Exponential random graph modeling of a gonorrhea outbreak in an indigenous First Nations community in Alberta, Canada

J Steven Raquel

12/07/2021

Exponential random graph modeling of a gonorrhea outbreak in an indigenous First Nations community in Alberta, Canada

Background

- This dataset concerns a localised outbreak of *Neisseria gonorrehoeae* (gonorrhea) in an indigenous First Nations community located in Alberta, Canada.
- It was originally analyzed in a paper by Prithwish De, et al. (2004), in which they used measures of network centrality (e.g. information centrality) to determine the association between the risk of infection between members of the network and their position within the network itself.
- The data for the 2004 paper was sourced from an earlier 2001 study by the authors in which they formulated a plan to address the outbreak.
- The network consists of 89 individuals, both male and female, 17 of whom were found to be patrons of a local bar in the area.

Background

- This work expands upon the original 2004 analysis by applying exponential random graph modeling (ERGM) to the network in order to quantify the effect of various attributes of the network (e.g. gender, bar attendance), as well as looking at other less common measures of network centrality to quantify the effect of an individuals' position in the network and their connectedness to others, on the outbreak.

Neisseria gonorrheoeae

- Gonorrhea is a sexually transmitted disease/infection (STD/STI) which can be transmitted orally, vaginally or anally.
- Although it can have many serious side effects, it can also be symptomless, leading to individuals unknowingly infecting their partners.
- When untreated, it also makes HIV more susceptible to transmission, making gonorrhea itself a risk factor for the propagation of HIV.
- To this day, gonorrhea is the second most commonly reported STD in Canada.

First Nations

- As of 2016, approximately 7% of Albertans identify as First Nations, one of the aboriginal groups native to Canada. This is compared to approximately 5% throughout all of Canada.
- First Nations peoples experience a disproportionate prevalence of STDs in their population relative to other groups in Canada, due at least in part to cultural differences and lack of access to resources such as those in more urbanized areas and more populated by non-indigenous people.
- As such, proper intervention design on this population will require knowledge of an sensitivity to their cultural norms.

Data Collection

- The original study followed an *egocentric* social network design in which individuals provided information on themselves and others within their social networks (alters).
- Sexual networks can be more difficult to collect data on sociometrically since it would require having access to the entire sexual network, which is unrealistic.

