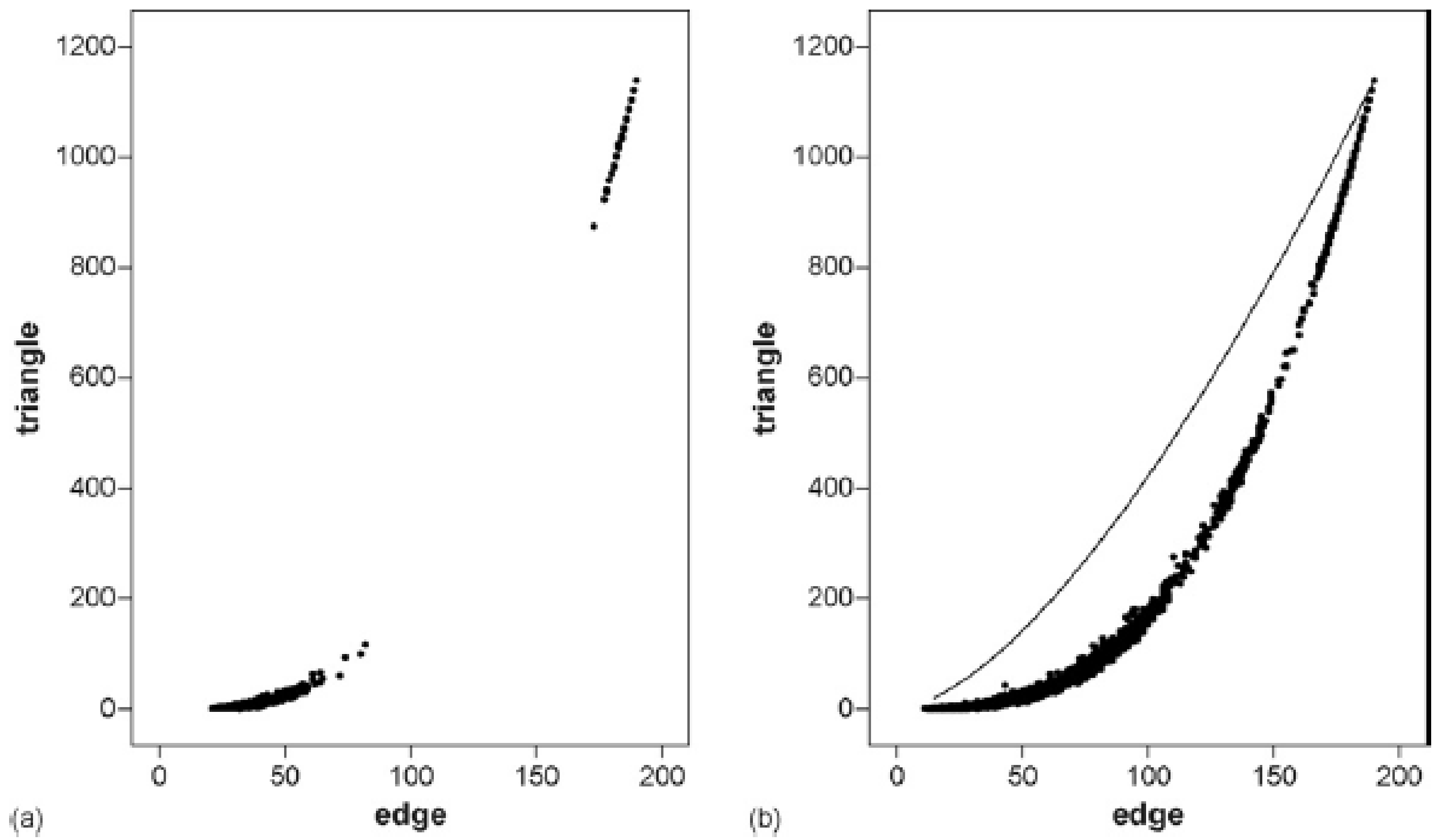
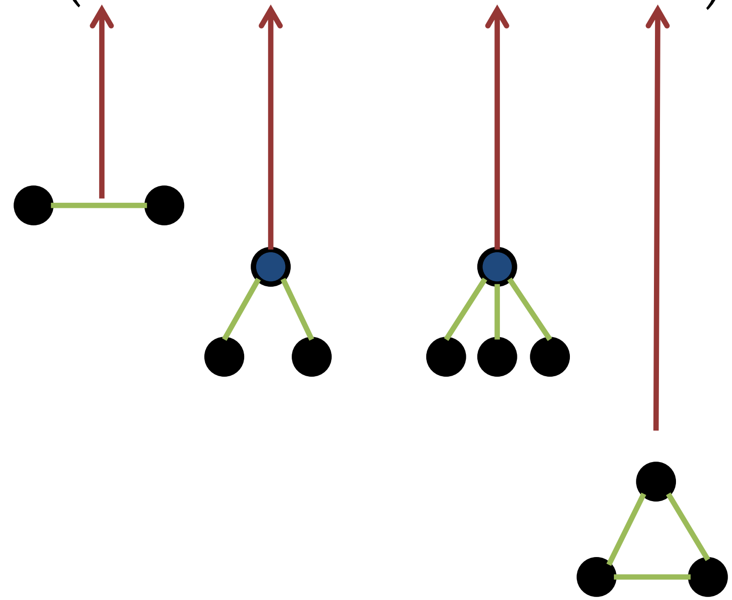


