Analyzing centralitty measures for a sexual network of gonorrhea transmission

SOC 280 Fall 2021

J Steven Raquel

Abstract

Introduction

Contact tracing for sexually transmitted diseases (STDs) such as gonorrhea, chlamydia, syphilis etc. is a persistent epidemiological problem, as it depends on individuals getting routinely tested as well as informing their sexual partners of their positive diagnosis should they receive one. Compounding this with the fact that many of these positive cases can be symptomless but still contagious creates a serious issue. Gonorrhea in particular is a disease that can be asymptomatic in both men and women who have it, that can go so far as causing infertility or lead to a life-threatening condition.

This dataset, constructed in the form of an adjacency matrix, contains 89 nodes, one of which is the "event" of attending a bar (i.e. when a node has a tie with this bar node, it means they attend the bar). Two of these nodes (denoted by `x` and `x2`) are missing information about their gender which is otherwise indicated by an `m` or `f` in the label of the respective node, followed by a number with which to differentiate them.

The network is directed, and the criteria for some node i to have a tie with some other node j must be that that ith node named node j as a prior sexual partner.

The data was collected from a series of adjacent aboriginal communities located in the province of Alberta, Canada, where public health officials took note of a local gonorrhea outbreak and confirmed that attendance at a local bar in one such community was associated with infection (De 2003).

Background

Data Collection

The sociometric approach for enumerating a sexual network entails an iterative process in which the subject names past sexual partners, who are them traced and interviewed to identify whether they are linked and also to identify other contacts in the network (Doherty 2005). Conversely, the egocentric approach bases the network entirely on the information volunteered by the original subject.

There are a number of shortcomings that are inherent to mapping out a sexual network, the first being that the design suffers from incomplete-network bias when partners cannot be traced or recruited for a variety of reasons (Doherty 2005). For example, the contact tracing of sexual partners is relegated to those who have a positive test result, and since such individuals may have markedly different positions within a network, tracing a network consisting solely of STI-positive partners has an inherent bias.

For example, consider that the prevalence of STI infection has a high negative correlation with condom usage i.e. using condoms infrequently, or not at all, comes with a higher risk of contracting an STI. If we assume individuals who engage in this kind of risky behavior on a regular basis are more likely to test positive for an STI, then the mapping of our sexual network may predominantly include only those individuals who partake in this behavior to some degree, and their partners with whom they had safe sex with (or simply did not transmit an STI to) will go unrecorded.

There are some behavioral considerations taken into account when considering this data, as is inherent with studies in which data is based on information nominated directly from the subjects. In a sexual network such as this the information regarding past sexual partners is volunteered on the part of the individual who represents the node in the graph, and as such, it's subject to an individual possibly withholding information, or simply not having the information at all. For example, an individual may be reluctant to disclose the identity of a sexual partner if they are in a monogamous relationship and to disclose this event would mean owning up to infidelity, or perhaps if the sex was in exchange for money i.e. sex work. It could also be that the sex occurred in an anonymous context and they simply do not have the information on the individual.

There is a social stigma attached to promiscuity (having a high number of sexual partners), as well as living with an STI, so it's important to note that the data may be skewed by dishonesty on the part of the individuals comprising the dataset. For example, a person who receives a positive test result for a sexually transmitted disease may refrain from naming *all* of their recent sexual partners, either to avoid having to communicate the uncomfortable truth of either having contracted or transmitted a disease, or to avoid judgment for divulging what may be perceived as a high number of sexual partners. Societal attitudes towards sex and sexual health both in a Western context and also in an indigenous/Aboriginal context can and should be kept in mind when drawing conclusions from this data.

Some of these biases could be mitigated if we had all of the complete information on infection/non-infection status and were able to recruit all members of a sexual network, but this is impractical, improbable, or even illegal in some areas. These biases are inherent to these types of models but these studies still are important in anticipating and modeling the spread of STIs due to the highly social component of transmission, compared to for example, the spread of influenza.

Wasserman and Faust (1994) define "cutpoints" and "bridges" as nodes and ties respectively

that cause the graph in which they are contained to have less components if they were to be taken out from the graph. In other words, they are the nodes or ties that connect what would be otherwise unconnected sub-graphs. These are crucial in sexual network analysis because these nodes and ties are the difference between whether a certain network may propagate an STI outbreak or not.