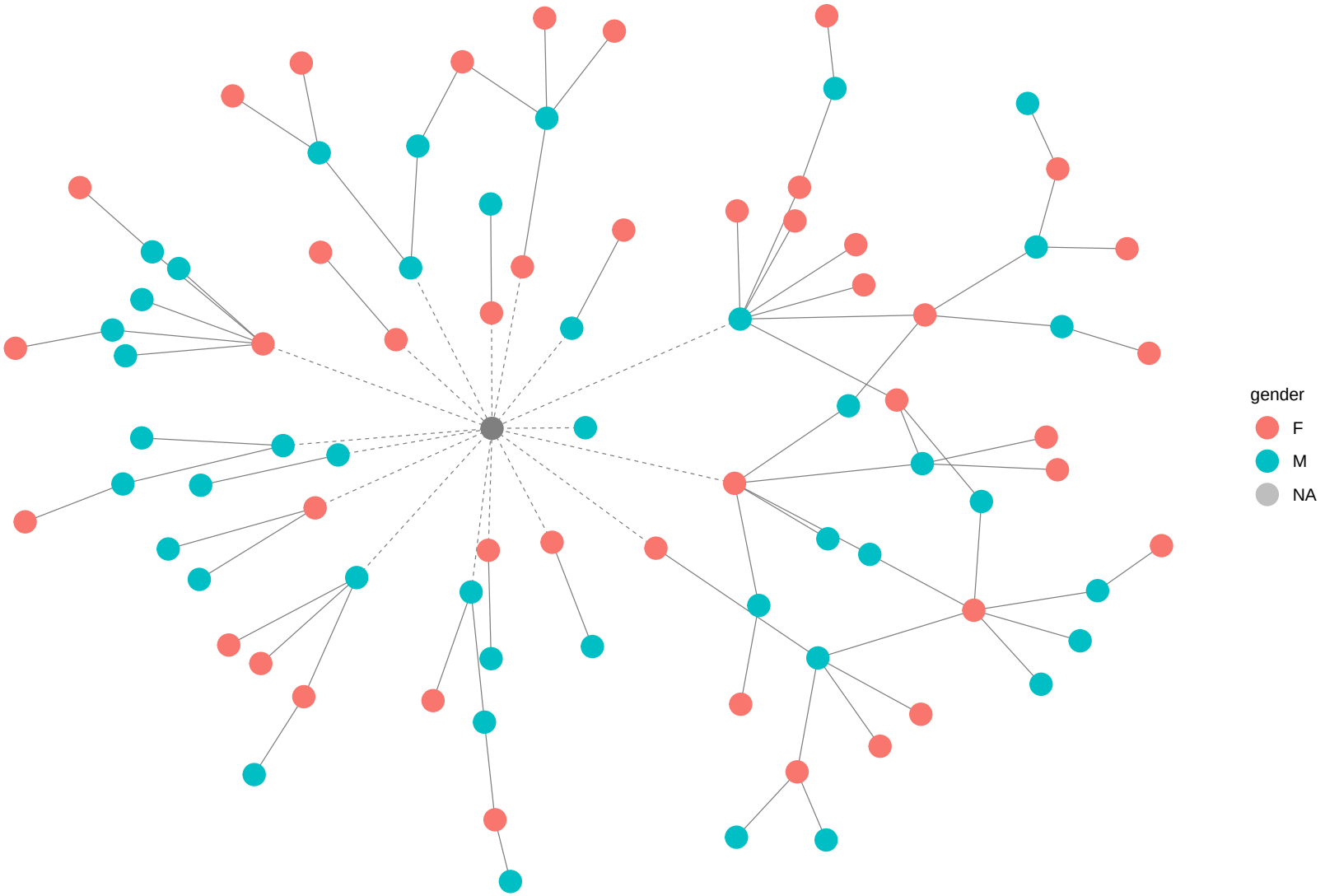
Data Collection

- Since STD clinical reports don't contain *complete* sexual contact information, the original index case could not be identified.
- Questions about sexual risk factors and drug use were omitted from the survey after a high proportion of respondents opted not to answer it.
- It was found in 2001 that certain individuals acquired their respective infections from partners whom they met outside of their own local community.

Data Collection

- As is typical for data collection processes in which information is nominated by the ego, and especially in the sensitive case where sexual partnering is involved, there is bound to be some information missing from the data.
- e.g. Self-reported behavior from an ego may not necessarily reflect their practices in reality.

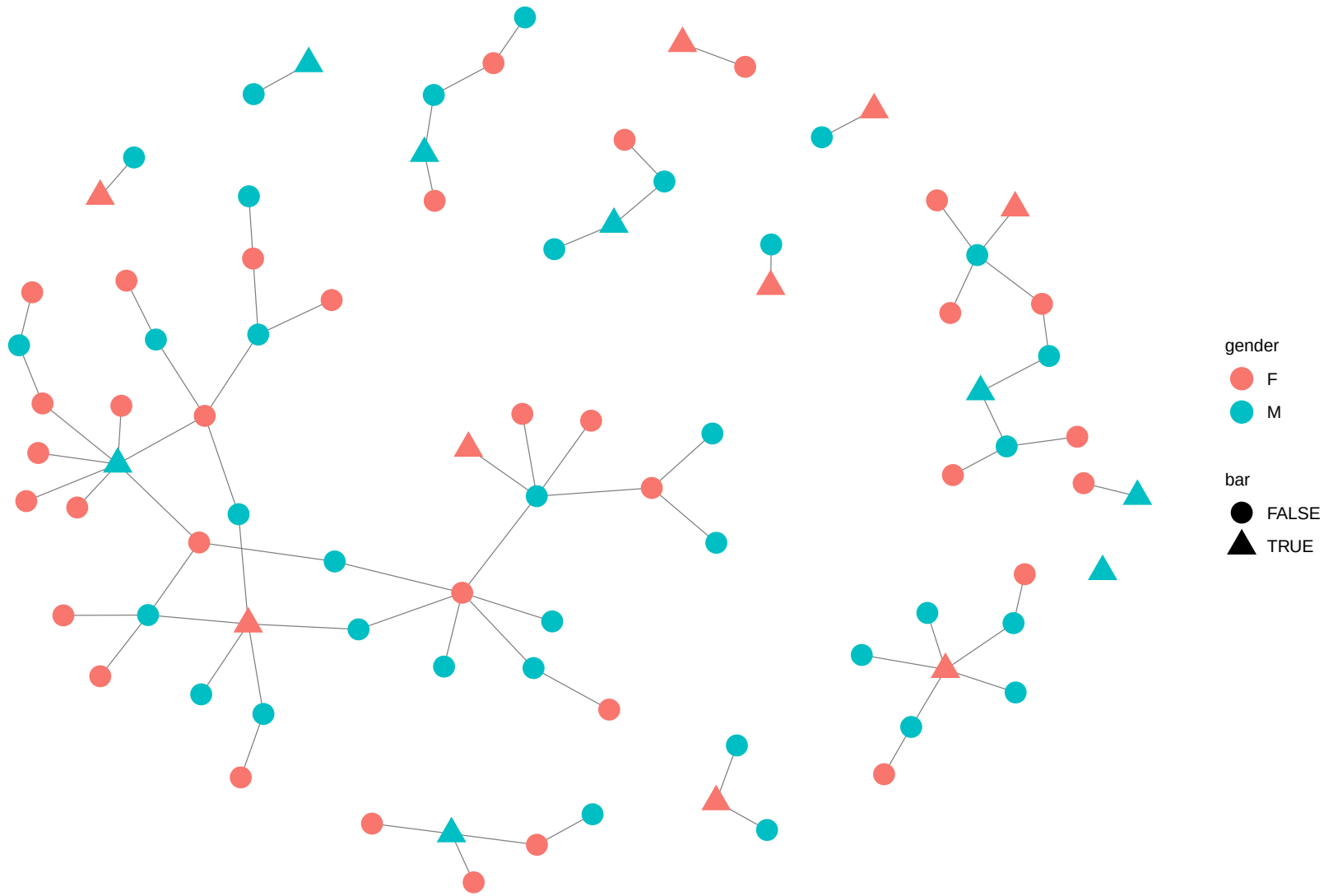
The connected network with the bar as a node



