Tapered Exponential-family Models for Social Networks

Mark S. Handcock

Department of Statistics
University of California - Los Angeles

Joint work with

Ian E. Fellows, Fellows Statistics
Andrea Wang, UCLA
Bart Blackburn, UCLA

Supported by NIH Grant HD041877, NSF awards CCF-2200197, MMS-0851555, SES-1357619, IIS-1546259.

UC-Irvine Statistics, May 18, 2023

Statistical Models for Social Networks

Notation

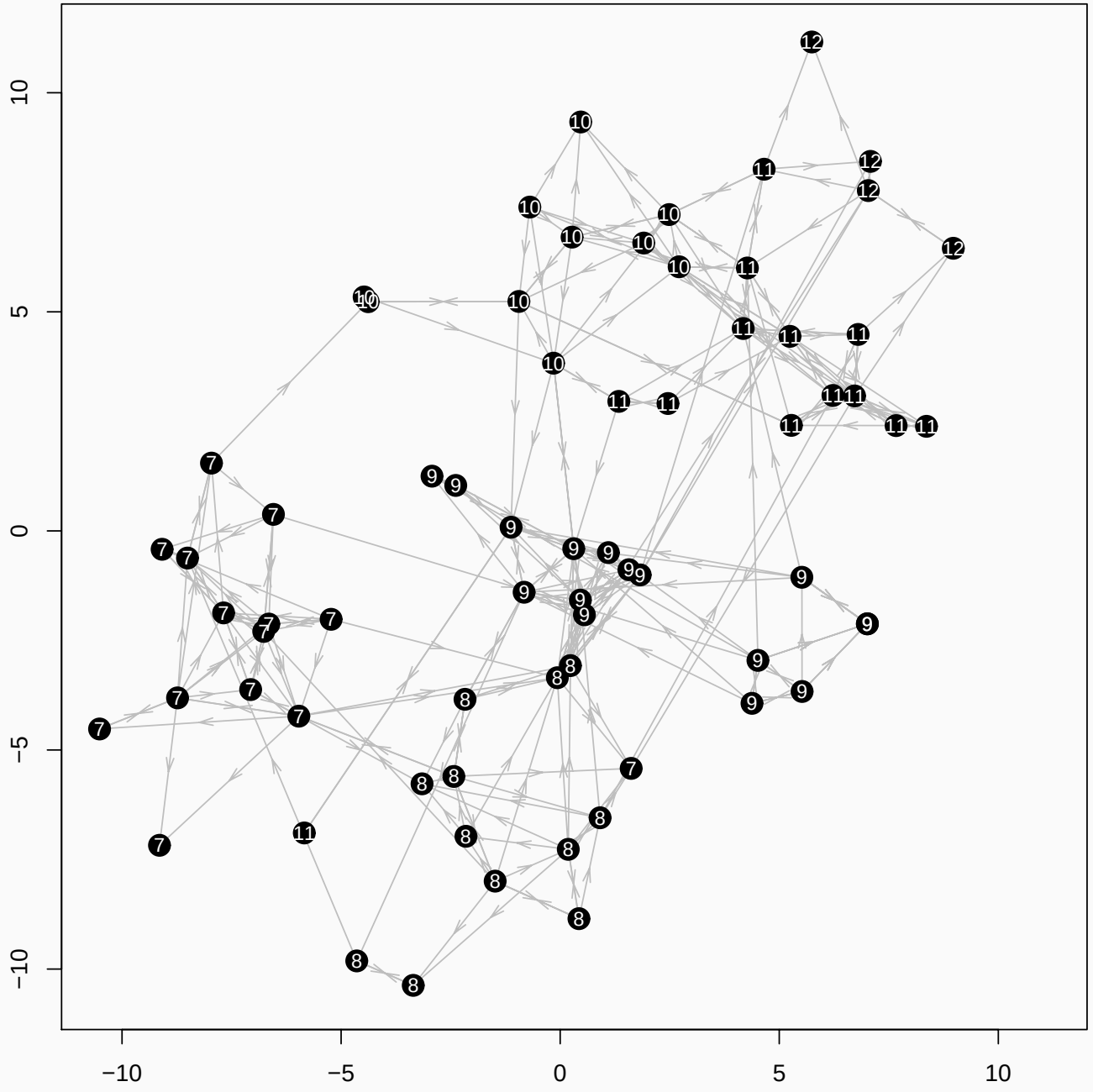
A *social network* is defined as a set of n social "actors", a social relationship between each pair of actors, and a set of variables on those actors/pairs.

$$Y_{ij} = \begin{cases} 1 & \text{relationship from actor } i \text{ to actor } j \\ 0 & \text{otherwise} \end{cases}$$

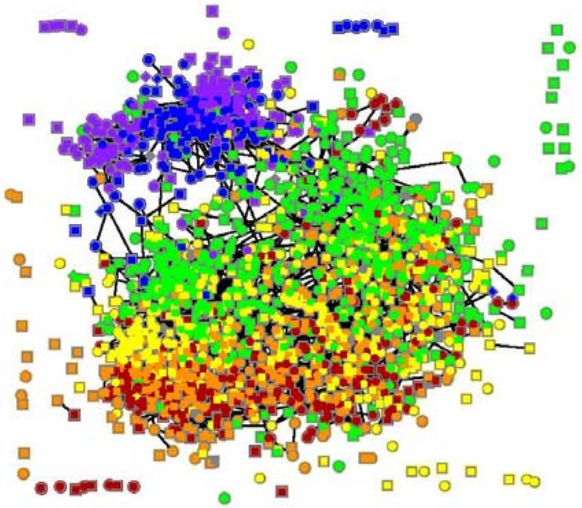
- call $Y \equiv [Y_{ij}]_{n \times n}$ a *graph*
  - a $N = n(n-1)$ binary array
- X be $n \times q$ matrix of actor variates
- call (Y, X) a *network*
- The basic problem of stochastic modeling is to specify a distribution for X, Y i.e., $P(Y = y, X = x)$

Examples of Friendship Relationships

- The National Longitudinal Study of Adolescent Health
  $\Rightarrow$ `www.cpc.unc.edu/projects/addhealth`
  - "Add Health" is a school-based study of the health-related behaviors of adolescents in grades 7 to 12.
- Each nominated up to 5 boys and 5 girls as their friends
- 160 schools: Smallest has 69 adolescents in grades 7–12

