A social network analysis of a gonorrhea outbreak in Alberta, Canada

J Steven Raquel

12/07/2021

Outline

- Background
- Methods
- Discussion
- Conclusion

Background

- This dataset concerns a localised outbreak of *Neisseria gonorrehoeae* in an indigenous community located in Alberta, Canada. it was originally analyzed in a paper by P 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 network consists of 89 individuals, both male and female, 17 of whom were found to be patrons of a local bar in the area.
- This work expands upon the original analysis by looking at other types of network centrality such as eigenvector centrality, as well as 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).

Background

- Gonorrhea is a sexually transmitted disease/infection (STD/STI) which can be transmitted orally, vaginally or anally.
- According to the Centers for Disease Control and Prevention, about 1 in 5 people in the United States have a STI, totalling nearly 68 million infections in 2018.
- Of the 26 million new infections in 2018, it is estimated about 1.6 million of them were gonorrhea.
- Although it can have many serious side effects, it can also be symptomless, leading to individuals unknowingly infecting their partners.

Background

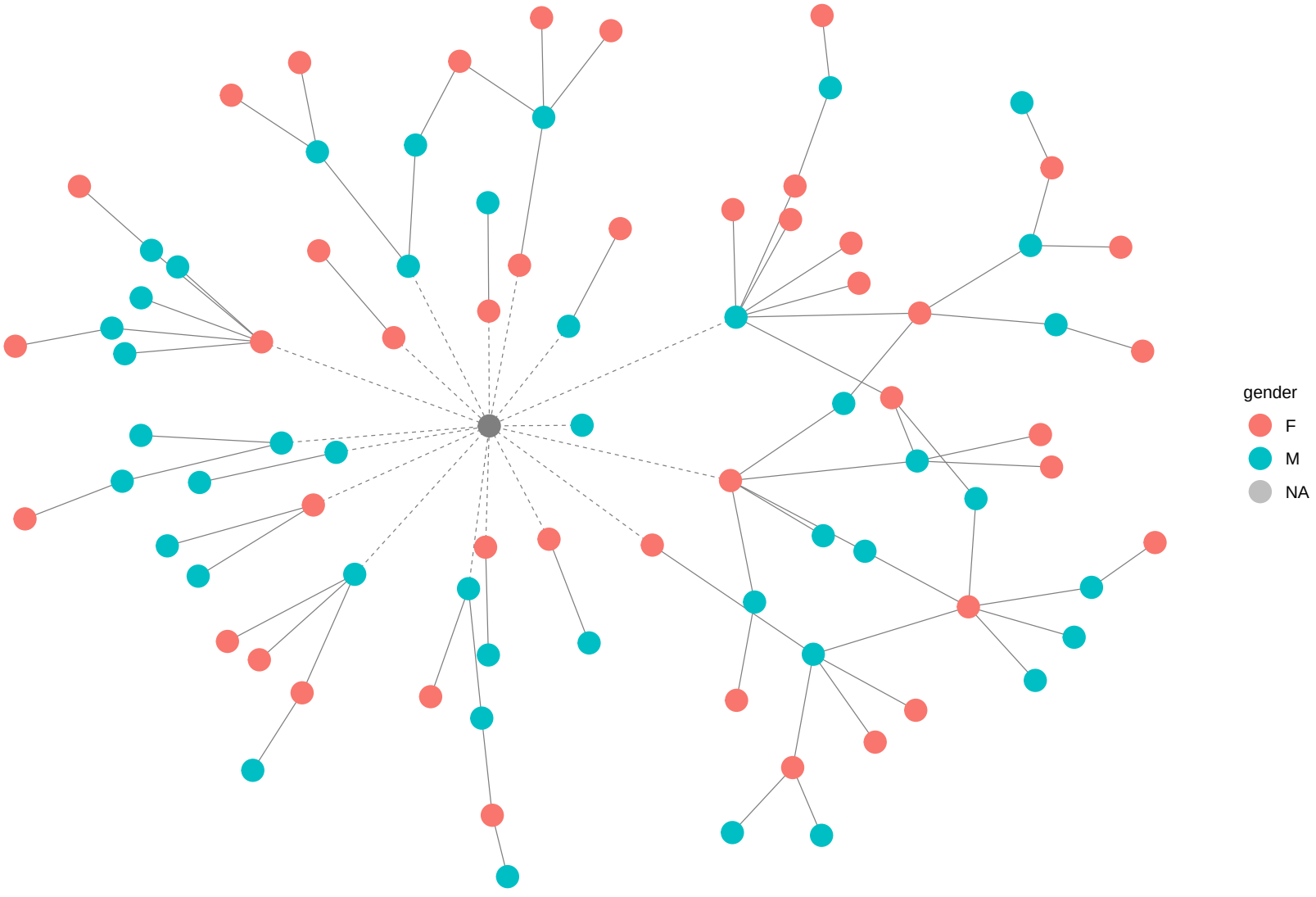
- Although the adjacency matrix for the data allowed us to analyze the network as a directed network, we thought it prudent to symmetrize the matrix (as was done in P De et al.)
- 2 of the datapoints were missing information regarding their gender so we dropped them from the dataset. They both had a degree of 1 so it wasn't thought that their exclusion would be impactful. This left 46 males and 46 females in the network.