The connected network with the bar as a node

- There are 91 total edges in this network, with 17 of them being between the bar and individuals, leaving 74 of them to be between individuals.
- Of these 74, 67 (approx. 90.5%) of them are M-F edges, 6 of them are M-M, and 1 of them is F-F.
- Of these 74, 36 (approx. 48.6%) of them are between bar-attendees and non-attendees, and 38 (51.3%) are between non-attendees and other non-attendees only.
- Note that none of the bar attendees have ties to other bar attendees at all.

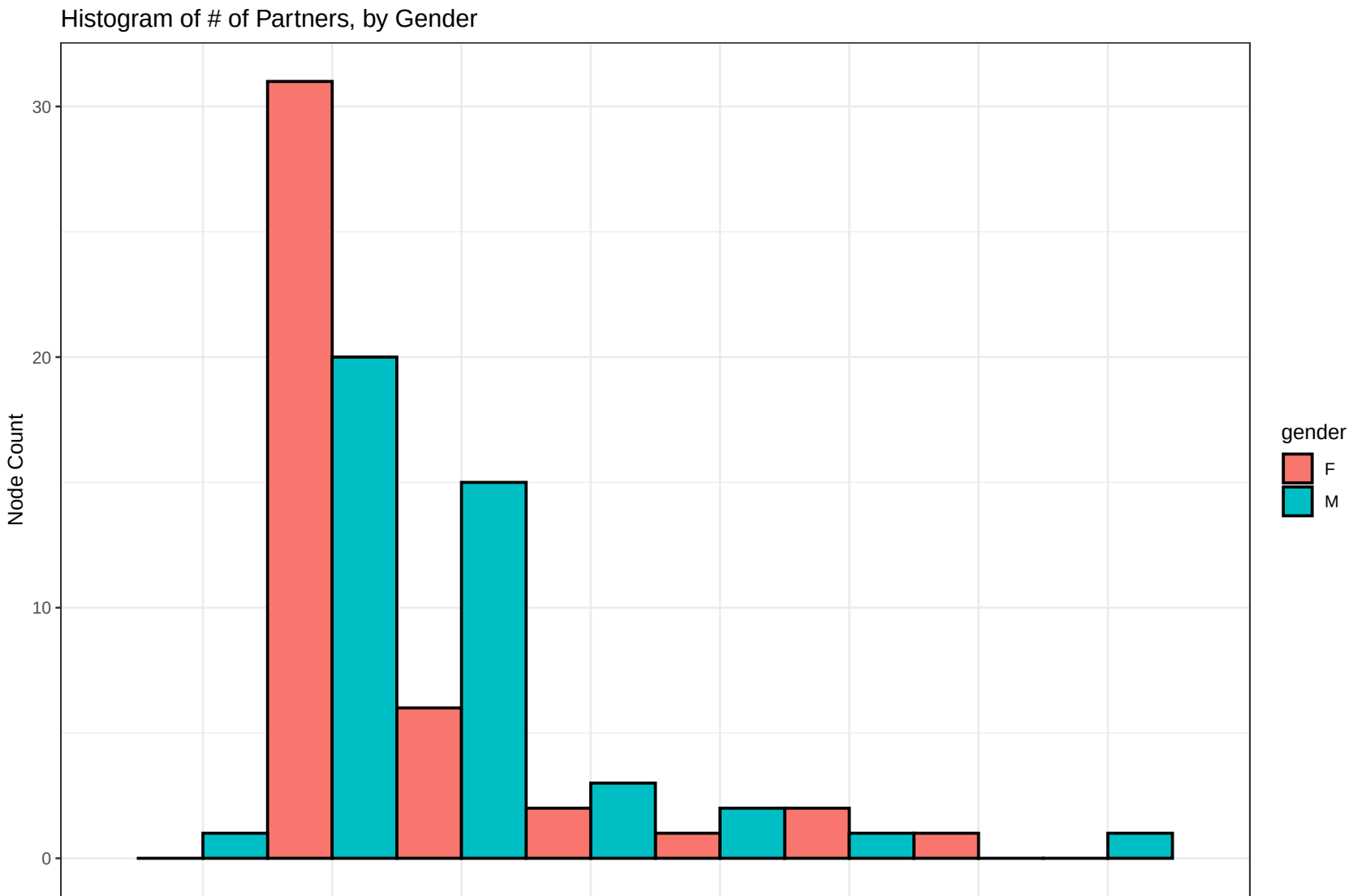
The disconnected network



The disconnected network

- There are about 9 or so subcomponents of this graph; notice that the largest among them has 3 bar attendees within it, 2 of whom have at least 5 ties to other individuals.
- Notice there is one isolate where a male node attended the bar but otherwise has no ties to any other node.
- There also a a series of smaller sub-components where there is only one edge between two nodes, one of whom is a bar attendee.
- Note the presence of a "6-cycle" in the largest subcomponent, in which 6 separate nodes are jointly connected through one another. This is the only such appearance of a k-cycle in the entire network.
- The largest subcomponent of the graph includes 38 of the 86 nodes in the network, or approximately 44% of all egos.

Distribution of ties



Centrality

- De et al. (2004) looked at convention measures of centrality e.g. degree centrality, information centrality, while analyzing the connected network that incorporated the bar as a node.
- While we will touch on these measures of centrality, we also wanted to expand upon them by looking at the *disconnected* network in which this bar node is absent, while retaining the information about bar attendance via a vertex attribute.

Imputing finite distance values between disconnected nodes

- One issue when computing centrality measures on a disconnected network is that we do not have a measure for the distance between two nodes who are completely disconnected from each other (i.e. there are no edges between them that link them together).
- As a finite approximation, we used $D + 1$, in which D is the size of the longest finite distance between two nodes in the network.

Exponential Random Graph Model (ERGM)

- While measures such as network centrality and network density can provide us some information about the characteristics of a network, in the case of a sexual network in which ties have an explicitly social component to them, an ERGM which incorporates the attributes of actors within the network can be more informative.