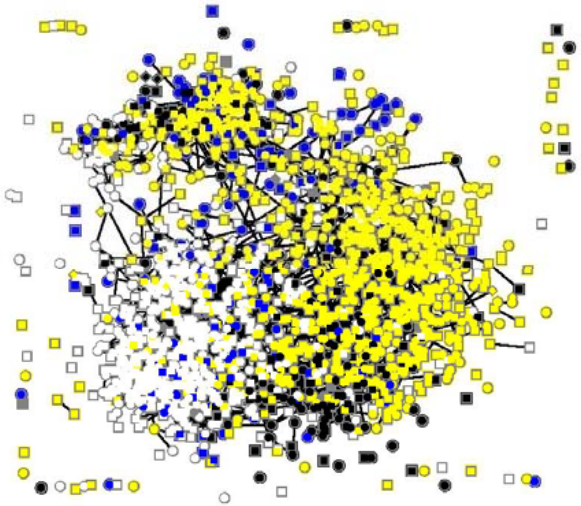

10
5
0
−5
−10
−10
−5
0
5
10

School Community Stratum 44
mutual friendships by Grade



2209 Students

School Community Stratum 44
mutual friendships by Race



2209 Students

- **Grade 7**
- **Grade 8**
- **Grade 9**
- **Grade 10**
- **Grade 11**
- **Grade 12**
- **Grade NA**

- **White** (non-Hispanic)
- **Black** (non-Hispanic)
- **Hispanic** (of any race)
- **Asian / Native Am / Other** (non-Hispanic)
- **Race NA**

Common Features of Social Networks

- *Mutuality* of ties
- *Individual heterogeneity* in the propensity to form ties
- *Homophily* by actor attributes

  $\Rightarrow$ Lazarsfeld and Merton, 1954; Freeman, 1996; McPherson et al., 2001

  - higher propensity to form ties between actors with similar attributes
    e.g., age, gender, geography, major, social-economic status
  - attributes may be observed or unobserved
- *Transitivity* of relationships
  - friends of friends have a higher propensity to be friends
- *Social Context* is important $\Rightarrow$ Simmel (1908), Heider (1946)
  - triad, not the dyad, is the fundamental social unit
- dependence on nodal and dyadic attributes

The ERGM Framework for Graph Modeling

Let $\mathcal{Y}$ be the sample space of Y e.g. $\{0,1\}^N$.

- $g(y)$, $y \in \mathcal{Y}$ d-vector of *graph statistics*
- represent graph features of interest (e.g., density, transitivity)
- desire $g(Y)$ to be jointly sufficient for the model

The ERGM Framework for Graph Modeling

Let $q(y)$ be a probability mass function over $\mathcal{Y}$.
Recall the maximum entropy motivation for exponential-families:

$$\underset{q}{\text{maximize}} \quad \sum_{y} q(y) \log(q(y))$$

subject to

$$E_q(g_i(Y)) = \mu_i, \qquad \forall\ i \in \{1, \ldots, d\}$$

Leads to:

$$P_\eta(Y = y) = \frac{\exp\{\eta \cdot g(y)\}}{c(\eta, \mathcal{Y})} \qquad y \in \mathcal{Y}$$

$$E_\eta(g(Y)) = \mu$$

The ERGM Framework for Network Modeling

Let $\mathcal{Y}$ be the sample space of Y e.g. $\{0,1\}^N$
and $\mathcal{X}$ be the sample space of X.

Model the multivariate distribution of Y given X via:

$$P_\eta(Y = y|X = x) = \frac{\exp\{\eta \cdot g(y|x)\}}{c(\eta, x, \mathcal{Y})} \qquad y \in \mathcal{Y}, \quad x \in \mathcal{X}$$

Frank and Strauss (1986)

- $\eta \in \Lambda \subset R^d$ d-vector of parameters
- $g(y|x)$ d-vector of *graph statistics*.
  $\Rightarrow$ $g(Y|x)$ are jointly sufficient for the model
- $c(\eta, x, \mathcal{Y})$ distribution normalizing constant

$$c(\eta, x, \mathcal{Y}) = \sum_{y \in \mathcal{Y}} \exp\{\eta \cdot g(y|x)\}$$

Extensive development of conditional models

- Classes of $g(y|x)$ (Generative Theory, Structural signatures)
- Inference on the log-likelihood function,

$$\ell(\eta|y_{\text{obs}}; x_{\text{obs}}) = \eta \cdot g(y_{\text{obs}}|x_{\text{obs}}) - \log c(\eta|x_{\text{obs}})$$

$$c(\eta|x_{\text{obs}}) = \sum_{z \ in \ \mathcal{Y}} \exp\{\eta \cdot g(z|x_{\text{obs}})\}$$

- For computational reasons, approximate the likelihood via Markov Chain Monte Carlo (MCMC)

How can we tell if a model class is useful?

Many aspects:

- Is the model-class itself able to represent a range of realistic networks?
  - *model degeneracy:* small range of graphs covered as the parameters vary

    Much work: Strauss, 1986; Jonasson, 1999; Handcock, 2003; Rinaldo, Fienberg and Zhou, 2009; Schweinberger 2011; Schweinberger ...

Model Degeneracy

idea: A random graph model is *near degenerate* if the model places almost all its probability mass on the boundary of the convex hull of $\{g(y|x) : y \in \mathcal{Y}\}$.

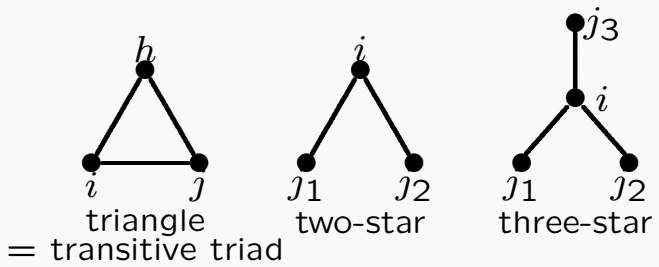
e.g. empty graph, full graph, no 2−stars, mono-degree graphs

- Example: The *triangle* model

$$P_\eta(Y = y) = \frac{\exp\{\eta_1 \text{edge}(y) + \eta_2 \text{triangle}(y)\}}{c(\eta_1, \eta_2)} \qquad y \in \mathcal{Y}$$

  is near-degenerate for most values of $\eta_2 > 0$

$$\text{edge}(y) = \sum_{i<j} y_{ij} \qquad \text{triangle}(y) = \sum_{i<j<k} y_{ij} y_{ik} y_{jk}$$