Fig. 1. Scatterplot for simulation study of edge/triangle Markov graph model: number of edges plotted against number of triangles for $\tau = 0.0$–1.0 in steps of 0.1. Left panel (a) $\theta = -1.5$; right panel (b) $\theta = -2.0$ to 0.0 in steps of 0.5.

Robins G, Snijders TAB, Wang P, Handcock MS, Pattison P. Recent Developments in Exponential Random Graph (p*) Models for Social Networks. *Social Networks* 2007;29:192-215

- Remember the Markovian assumption:
  - Edges conditional iff they share a node
  - This model is then comprised of ties, stars & triangles

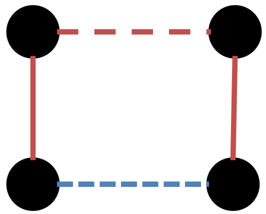
$$p(Y=y)=\frac{1}{\kappa(\theta)}\exp\left(\theta L+\sigma_2 S_2+\sigma_3 S_3+\tau T\right)$$



- We estimate parameters:
  - Edge (θ) - propensities for the graph to have L edges
  - Star (σ_k) – propensity for individuals to have connections to k multiple partners
  - Triangles (τ) – propensity for local clustering

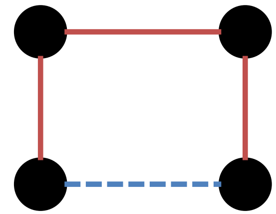
Exponential Random Graph Model
(Re-)Parameterization

Markov-Dependence assumption



Edges conditional iff they share a node

Partial-Conditional Dependence assumption

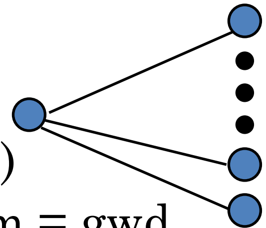


Edges conditional iff they share a node

Snijders TAB, Pattison PE, Robins GL, Handcock MS. New Specifications for exponential Random Graph Models. *Sociological Methodology* 2005;35:99-153.

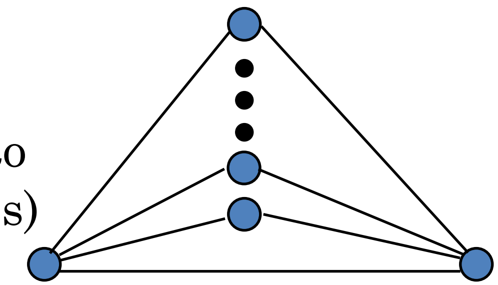
Exponential Random Graph Model
(Re-)Parameterization

k-star
(degree distribution)



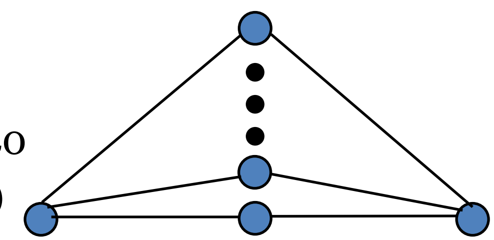
ergm = gwd

k-triangles
(propensity to
close triangles)



ergm = gwesp

k-2paths
(propensity to
close cycles)



ergm = gwdsp

- In ergm, these are specified with a series of geometrically weighted terms for degree, edge-wise shared partners and dyad-wise (regardless of tie), shared partners.
- The weighting accounts for the size of k.
- Moreover, it is specified with a tuning factor (α) which allows you temper the effect of triad closure on tie probability for an increasing number of shared partners.

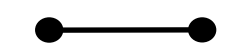
Snijders TAB, Pattison PE, Robins GL, Handcock MS. New Specifications for exponential Random Graph Models. *Sociological Methodology* 2005;35:99-153.