- ERGMs model networks as a function of network statistics, by imagining the network as one instance of a set of possible, similar networks, i.e. the outcomes of a stochastic (random) process.
- They are used to predict the probabilities of ties between nodes, similar to logistic regression in the use of log-odds to model probability.

Exponential Random Graph Model (ERGM)

- $\log\left[\frac{\Pr(Y_{ij}=1|y_{ij}^{C})}{\Pr(Y_{ij}=0|y_{ij}^{C})}\right] = \sum A(Y_{ij})\eta_A d_A(y)$
- y_{ij}^{C} = all ties in y, excepting y_{ij}
- $A(Y_{ij})$, all effects in model
- η_A, the parameter for effect A
- $d_A = z(y_{ij}^{+}) - z(y_{ij}^{-})$, difference score for effect z if the tie between i and j were added

Exponential Random Graph Model (ERGM)

- When modeling the ERGM for this network, the bar node was redacted so that the network could be a one-mode network.
- Attendance of the bar was assigned to the vertices as an attribute, as was gender.

Model Selection

- When fitting the ERGM model, first we started with the simplest model that only models the number of edges (this would be akin to an intercept-only model in logistic regression). We then built several more models which only had one parameter, and then compared the Akaike Information Criteria (AIC) of the models to determine which among these model parameters were worth including.
- Among the model parameters fit were homophily (`nodematch`), heterophily (`nodemix`) and degree (`degree`).

Akaike Information Criterion (AIC)

- The Akaike Information Criterion (AIC) is an estimator of out-of-sample prediction error.
- We use it as a means of model selection, where we generally opt to select the model with the lowest AIC.
- $$AIC = -2\ln(\mathcal{L}) + 2k$$
- Where L is the likelihood of the model, and k is the number of parameters.

Model Selection

Table 1: Table of Akaike Information Criteria for each ERGM model

formula	aic
gonnet_net_nobar ~ edges	725.6599
gonnet_net_nobar ~ edges + nodematch("bar")	718.7543
gonnet_net_nobar ~ edges + degree(d = c(1:3))	681.9365
gonnet_net_nobar ~ edges + nodemix("gender")	669.8531
gonnet_net_nobar ~ edges + nodefactor("bar")	725.7581
gonnet_net_nobar ~ edges + nodemix("bar")	902.9734

Likelihood Ratio Test

- A likelihood ratio test is a method to compare a full model to a nested model, where the nested model has some subset of the same parameters as the full model.
- When comparing a full model to a nested model, we can conduct a Likelihood Ratio Test, in which we take the ratio of the maximum likelihoods of both the full model and the nested model, in order to determine whether the nested model explains the data as well as the full model.