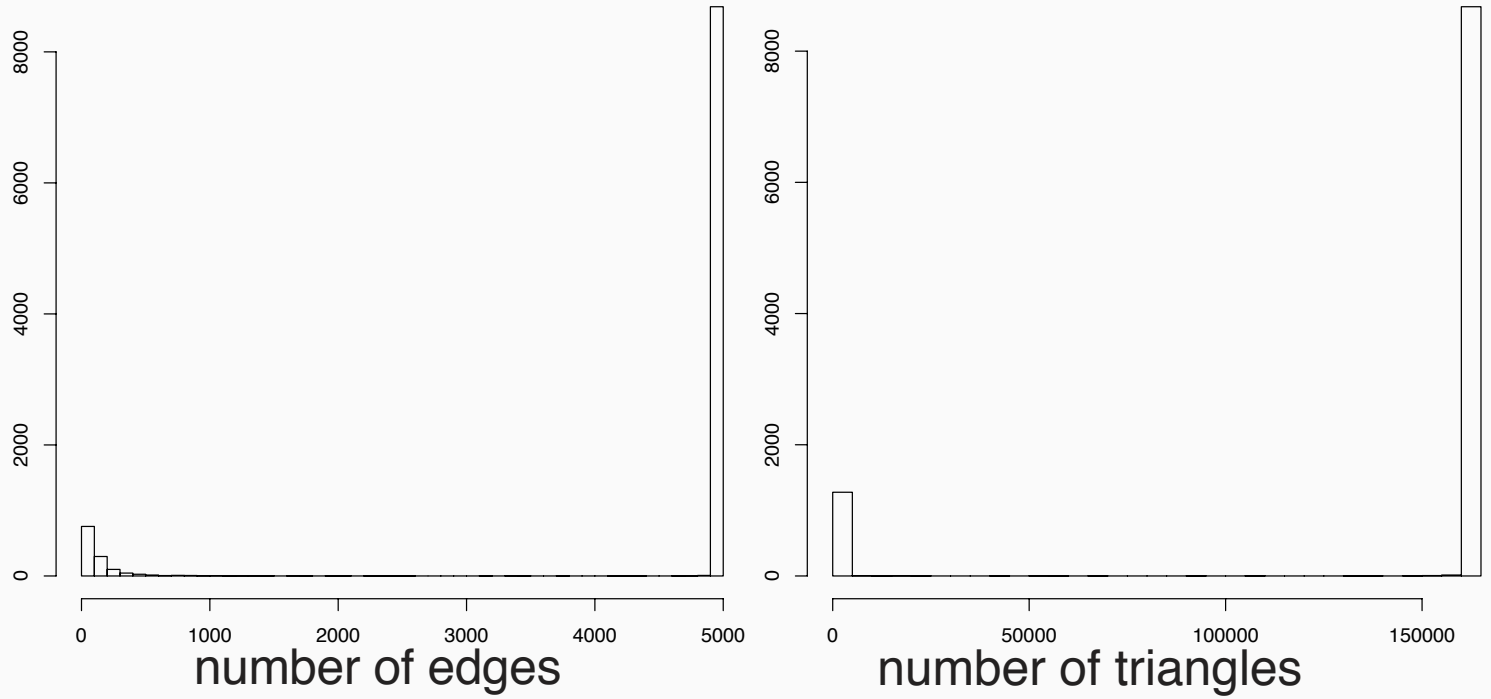


Degeneracy: The triangle model

$$P(Y = y) = \int p(\eta)\, P_\eta(Y = y) d\eta$$

where $p(\eta)$ denotes a distribution over η.

Prior predictions of the statistics under the triangle model
N=4,950 edge variables. Note the extreme polarisation.



Modeling Dependence: Challenges

- ERGMs similar to models in physics, spatial statistics, time-series
- lack of a natural neighborhood structure to bound dependence
- important to exploit X to explain variation
- Schweinberger and Handcock (2015): important to use hierarchical specification to "localize dependence"

Tapered ERGM

Consider the simple modification adding *variation constraints*:

$$\underset{q}{\text{maximize}} \quad \sum_y q(y) \log(q(y))$$

subject to

$$E_q(g_i(Y)) = \mu_i, \quad E_q((\mu_i - g_i(Y))^2) \leq \sigma_i^2, \quad \forall\ i \in \{1, \ldots, d\},$$

Leads to:

$$q(y|\theta, \tau) = \frac{1}{Z(\theta, \tau)} e^{\sum_k \theta_k g_k(y) - \sum_k \tau_k (\mu_k(\theta,\tau) - g_k(y))^2}, \tag{1}$$

where $\mu(\theta, \tau) = E_q(g(Y))$

Tapered ERGM

A "new" family of exponential-family models

$$q(y|\theta,\tau) = \frac{1}{Z(\theta,\tau)} e^{\sum_k \theta_k g_k(y) - \sum_k \tau_k (\mu_k(\theta,\tau) - g_k(y))^2}. \quad (2)$$

where $\tau > 0$ are vectors of hyper parameters

$$\text{mean value parameters :} \quad \mu(\theta,\tau) = E_q(g(Y))$$

$$\text{tapering parameters :} \quad \sigma^2 = \mathbf{V}(\theta,\tau) \equiv \mathbf{V}_q(g(Y))$$

the augmented term *tapers* the likelihood of configurations far from the mean μ.

τ_k interpreted as the strength of the attractive force to the mode
τ is determined by σ^2 via

$$\mathbf{V}(\theta,\tau) = \mathbf{V}_q(g(Y)) = \sigma^2 = (\sigma_1^2, \ldots, \sigma_d^2)$$

Degeneracy of Tapered ERGM

When near-degeneracy occurs, the ERGM $q(y|\theta)$ is plagued by multimodality in $g(Y)$.

One way to ensure $q(y|\theta)$ is unimodal is to require it does not have any local minima or saddle points for any θ.

Using results of Horvat, Czabarka, and Toroczkai (2015), Blackburn and Handcock (2022) show:

Theorem

Let chull(T) be the convex hull of the sample space of statistics, T. For any vector μ of mean parameters in chull(T), there exists a vector of tapering parameters $\tau \in \mathbb{R}^d_{\geq 0}$ such that the Tapered ERGM with tapering center μ is non-degenerate.

Degeneracy of Tapered ERGM

Let $s \in \{g(y,x) : y \in \mathcal{Y}, x \in \mathcal{X}\}$ be a possible network statistic.
Let $N(s)$ be the number of networks with statistic s.
Let $N^*(s)$ be a smoothed twice differentiable approximation of $N(s)$.
Let $q^*(s) = \frac{N^*(s)}{N(s)} q(s)$ be a smoothed version of the tapered ERGM PMF.

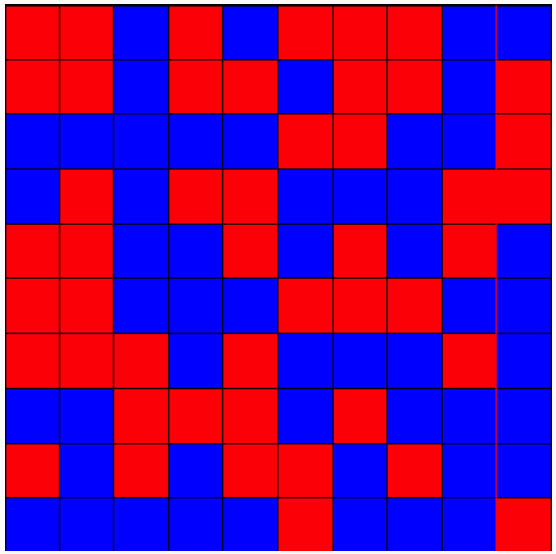
Theorem

There exists $\alpha > 0$ such that for all $\alpha > \tau > 0$, the smooth $q^*(s)$ has no local minima, nor saddle points for all θ.