Background

- Data on sexual networks can be collected either egocentrically or sociometrically.
- In the former case, the method involves interviewing the ego, and getting information regarding their alters (sexual partners) without in turn contacting their alters.

On studying sexual networks

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.

Distribution of ties

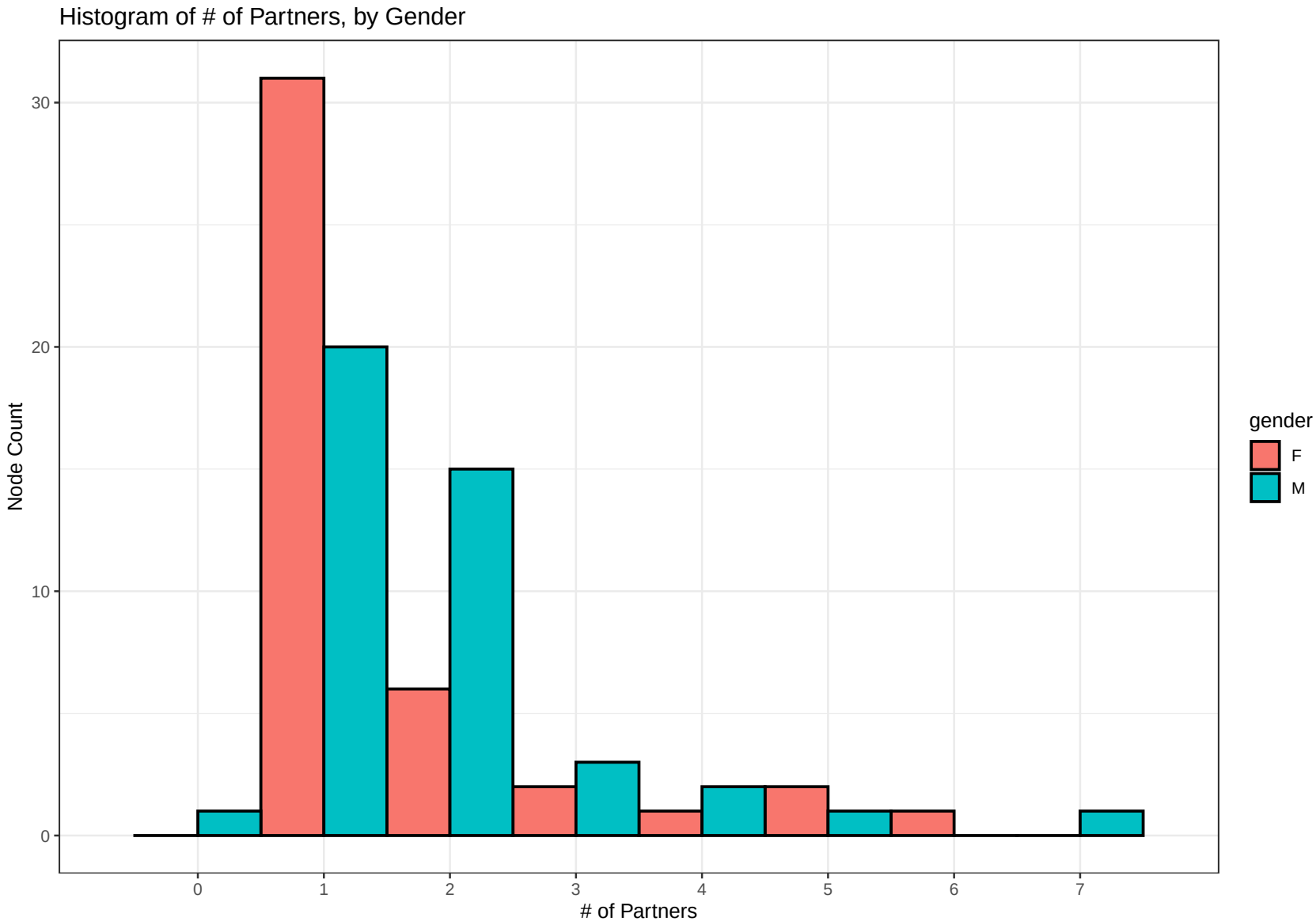


Figure 1: Histogram of partners, by gender.

Methods

- Degree Centrality
- Eigenvector Centrality
- Katz Centrality
- Average Distance
- Exponential Random Graph Model (ERGM)

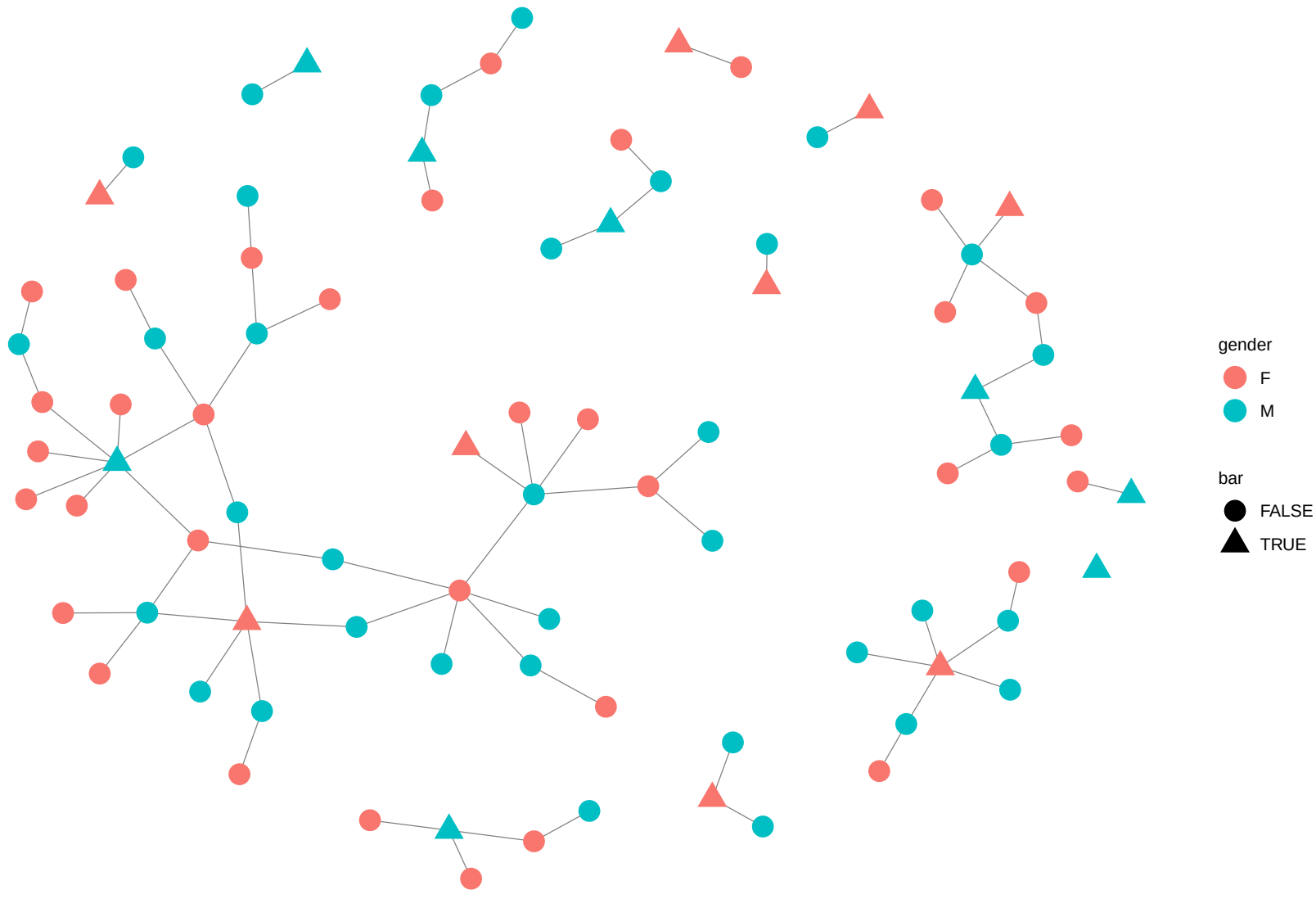
Exponential Random Graph Model (ERGM)