- The number of terms available is large & growing, and their justifications/dependencies increasingly complicated.
  - best motivated/selected ***theoretically***
- The best thing to do is to read examples, and work your way through the available tutorials/help files to make sense of what makes sense.
  - See especially
    - Morris et al 2008 and
    - help("ergm-terms", package=ergm)
- You can even construct your own

Exponential Random Graph Model (Re-)Parameterization

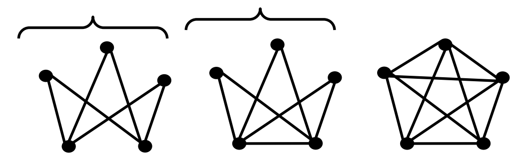
skip?

- Bernoulli:
  - $Z_0\{i,j\} = Z_0\{k,l\}$

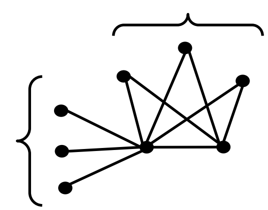


- Markov:
  - $Z_0\{i,j\} \cap Z_0\{k,l\} \neq \varnothing$

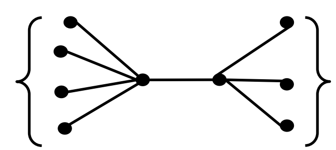
- Partial conditional Dependence:
  - $Z_1\{i,j\} \supseteq Z_0\{k,l\}$ and $Z_1\{k,l\} \supseteq Z_0\{i,j\}$



- Degree/closure interaction:
  - $Z_1\{i,j\} \supseteq Z_0\{k,l\}$ or $Z_1\{k,l\} \supseteq Z_0\{i,j\}$



- Three-path:
  - $Z_1\{i,j\} \cap Z_0\{k,l\} \neq \varnothing \Leftrightarrow$
    $Z_1\{k,l\} \cap Z_0\{i,j\} \neq \varnothing$



Source: Pip Pattison

Exponential Random Graph Model
"Real" World Example

Table 1. *ERG Model of Glee & Jefferson High Romantic Relationships/Hookups.*

	Glee's "McKinley High"	Add Health's "Jefferson High"
Edges	1.45	–1.05
	(1.77)	*(0.92)*
Same Gender	–2.90***	–4.53***
	(0.70)	*(0.49)*
Degree = 1	3.33*	3.56***
	(1.57)	*(0.57)*
2-Star	0.65	1.35***
	(1.00)	*(0.28)*
3-Path	–0.03	–0.02
	(0.19)	*(0.03)*
GWNSP	–0.40*	–0.53***
	(0.19)	*(0.15)*

Note: Presented are coefficients and *(standard errors)*. ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.