It is *uniphase* in the sense of Horvat, Czabarka, and Toroczkai (2015).

Illustration: Ising Model for Political Polarization

Consider a 10×10 grid of cells linked via contiguity. We consider the links as fixed and cell values, y_{ij}, as random {red, blue}



$$q(y|\theta,\tau) = \frac{1}{Z(\theta,\tau)} e^{\sum_k \theta_k g_k(y) - \sum_k \tau_k (\mu_k(\theta,\tau) - g_k(y))^2}. \quad (3)$$

Choose

$$g_1(y) = \sum_{i,j} y_{ij} \qquad g_1(y) = \sum_{i,j} y_{ij}(y_{(i+1)j} + y_{i(j+1)}) \qquad \text{“shared spin”}$$

Illustration: Ising Model for Political Polarization

Simulations from traditional Ising model ($\tau = 0$)

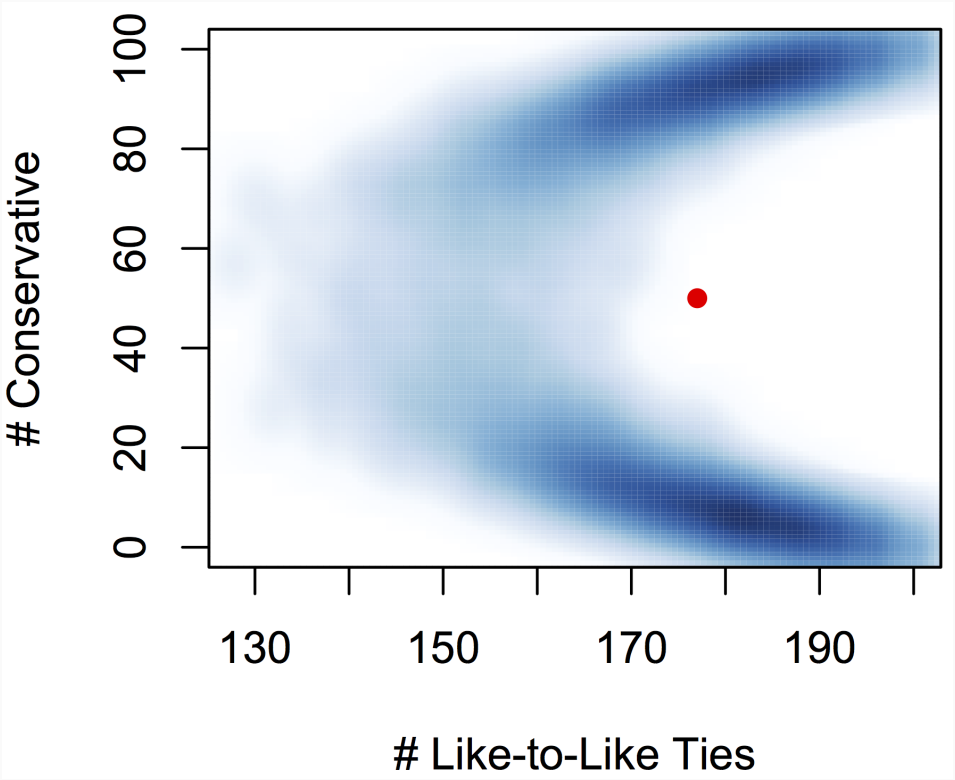


Figure 1: Simulations from the Ising model at the MLE with $\hat{\theta}_{\text{mle}} = (0, 0.45)$. The observed configuration is marked as a red point. Note that the configurations similar to the observed are extremely uncommon under the maximum likelihood model.

The red point is a society where 177 of 200 connections are with like.

Illustration: Ising Model for Political Polarization

Simulations from tapered Ising model ($\tau = 0$ to 1.41)

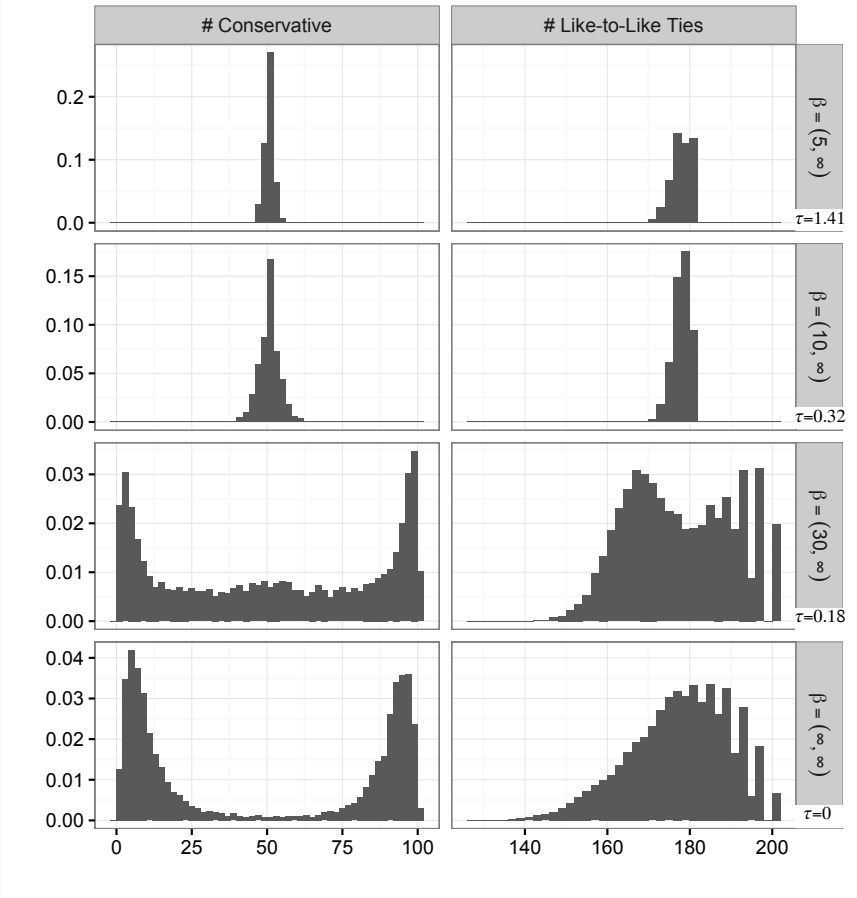


Figure 2: Histogram density estimates based on 10,000 simulations at each β value from the MLE fit to the configuration with 50 conservatives and 177 like affiliation ties. Smaller values of β remove the bimodality of the phase transition.