- 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.

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, and instead attendance of the bar was assigned to the vertices as an attribute, as was gender. This model was saved as `gonnet_net_nobar` in `R`.

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.

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.

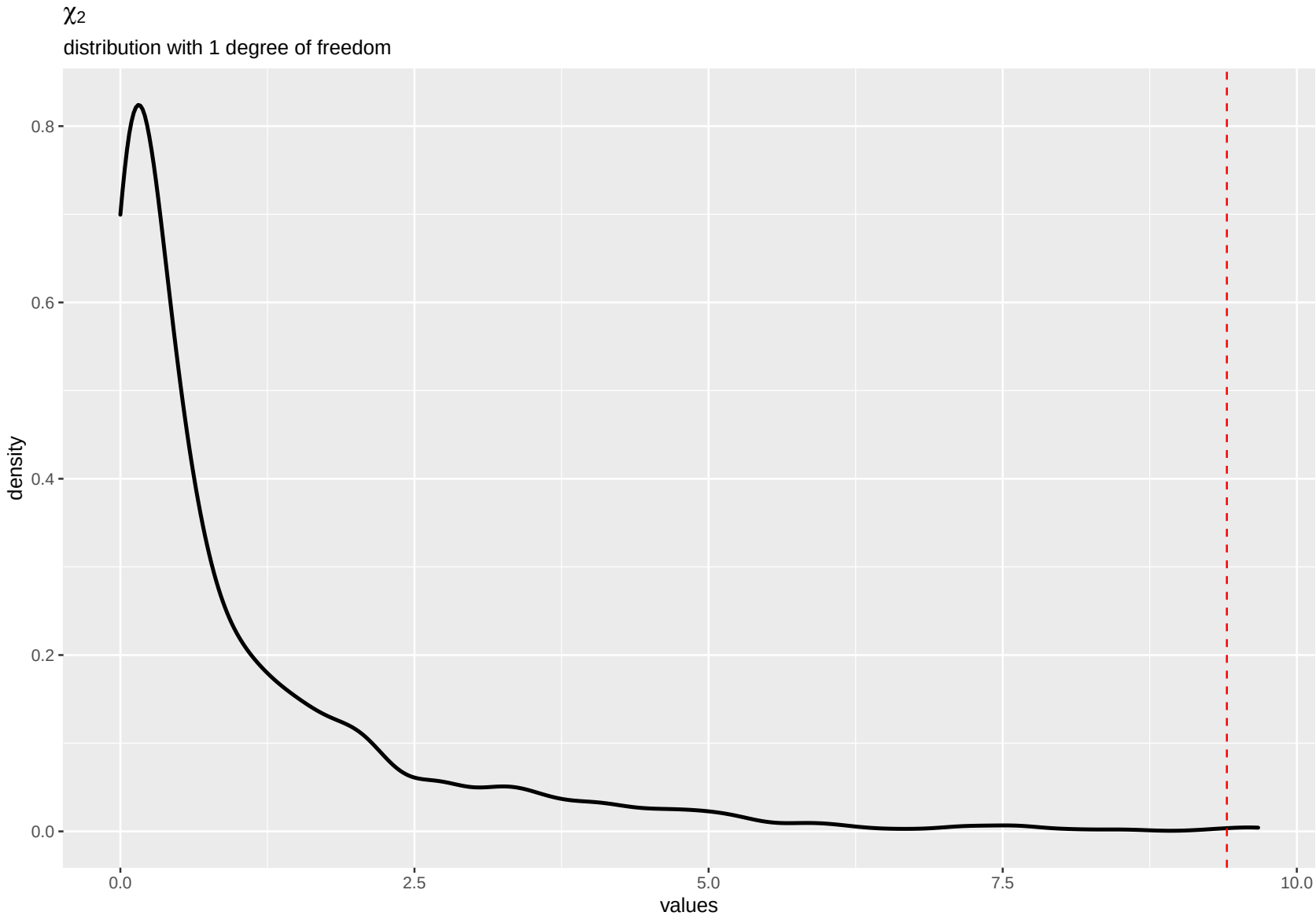
Likelihood Ratio Test

```