Likelihood Ratio Test

- $\lambda = \frac{\mathcal{L}_s(\hat{\theta})}{\mathcal{L}_g(\hat{\theta})}$
- $\text{LRT} = -2 \ln \lambda$
- Where $\mathcal{L}_s(\hat{\theta})$ is the maximized log-likelihood for the nested model, and $\mathcal{L}_g$ is the maximized log-likelihood for the full model.

Likelihood Ratio Test

- H_0: The nested model fits the data as well as the full model. i.e. The nested model is preferred.
- H_A: The nested model does not fit the data as well as the full model. i.e. The full model is preferred.
- We reject the null hypothesis H_0 when the test statistic falls within the rejection region of a χ^2 distribution with degrees of freedom k where k is the difference in the number of parameters between the two models. Otherwise, we fail to reject H_0.

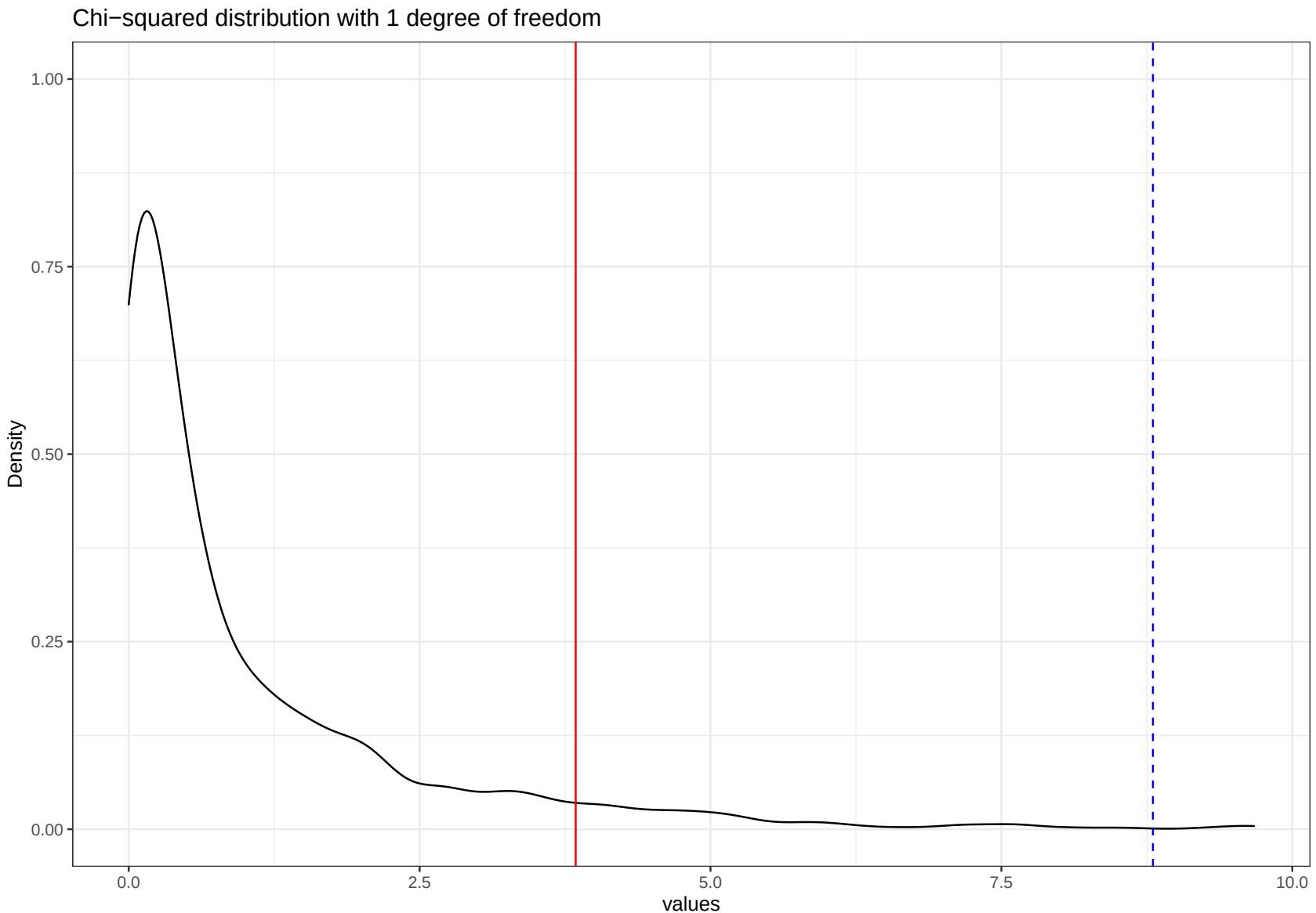
Likelihood Ratio Test

```
## Likelihood ratio test
##
## Model 1: gonnet_net_nobar ~ edges + degree(d = c(1:2)) + nodemix("gender") +
##     nodematch("bar")
## Model 2: gonnet_net_nobar ~ edges + degree(d = c(1:2)) + nodemix("gender")
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -302.24