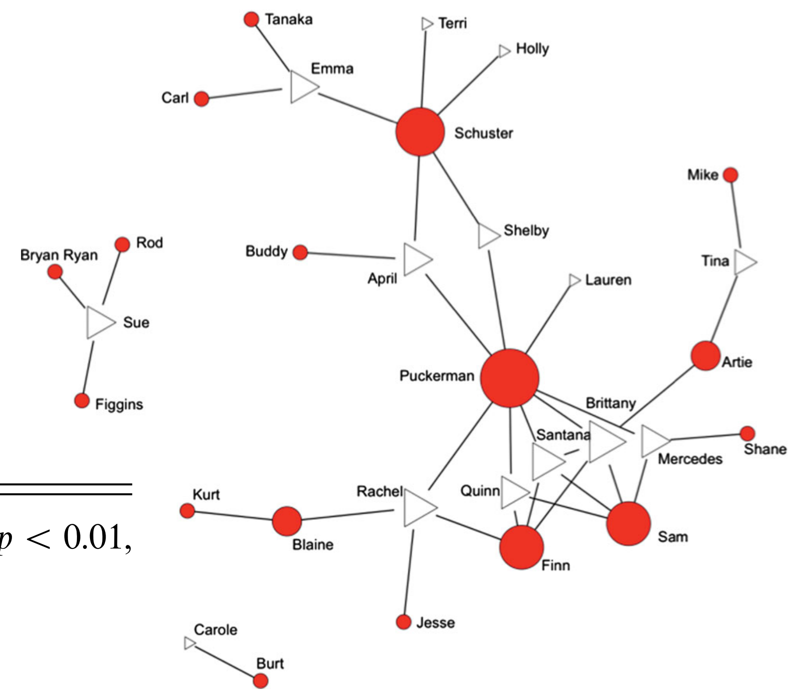
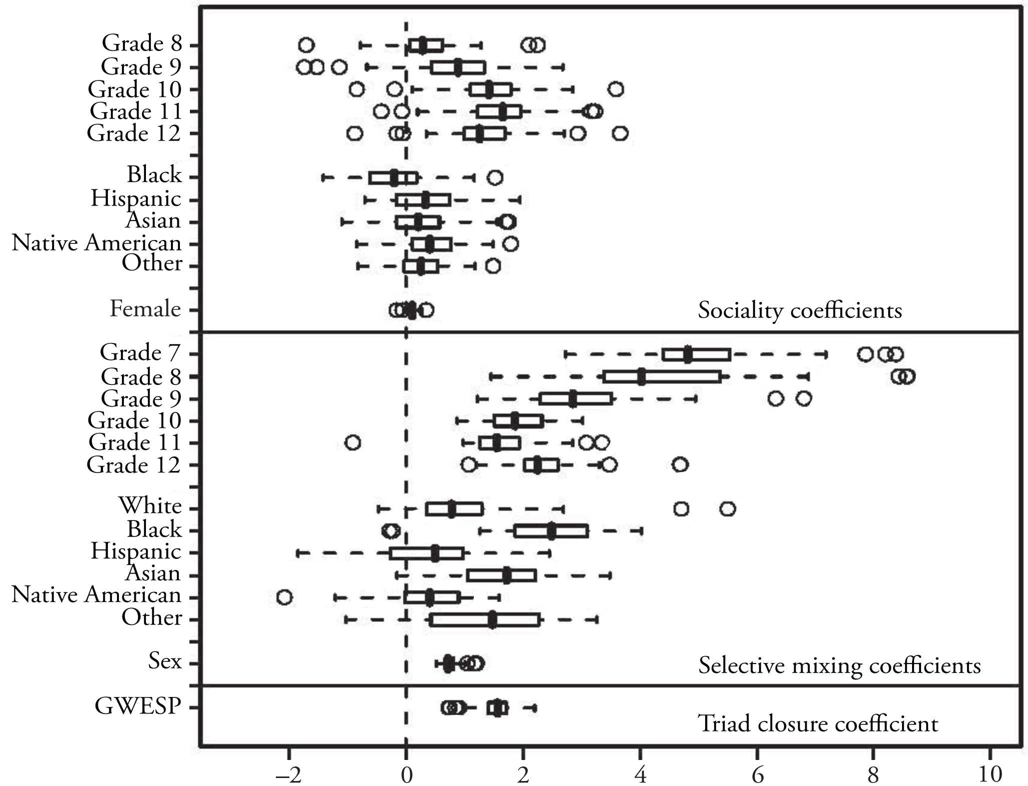


Fig. 1. Romantic and hook-up ties in seasons 1–2 of *Glee*.
Node shading/shape represents gender: male = filled/circle, female = open/triangle; node size is proportional to the natural log of degree. (color online)

adams j. 2015. "Glee's McKinley High: Following Middle America's Sexual Taboos." *Network Science* 3(2): 293-295.

Exponential Random Graph Model
Real World Example

Figure 3. Coefficients From the Full Model, Plotted Across All 59 Schools



Notes: Boxplots follow the Tukey method. Boxes represent quartiles; whiskers extend to the most extreme data point within 1.5 times the interquartile range from the edge of the box; and points represent outliers.

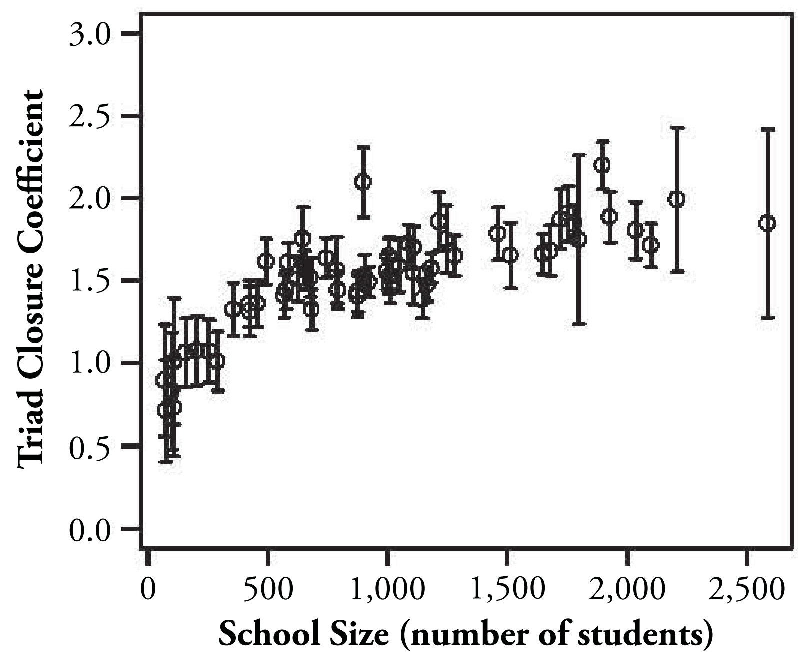
Goodreau SM, Kitts JA, Morris M.