## Likelihood ratio test
##
## Model 1: gonnet_net_nobar ~ edges + degree(d = c(1:3)) + nodemix("gender") +
##     nodematch("bar")
## Model 2: gonnet_net_nobar ~ edges + degree(d = c(1:3)) + nodemix("gender")
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   7 -302.05
## 2   6 -306.76 -1 9.4055   0.002163 **
## ---
## 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.

Likeilhood Ratio Test



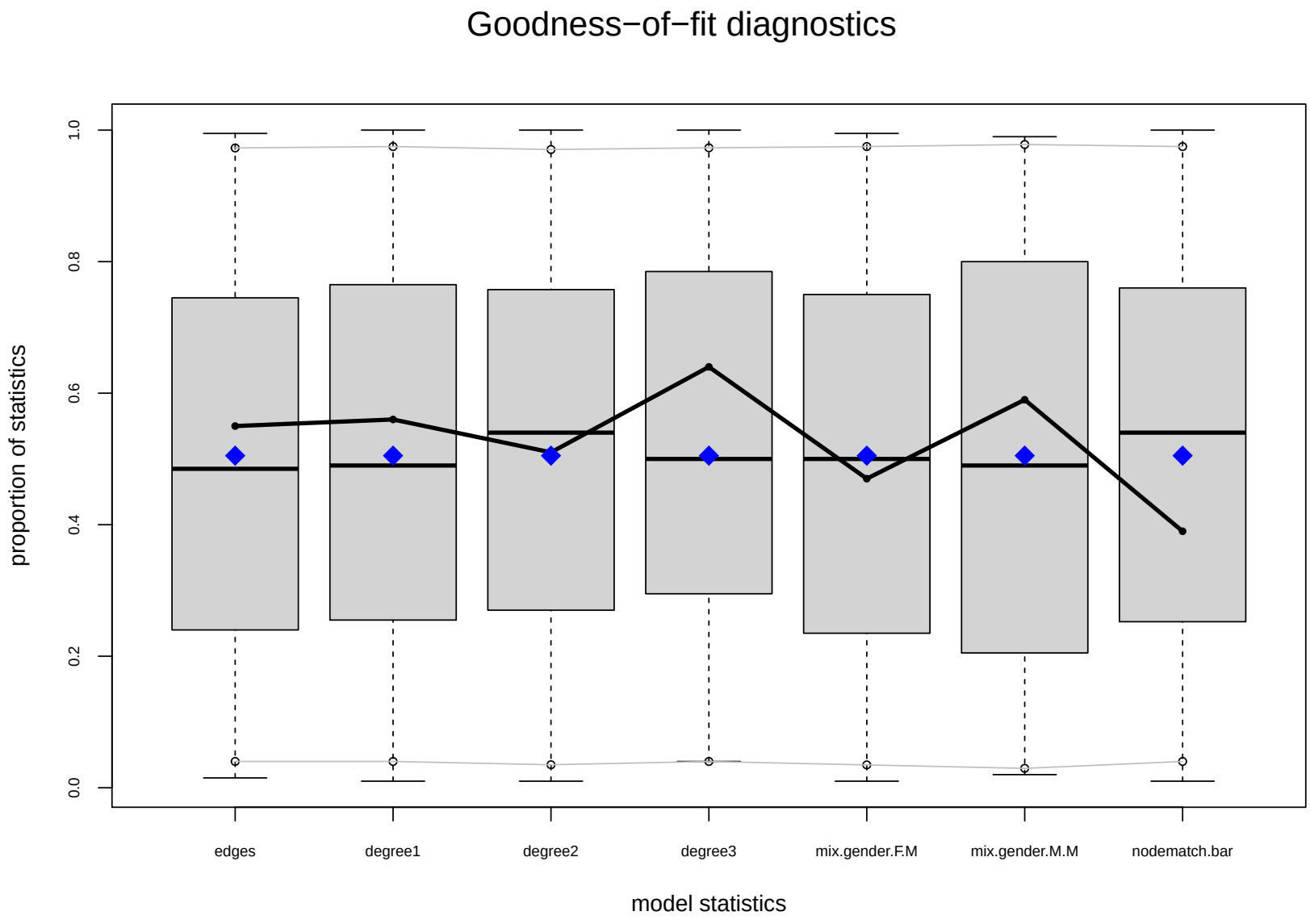
Markov Chain Monte Carlo (MCMC)

Model Diagnostics