Inference for Tapered ERGM

The first derivative of the log likelihood is

$$\frac{\delta \ell}{\delta \theta_i} = (g_i(x) - \mu_i(\theta, \tau)) - 2 \sum_k \tau_k \frac{\delta \mu_k(\theta, \tau)}{\delta \theta_i} (\mu_k(\theta, \tau) - g_k(x))$$

The second derivative of the log likelihood at the MLE is

$$\left. \frac{\delta \ell}{\delta \theta_i \delta \theta_j} \right|_{\hat{\theta}_{\text{mle}}} = -\frac{\delta \mu_i(\theta, \tau)}{\delta \theta_j} - 2 \sum_k \tau_k \frac{\delta \mu_k(\theta, \tau)}{\delta \theta_i} \frac{\delta \mu_k(\theta, \tau)}{\delta \theta_j}. \tag{4}$$

where

$$\frac{\delta \mu(\theta, \tau)}{\delta \theta_i} = (I - B)^{-1} c^i,$$

$$B_{rk} = 2 \beta_k^{-2} \text{cov}(g_r(X), g_k(X))$$

and c^i be a vector with elements

$$c_r^i = \text{cov}(g_r(X), g_i(X)),$$

Inference for Tapered ERGM

So finding the MLE of θ reduces to finding the MLE with $\mu(\theta, \tau) = g(x_{\text{obs}})$ and can be computed as simply as a standard ERGM (via MCMC or otherwise).

Nominal standard errors and likelihood ratios can be computed using the above formulas

Add Health Network

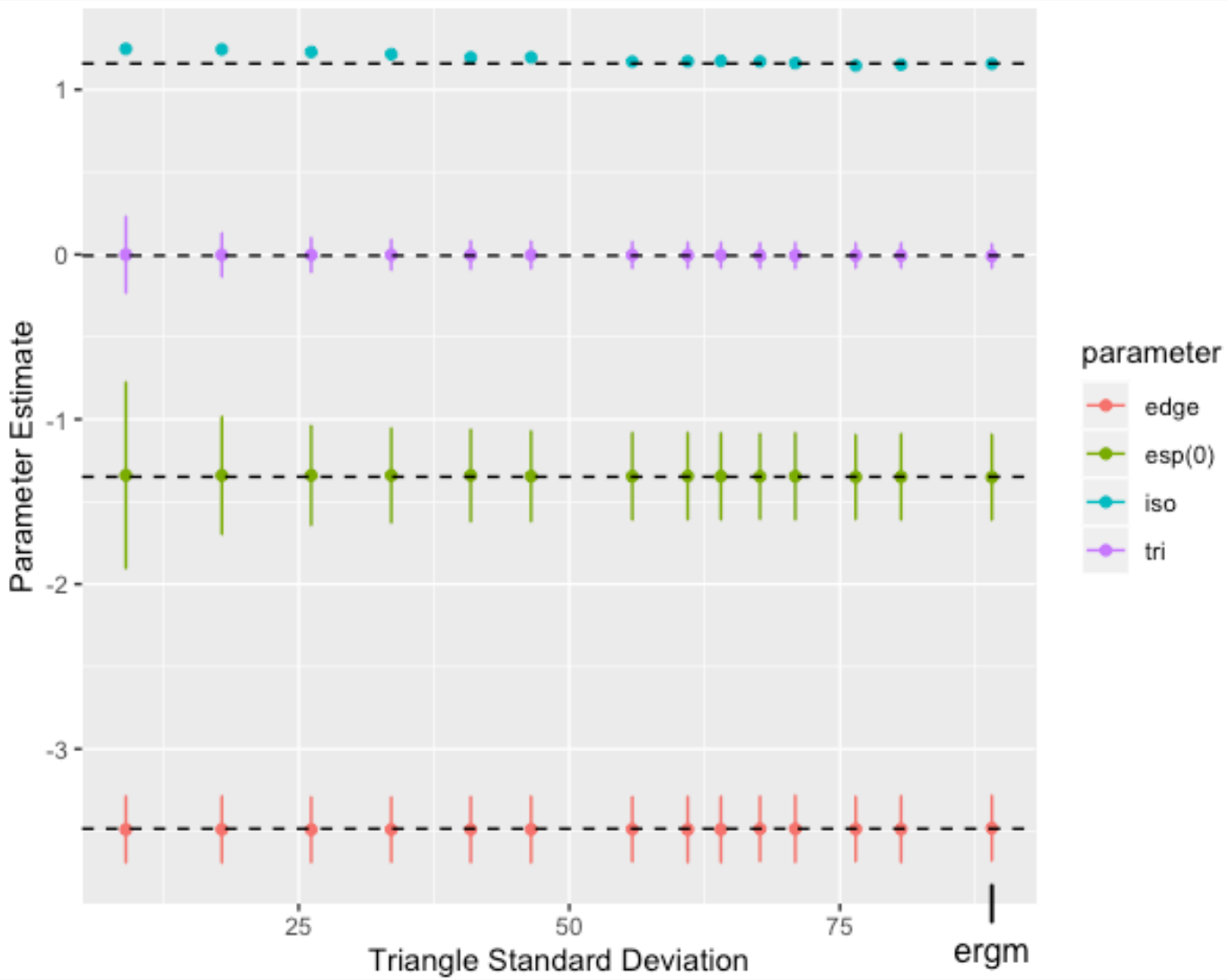
- "Add Health" is a school-based study of the health-related behaviors of adolescents in grades 7 to 12.
- Each nominated up to 5 boys and 5 girls as their friends
- Synthetic data (Faux Desert High)
- $n = 107$ students across six grades
- Covariates: grade (7 through 12) and race
- There are 439 directed friendship edges, 677 triangles

Comparing an ERGM to a Tapered ERGM

Table 1: ERGM fit vs Tapered ERGM fit on Faux Desert High Network. In the Tapered ERGM, tapering was done on the dyad-dependent terms.

Term	ERGM	SE	Tapered ERGM	
# edges	-3.48	(0.10)	-3.49	(0.10)
# triangles	-0.008	(0.038)	-0.002	(0.054)
# isolates	1.16	(0.47)	1.20	(0.63)
# no shared	-1.35	(0.13)	-1.35	(0.15)
# homophily on 7	2.22	(0.23)	2.19	(0.24)
# homophily on 8	2.07	(0.17)	2.05	(0.17)
# homophily on 9	1.99	(0.16)	1.98	(0.16)
# homophily on 10	1.57	(0.11)	1.57	(0.11)
# homophily on 11	1.78	(0.15)	1.77	(0.15)
# homophily on 12	1.28	(0.28)	1.28	(0.28)

How much does tapering change the parameter estimates?