Birds of a Feather or Friend of a Friend? Using Exponential Random Graph Models to Investigate Adolescent Social Networks.

Demography 2009;46(1):103-125.

Exponential Random Graph Model
Real World Example

Figure 4. Triad Closure (GWESP) Coefficient: Full Model



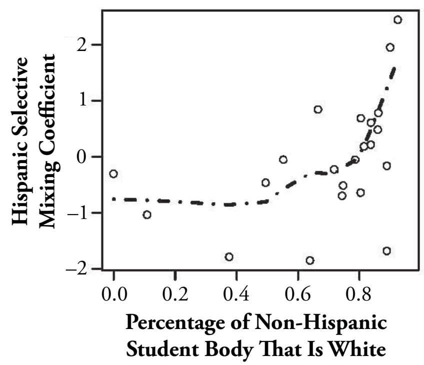
Goodreau SM, Kitts JA, Morris M.

Birds of a Feather or Friend of a Friend? Using Exponential Random Graph Models to Investigate Adolescent Social Networks.

Demography 2009;46(1):103-125.

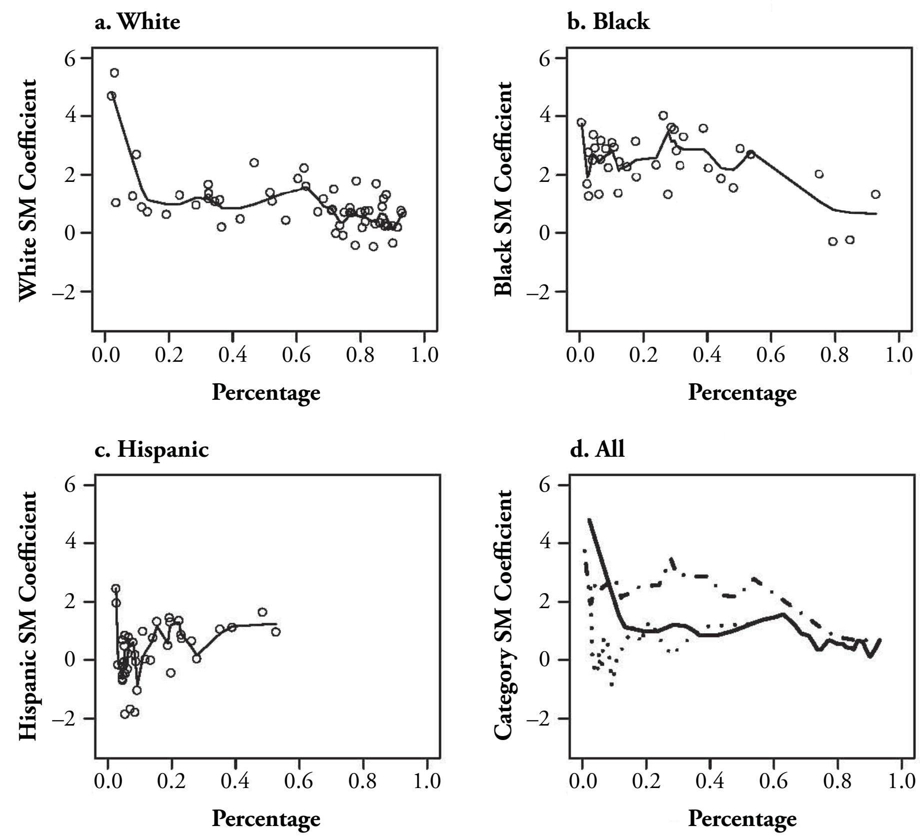
Exponential Random Graph Model
Real World Example

Figure 7. Hispanic Selective Mixing, by Proportion White: Full Model



Note: The line represents a lowess curve.

Figure 6. Selective Mixing (SM) Coefficients, by Proportion of School in Category: Full Model



Note: Lines represent lowess curves.

Goodreau SM, Kitts JA, Morris M.

Birds of a Feather or Friend of a Friend? Using Exponential Random Graph Models to Investigate Adolescent Social Networks.

Demography 2009;46(1):103-125.

We'll approach the model in the following steps:

- The general form of the model

Estimation

1. Available Terms
2. Model Diagnostics

Goodness of Fit

3. Simulation & Visualization
4. GoF Tests