```
## Call:
## .f(formula = ..1)
##
## Monte Carlo Maximum Likelihood Results:
##
##                Estimate Std. Error MCMC % z value Pr(>|z|)
## edges           -5.1709     1.0264      0  -5.038  < 1e-04 ***
## degree1          2.8404     0.7129      0   3.984  < 1e-04 ***
## degree2          1.5528     0.5867      0   2.646 0.008134 **
## degree3          0.1393     0.6371      0   0.219 0.826926
## mix.gender.F.M   3.4620     0.9892      0   3.500 0.000465 ***
## mix.gender.M.M   1.7712     1.0855      0   1.632 0.102734
## nodematch.bar   -0.7419     0.2462      0  -3.013 0.002584 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      Null Deviance: 5066.9  on 3655  degrees of freedom
##  Residual Deviance:  604.1  on 3648  degrees of freedom
##
## AIC: 618.1  BIC: 661.5  (Smaller is better. MC Std. Err. = 0.3362)


##                obs min  mean max MC p-value
## edges           74  60 74.69  92       1.00
## degree1         51  36 51.11  63       1.00
## degree2         21  10 20.57  32       1.00
## degree3          5   1  5.14  11       1.00
## mix.gender.F.M  67  53 67.25  81       0.94
## mix.gender.M.M   6   1  6.15  12       1.00
## nodematch.bar   38  28 36.98  49       0.78
```

Plotting Goodness of Fit



MCMC Diagnostics

```
## Sample statistics summary:
##
## Iterations = 245760:1069056
## Thinning interval = 2048
## Number of chains = 1
## Sample size per chain = 403
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                     Mean    SD Naive SE Time-series SE
## edges            1.06948 7.508   0.3740         0.3740
## degree1         -0.45658 5.120   0.2551         0.2551
## degree2         -0.09677 3.901   0.1943         0.1943
## degree3          0.05459 2.181   0.1087         0.1087
## mix.gender.F.M   1.18859 7.224   0.3598         0.3598
## mix.gender.M.M -0.13151 2.363   0.1177         0.1177
## nodematch.bar    0.21092 4.953   0.2467         0.2467
##
## 2. Quantiles for each variable:
##
##                2.5% 25% 50% 75% 97.5%
## edges           -11  -5   1   6 17.95
## degree1         -10  -4  -1   3  9.95
## degree2          -8  -3   0   3  7.00
## degree3          -4  -2   0   1  5.00
## mix.gender.F.M  -11  -4   1   6 16.00
## mix.gender.M.M   -4  -2   0   1  5.00
## nodematch.bar    -9  -3   0   4  9.95
##
##
## Are sample statistics significantly different from observed?
##                   edges     degree1     degree2    degree3 mix.gender.F.M
## diff.      1.069478908 -0.45657568 -0.09677419 0.05459057   1.1885856079
## test stat. 2.859487043 -1.79002141 -0.49802123 0.50239871   3.3030672671
```

Model Interpretation