We see that regardless of how much tapering we apply, the parameter estimates and standard errors are similar to standard ERGM.

How do we interpret the parameters of the tapered ERGM?

If the tapering parameters τ are zero, then the Tapered model is identical to the standard ERGM and an interpretation of the θ parameters is as conditional log-odds. However, non-zero τ has an effect on the interpretation of the parameters. To see this, Let $P(Y_{ij} = 1 | Y_{ij}^c = y_{ij}^c) \equiv P(Y_{ij}^+)$ and $P(Y_{ij} = 0 | Y_{ij}^c = y_{ij}^c) \equiv P(Y_{ij}^-)$.

Under the Tapered ERGM the log-odds of a tie conditional on Y_{ij}^c is

$$\log\left(\frac{P(Y_{ij}^+)}{P(Y_{ij}^-)}\right) = \sum \Delta t_k(Y_{ij}) \left[\theta_k + \tau_k \delta_{kij}\right]$$

where

- $\Delta t_k(Y_{ij}) = t_k(Y_{ij}^+) - t_k(Y_{ij}^-)$ is the change statistic
- $\delta_{kij} = (\mu_k - t_k(Y_{ij}^+)) + (\mu_k - t_k(Y_{ij}^-))$ is the sum of the differences from the mean.
- δ_{kij} is a measure of the deviation of the network statistics from their mean.

How do we interpret the parameters of the tapered ERGM?

When θ_k is the MLE, the log-odds of a tie conditional on Y_{ij}^c is

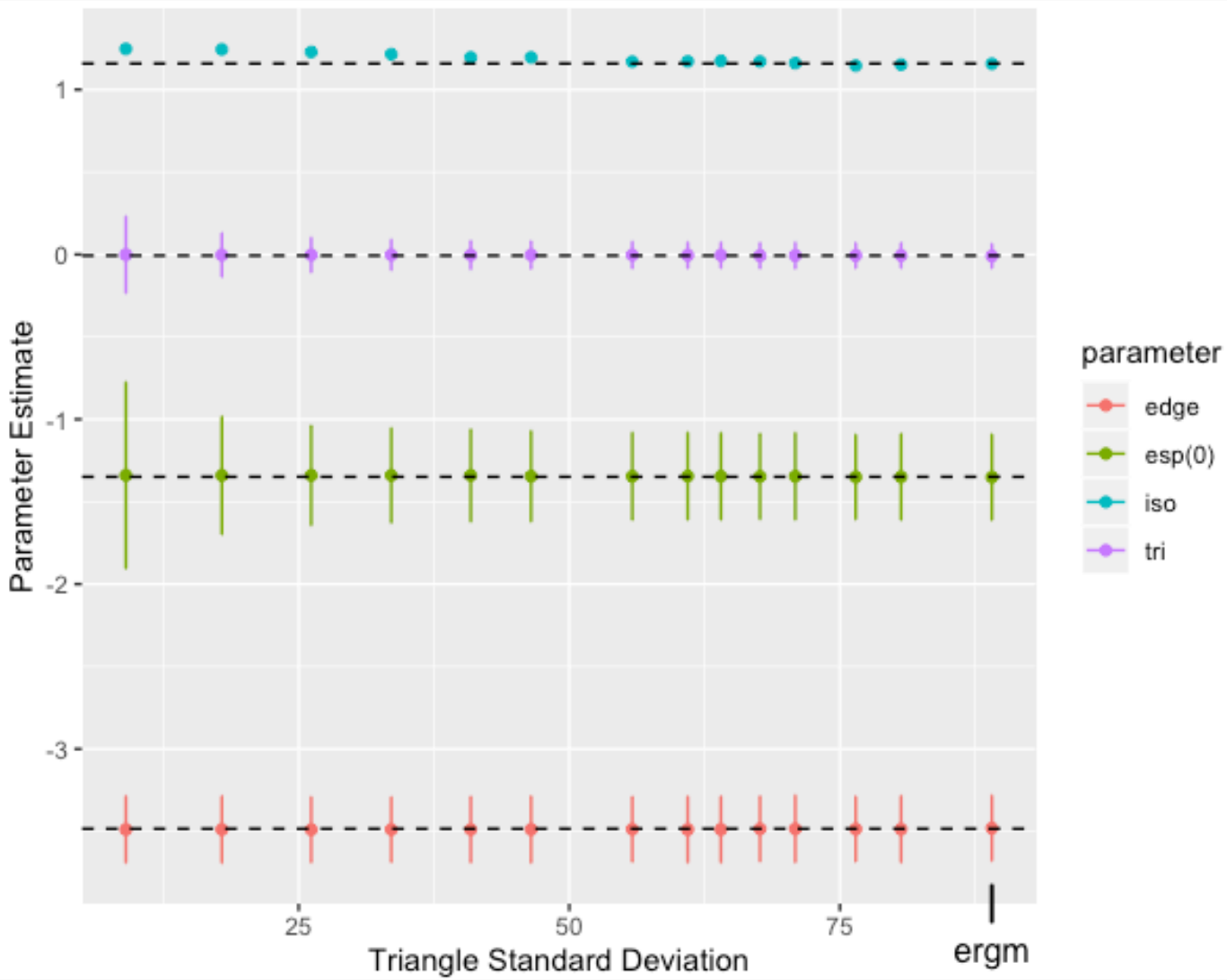
$$\sum_k \Delta t_k(Y_{ij}) \left[\hat{\theta}_k + \tau_k(2Y_{ij} - 1)\Delta t_k(Y_{ij})\right]$$

.

The last expression suggests a measure of the bias in the Tapered ERGM parameter estimate $\hat{\theta}_k$, as an estimate of the conditional log-odds, is the average over the dyads (i, j) in the network of the penalty term:

$$-\tau_k \sum_{ij} (2Y_{ij} - 1)\Delta t_k(Y_{ij})$$

How much does tapering change the parameter estimates?



We see that regardless of how much tapering we apply, the parameter estimates and standard errors are similar to standard ERGM.

How to choose the tapering?

- Fellows and Handcock suggest $\tau = \frac{1}{r^2 \mu}$ where r is a user specified multiplier so that observations r standard deviations from the mean are tapered most.
- In particular, a reasonable default assumes Poisson-like variation in $g(Y)$ so that $\tau = \frac{1}{r^2 g(y_{\text{observed}})}$.
- This usually leads to light tapering of graphs unexpectedly far away from the mean

How, really, should we choose the tapering?