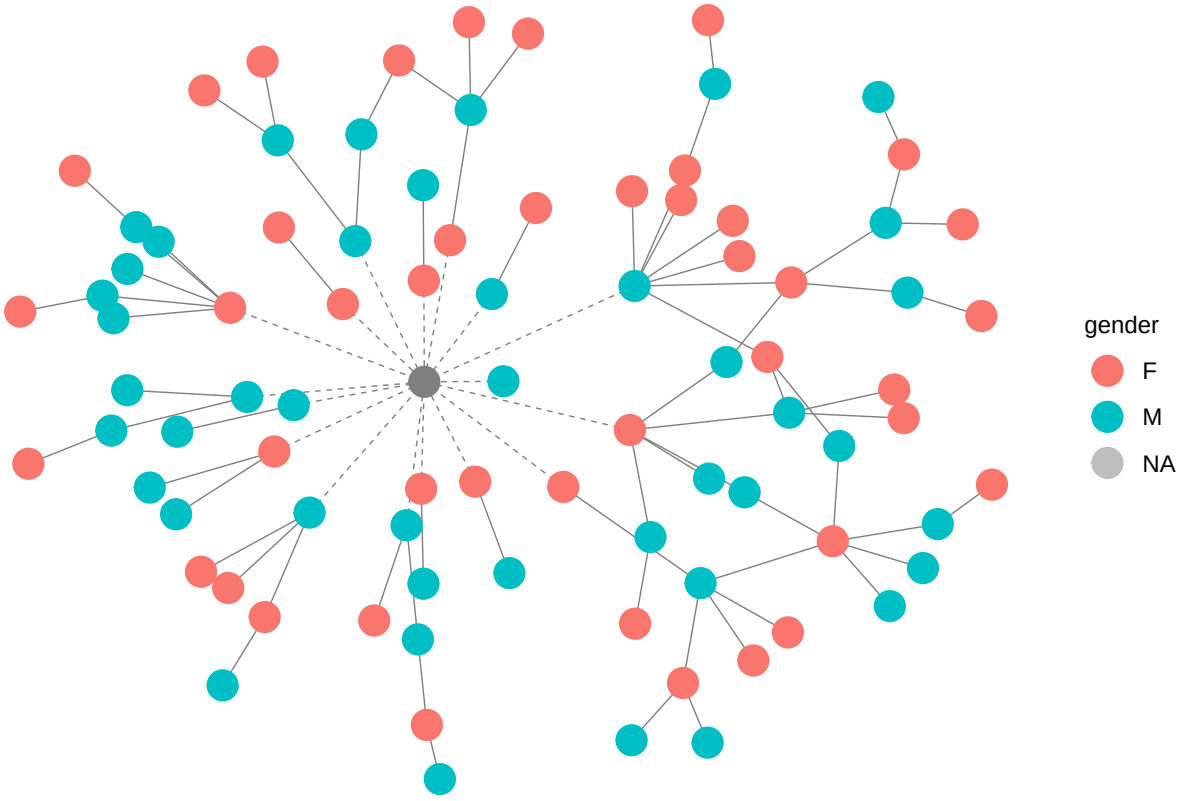


Figure 1: The network with bar as a node.

Looking at this initial sociogram, where the bar node is firmly in the center, we see that it has outgoing connections to 17 nodes, which themselves have connections to at least one other node in the rest of the network.

The majority of ties are between individuals of the opposite sex, e.g. male-to-female or female-to-male, but there are a minority of instances where individuals have a tie to individuals of

the same sex, e.g. `m112` has ties with both `m106` and `m107`, who both in turn have ties to at least one female node.

The node `m010` is also unique in that it happens to have ties to one female node (`f024`) and one male node (`m018`), the latter of which in turn has a tie with a female node (`f023`). The idea of men who have sex with men (MSM) or women who have sex with women (WSW) acting as bridging nodes between otherwise disparate sexual networks was something considered in the exploratory analysis but there just wasn't enough data to delve deeper into this subject.

Centrality

According to Borgatti and Everett (2006), *centrality* is a summary index of a node's position in a graph, based on sums or averages of one of several things: 1) the number of edges the node has, 2) the length of the paths that end up at the node, or 3) the proportion of paths that contain the node inside of it (not as an endpoint).

Different measures of centrality depend on functions of one of these aspects and communicate different things about a node, depending on the algorithm for the centrality measure. In this paper, we're going to look different kinds of centrality as a means of quantifying the impact of individual nodes on the network.

Principal Component Analysis

Principal component analysis is a method for drawing out important variables from a dataset and is used frequently when dimensionality reduction is necessary.

Degree Centrality

For an undirected network such as this, then the degree of some node i is just the count of ties it has to other nodes. Supposing we had some adjacency matrix A, then the degree of

Gender	Mean Number of Partners
F	1.605
M	1.837

node i is equal to

$$D_i = \sum_j a_{ij}$$

where j is the number of nodes in the network and $a_{ij} = 1$ given that node i has a tie with node j, and zero otherwise. Note that in this network those who attended the bar have a tie with the bar node to represent that relationship, hence in order to treat degree as a means of quantifying sexual partners, we added in an adjustment in the form of a new variable called `partners` which subtracts 1 from the degree of those bar attendees.

As noted earlier in Figure 3, the largest component contains 2 nodes (`m017` and `f011`) with degrees of 6 and 7 respectively. Noting their respective positions in the network, we see that a high degree centrality doesn't necessarily imply a large impact on the network in general, as long as an ego's alters do not themselves have a high degree centrality. The converse is also true where an ego can have a small degree but still be connected to an alter than itself has a high degree, e.g. `f014` was a bar attendee with a degree of one, connecting itself to `m208` who accordingly has 4 other alters, some of whom have more among them.