## 2   5 -306.64 -1 8.8022   0.003009 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Likelihood Ratio Test

- The full model has 1 more parameter than the nested model, so we are comparing our LRT test statistic to a χ^2 distribution with 1 degree of freedom.
- The χ^2 test statistic for the likelihood ratio test had a p-value of ≈ 0.002, which is less than our significance level $\alpha = 0.05$.
- Hence we have enough evidence to reject the null hypothesis and conclude that full model, which has the homophily attribute for bar attendance in addition to the heterophily attribute for gender and the degree attribute, is the preferred model.

Likelihood Ratio Test



Markov Chain Monte Carlo (MCMC)

- A Markov Chain is a stochastic process that goes from one state to another, with the future state X_{t+1} depending only on the current state X_t at time t.
- $\Pr(X_{t+1} = x | X_t = x_t)$
- A Monte Carlo process refers to a simulation that samples many random values from a posterior distribution of interest.
- Hence a Markov Chain Monte Carlo (MCMC) simulation is a simulation of a random process whose future value depends only on the current value.

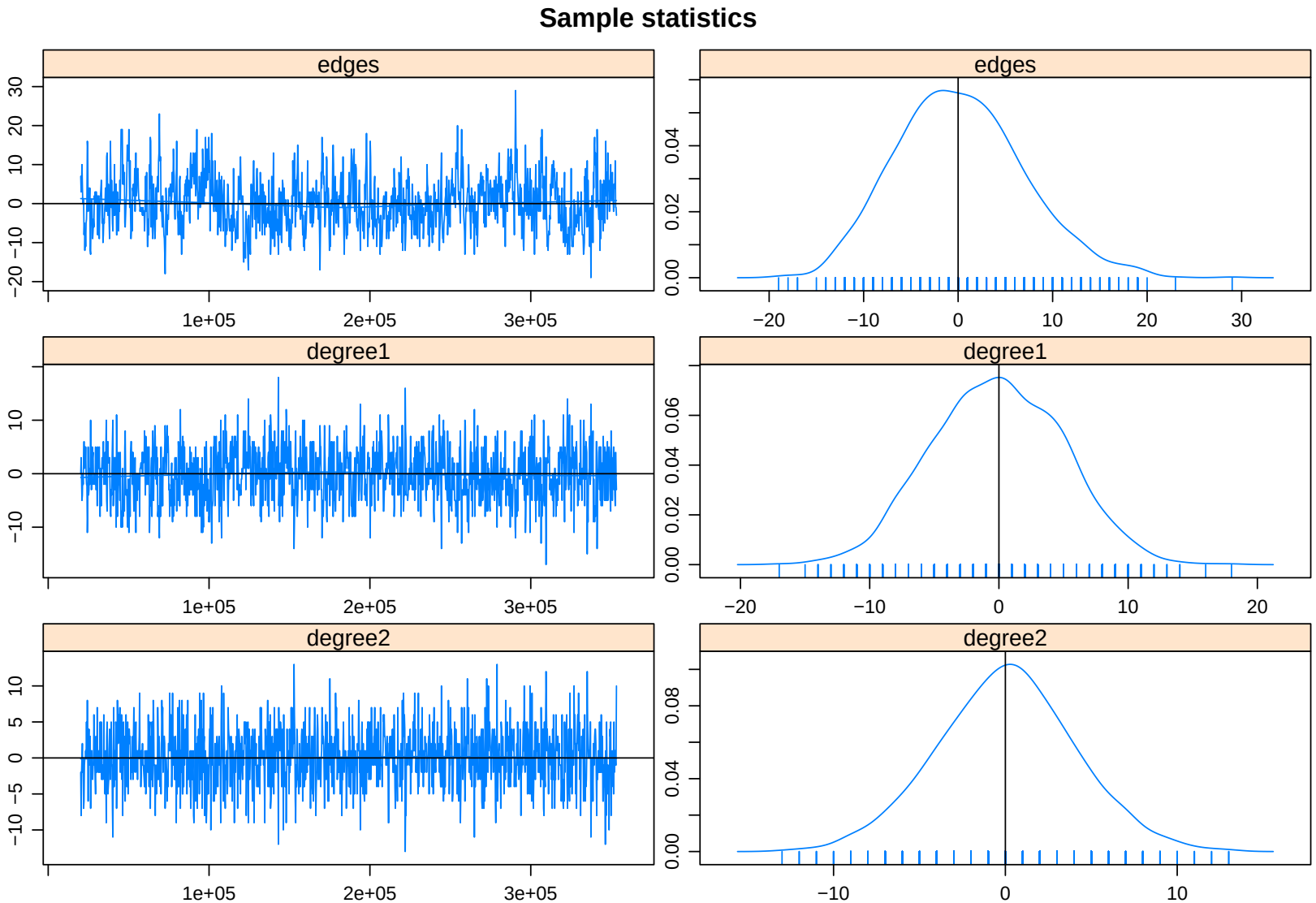
Metropolis-Hastings Algorithm