- One of the hallmarks of near-degeneracy is bi/multimodality.
- How can we measure the bimodality of a distribution?
  Let Z be the standardized version of $g(Y)$ then the kurtosis is:

$$\text{Kurt}[g(Y)] \equiv \text{E}\left[Z^4\right] = \frac{\mu_4}{\mu_2^2}$$

- $\text{Kurt}[g(Y)] \geq 1$
- Gaussian: $\text{Kurt}[g(Y)] = 3$; Uniform: $\text{Kurt}[g(Y)] = 9/5$;
  Poisson: $\text{Kurt}[g(Y)] = 3 + \frac{1}{\mu}$
- Blackburn and Handcock (2022) argue that we can interpret kurtosis as a measure of bimodality in the context of network modeling

Penalized Likelihood via the Kurtosis

If we set a target kurtosis value we can simply maximize the log-likelihood subject to a penalty on how far the kurtosis deviates from the target value plus a penalty on the magnitude of τ:

$$\hat{\tau} = \arg\max_{\tau} \left[l(\theta, \tau, ; y_{\text{observed}}) - \tau - \gamma \text{ penalty on } K[g(Y)|\theta, \tau]\right]$$

where K_T is a target kurtosis and K_σ and γ are scale parameters.

Penalized Likelihood via the Kurtosis

If we set a target kurtosis value we can simply maximize the log-likelihood subject to a penalty on how far the kurtosis deviates from the target value plus a penalty on the magnitude of τ:

$$\hat{\tau} = \arg\max_{\tau} \left[l(\theta, \tau, ; y_{\text{observed}}) - \tau - \gamma \left(\frac{K[g(Y)|\theta, \tau] - K_T}{K_\sigma} \right)^2 \right]$$

where K_T is a target kurtosis and K_σ and γ are scale parameters.

Sensible default values are

- $K_T = 3$ (Gaussian)
- $K_\sigma = 0.6$, half the distance from 3 to 1.8 (Uniform).
- $\gamma = \frac{1}{2}$

To simplify, take the average penalty over all tapered terms and reexpress as $\hat{r} = \frac{1}{\sqrt{\hat{\tau}\mu}}$.

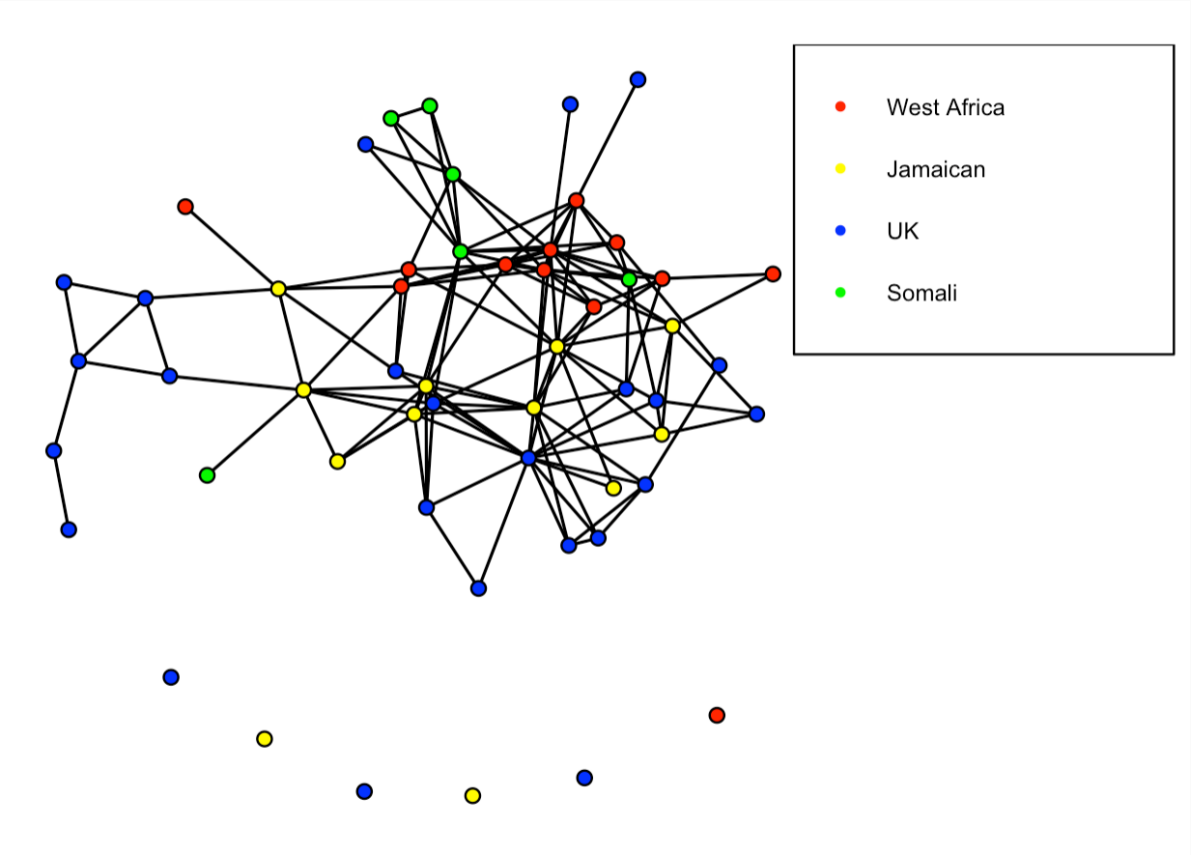
Example: Add Health Network

The optimal $\hat{r} = 2.48$, that is taper at over two standard deviations - very weak.

Example: Ethnic heterogeneity of a London street gang

- The members of a London gang between 2006 and 2009.
- A tie exists between two gang members if they were arrested together for committing a crime at least once.
- undirected network with 54 vertices and 133 ties
- Studied by Grund and Densley (2015)

Co-offending network of a London street gang



A tie exists between two gang members if they have committed at least one crime together. All gang members are Black but the gang is comprised of four distinct ethnicities, categorized by the authors as their countries of origin.

Modeling Ethnic heterogeneity of a London street gang

Grund and Densley (2015) posit that who co-offends with whom is driven by ethnic homophily and homphilous triad-closure.