- A common MCMC algorithm is the Metropolis-Hastings algorithm, in which the process "randomly walks" starting from some θ_0.
- To determine whether the process advances, a "candidate value" θ is generated by sampling from the proposal distribution $g(\theta|\theta_{s-1})$ at iteration s.
- We then derive the Metropolis-Hastings acceptance ratio η, which is a ratio of the estimated distribution at time θ and at the proposed time θ_n.
- Then we derive the acceptance probability ρ by taking the minimum between the acceptance ratio η and 1, i.e. $\rho = \min(1, \eta)$
- We sample $U \sim \text{Uni}(0, 1)$ to and compare it to ρ. If $U < \rho$ then we accept the candidate value such that $\theta_s = \theta$
- Otherwise we repeat another iteration with θ_s as before.

MCMC Diagnostics

- When using MCMC chains our goal is for it to *converge*.
- We can diagnose this by looking at traceplots of the MCMC chain as well as the density curve.
- We want to see the traceplot have constant variance (homoskedascticity) and for the density plot to be centered at 0 and shaped like a bell curve.

MCMC Diagnostics

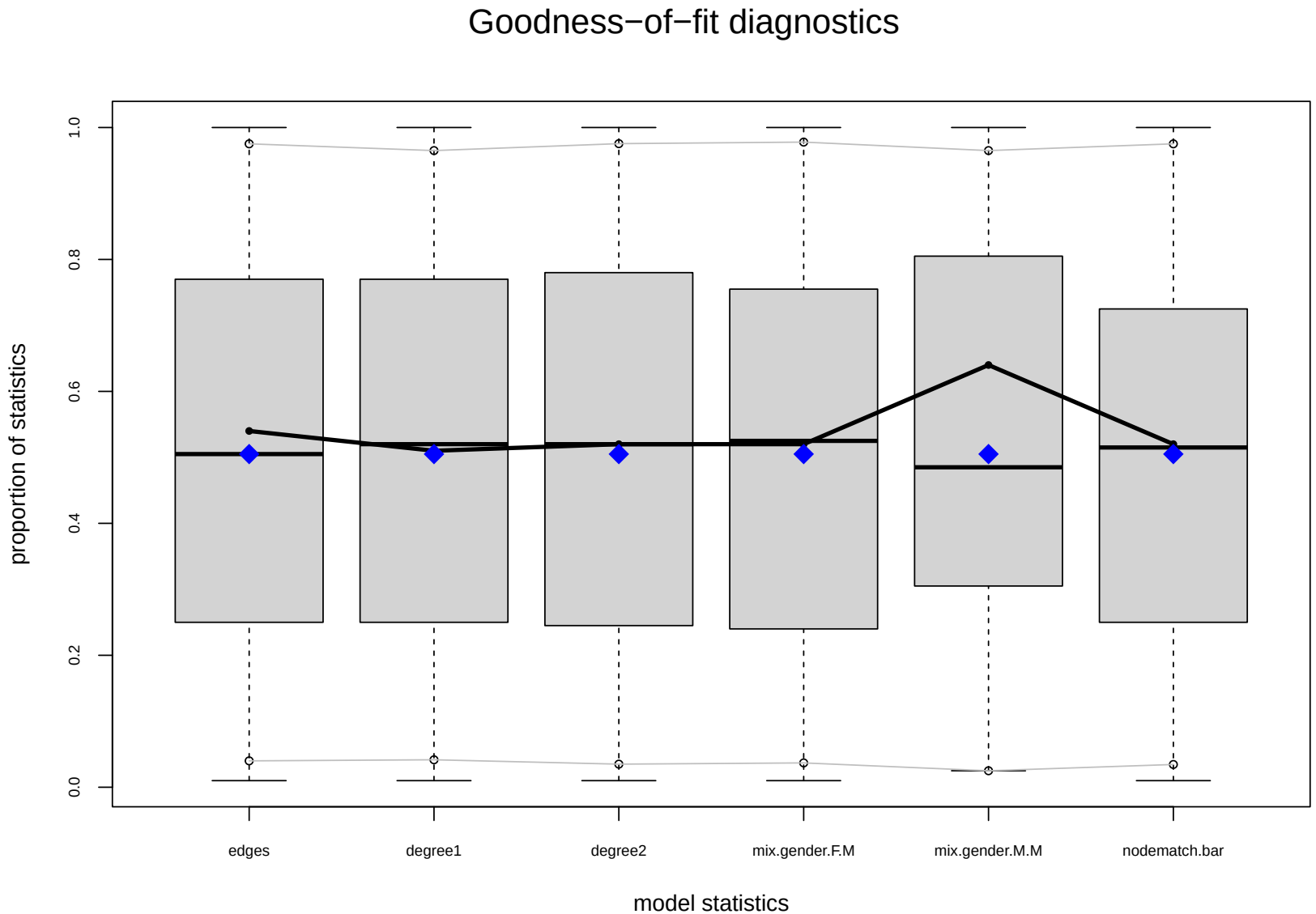


Goodness of Fit

```
## Call:
## .f(formula = ..1)
##
## Monte Carlo Maximum Likelihood Results:
##
##                Estimate Std. Error MCMC % z value Pr(>|z|)
## edges           -5.2893     1.0010      0  -5.284  < 1e-04 ***
## degree1          2.7322     0.5496      0   4.971  < 1e-04 ***
## degree2          1.4581     0.4510      0   3.233 0.001225 **
## mix.gender.F.M   3.5327     0.9546      0   3.701 0.000215 ***
## mix.gender.M.M   1.8113     1.0136      0   1.787 0.073939 .
## nodematch.bar   -0.6958     0.2211      0  -3.146 0.001653 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      Null Deviance: 5066.9  on 3655  degrees of freedom
##  Residual Deviance:  604.5  on 3649  degrees of freedom
##
## AIC: 616.5  BIC: 653.7  (Smaller is better. MC Std. Err. = 0.7181)
```

```
##                obs min  mean max MC p-value
## edges           74  59 74.28  96          1
## degree1         51  37 51.08  63          1
## degree2         21  11 20.59  30          1
## mix.gender.F.M  67  55 67.07  87          1
## mix.gender.M.M   6   1  6.04  15          1
## nodematch.bar   38  24 38.07  53          1
```

Plotting Goodness of Fit



The Final Model

```
##
## Call:
## .f(formula = ..1)
##
## Last MCMC sample of size 1302 based on:
##          edges         degree1         degree2  mix.gender.F.M  mix.gender.M.M
##        -5.1948          2.7430          1.4678          3.4534          1.7138
##  nodematch.bar
##        -0.7061
##
## Monte Carlo Maximum Likelihood Coefficients:
##          edges         degree1         degree2  mix.gender.F.M  mix.gender.M.M
##        -5.2893          2.7322          1.4581          3.5327          1.8113
##  nodematch.bar
##        -0.6958
```

The Final Model's Effects

- Degree
- Nodemix (Gender)
- Nodematch (Bar)

The Final Model's Equation

- Recall the equation for the ERGM model
- $\log\left[\frac{\Pr(Y_{ij}=1|y_{ij}^C)}{\Pr(Y_{ij}=0|y_{ij}^C)}\right] = \sum A(Y_{ij})\eta_A d_A(y)$
- $\log\left[\frac{\Pr(Y_{ij}=1|y_{ij}^C)}{\Pr(Y_{ij}=0|y_{ij}^C)}\right] =$
  $-5.1948 + 2.7430\text{Degree}_1 + 1.4678\text{Degree}_2 + (3.4534 \times \text{Gender}_{FM}) + (1.7138 \times \text{Gender}_{MM}) + (-0.7061 \times \text{Bar})$

Degree

- This comprises 2 effects, each for the number of nodes in the network with degrees of 1 and 2 respectively.
- $\eta_2 \approx 2.7430$
- $\eta_3 \approx 1.4678$

Gender

- We used the `nodemix` effect in the ERGM which tests the effects of mixing and matching across the levels of gender, M and F.
- F-F is the reference class for this effect (note that there was only one F-F case in the dataset).
- $\eta_4 \approx 3.4534$
- $\eta_5 \approx 1.7138$

Bar

- $\eta_6 \approx -0.7061$

Remaining Work

- Centrality measures
- Treating links as nodes and nodes as links
- Suggestions?

References

De P, Singh AE, Wong T, et al. 2004. "Sexual network analysis of a gonorrhea outbreak." *Sexually Transmitted Infections* 80:280-285.

Carrington, P. and Scott, J., 2011. *The SAGE handbook of social network analysis. 1st ed.* Los Angeles [etc.]: SAGE Publications, pp.484-500.

Irene A. Doherty, Nancy S. Padian, Cameron Marlow, Sevgi O. Aral, Determinants and Consequences of Sexual Networks as They Affect the Spread of Sexually Transmitted Infections, The Journal of Infectious Diseases, Volume 191, Issue Supplement_1, February 2005, Pages S42–S54, https://doi.org/10.1086/425277

Wasserman, S. and Faust, K., 1994. Social network analysis: methods and applications. Cambridge: Cambridge University Press.

DE, PRITHWISH MHSc*; SINGH, AMEETA E. BMBS, MSc, FRCPC†; WONG, TOM MD, MPH, FRCPC*; YACOUB, WADIEH