- But the triangle term is near degenerate in standard ERGM.
- So they use homophilous GWESP terms and conclude homophilous triangle closure
- Tapered ERGM can fit the homophilous triangles directly

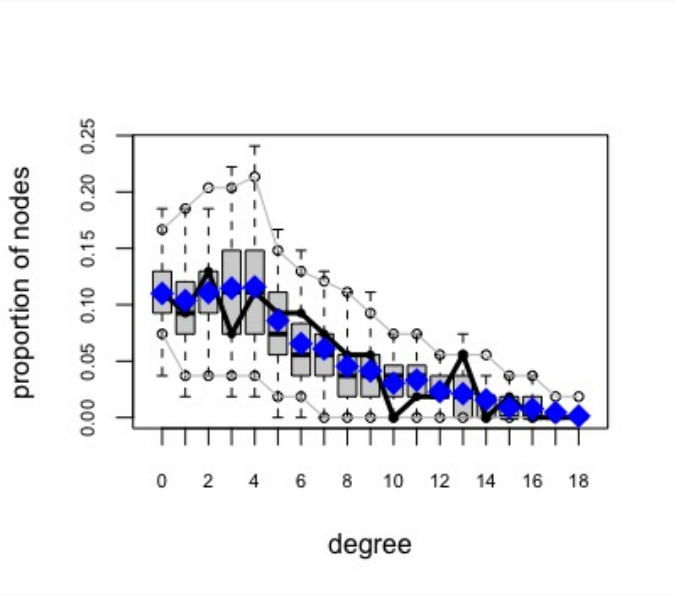
Tapered ERGMs fit on London Gang Network

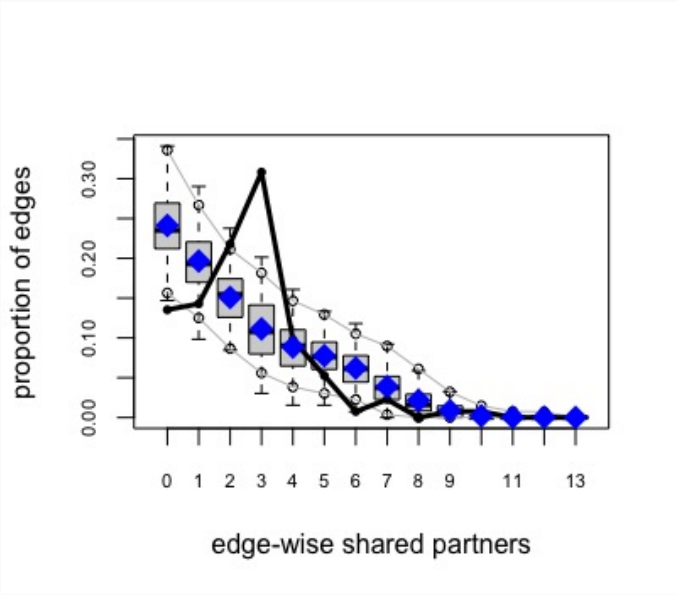
Term	Model 1	Model 2	τ	bias
edges	-3.23 (0.18)***	-3.34 (0.17)***	0.001	-0.0001
triangles	0.68 (0.10)***	0.71 (0.09)***	0.001	-0.0012
triangles(West Africa)	0.11 (0.38)	0.12 (0.37)	0.011	-0.0023
triangles(Jamaican)	0.17 (0.61)	0.41 (0.54)	0.027	0.0000
triangles(UK)	0.56 (0.38)	0.61 (0.42)	0.021	-0.0015
match(West Africa)	0.96 (0.60)	0.95 (0.56)	0.008	-0.0005
match(Jamaican)	1.35 (0.66)*	0.94 (0.55)	0.012	0.0006
match(UK)	0.27 (0.40)	0.31 (0.42)	0.007	-0.0004
match(Somali)	2.17 (0.59)***	2.33 (0.50)***	0.027	0.0004
isolates		0.98 (0.67)	0.027	-0.0027

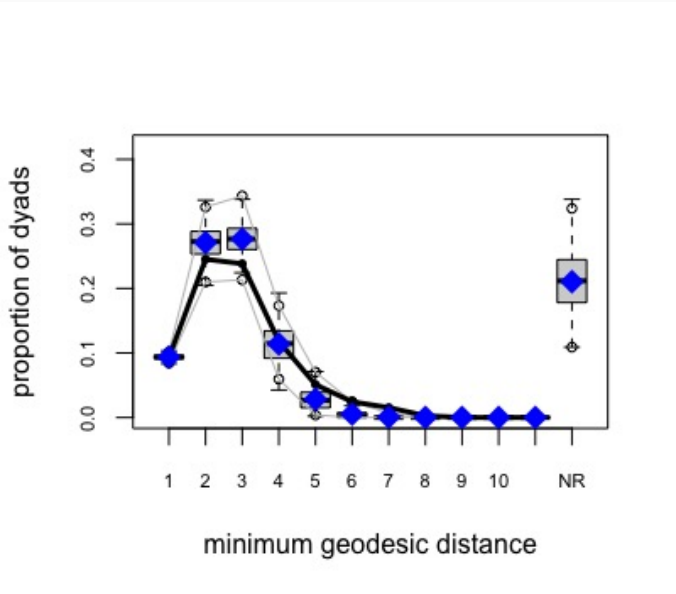
$^*p < .05$ $^{**}p < .01$ $^{***}p < .001$

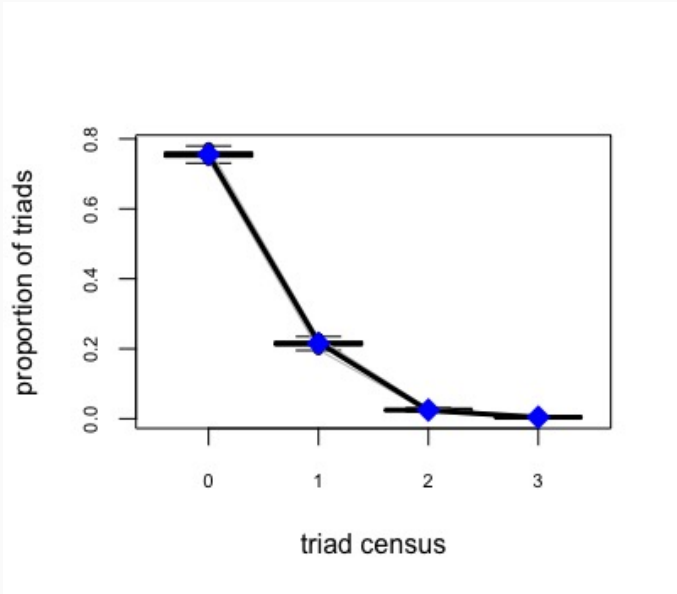
The conclusion is that overall triad closure is the main factor, not homophilous closure.

Goodness-of-fit diagnostic plots for Model 2









Conclusions on Tapering models

- Practical modeling via ERGMs has been hindered by concerns about near-degeneracy.
- Near-degeneracy constrains the space of ERGMs in that many intuitive or theory-driven terms, like the triangle, most often cannot be used
- The Tapered ERGM can incorporate any term with a guarantee of non-degeneracy.
- Frees modeler to choose most scientifically interpretable statistics
- Has theoretical guarantees of stability
- Has a simple and appealing interpretation (constrained max entropy)

Conclusions on Tapering models

- We have the developed a procedure to estimate the tapering needed for a non-degenerate model
- Parameter estimates are close to ERGM, when the later exist.
- The procedure usually chooses a standard ERGM when justified.
- Computationally stable: Can be used as a computational device
- Open-source, user-friendly software is available on GitHub in the ergm.tapered package