It's possible in fact for an ego to have a high degree but whose alters themselves have a very low degree, such that their respective sub-component is not very large, e.g. `f004` which has an outdegree of 5, but all of its alters have a degree of only 1 or 2.

Other forms of centrality are better at measuring this impact, for example, eigenvector centrality, which will be discussed later in this literature.

We observe from the data that on average, men have a slightly higher number of sexual partners than women in this sample.

We further observe that those who attend the bar end up having an average of about 1.6

Attended Bar	Mean Number of Partners
FALSE	1.623
TRUE	2.118

sexual partners compared to as many as 2.1 partners from those who did. Again, these two quantities are technically dependent (owing to the fact that these are not isolated samples, i.e. the bar attendees can be partnered with the non-attendees), so we can't formally compare them with a t-test, but it is something to take note of going forward. Later on in the literature we will bar attendance as a data attribute when developing an exponential random graph model (ERGM), in order to model this relationship, should it exist.

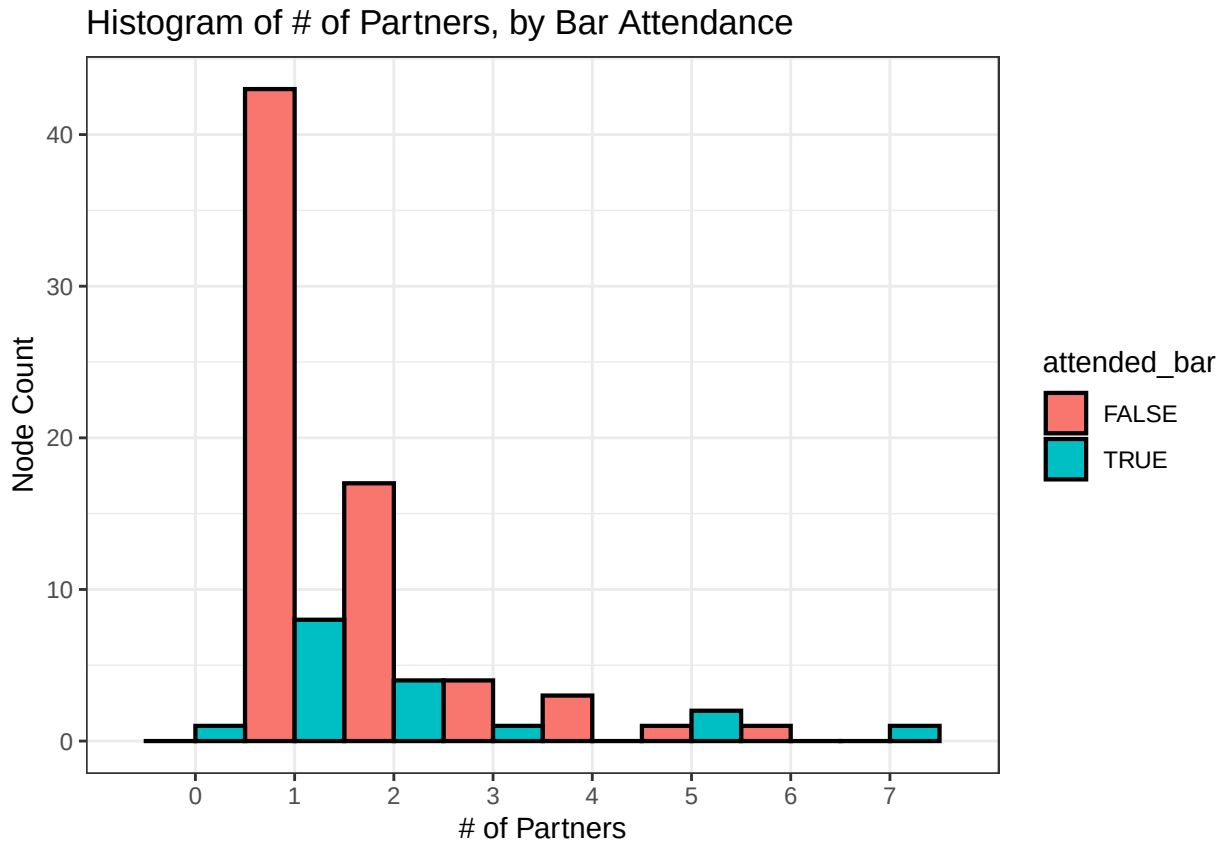


Figure 2: Distribution of number of partners, by bar attendance.

Figure 2, which depicts a histogram of number of partners, colored by bar attendance, shows that a large number of non-bar patrons only have one partner; this is mostly bolstered by

the fact that most of the members of the network do not frequent the bar, and also there are a sizeable number of individuals in the network who were simply alters of an ego who did attend the bar and do not have any other alters themselves.

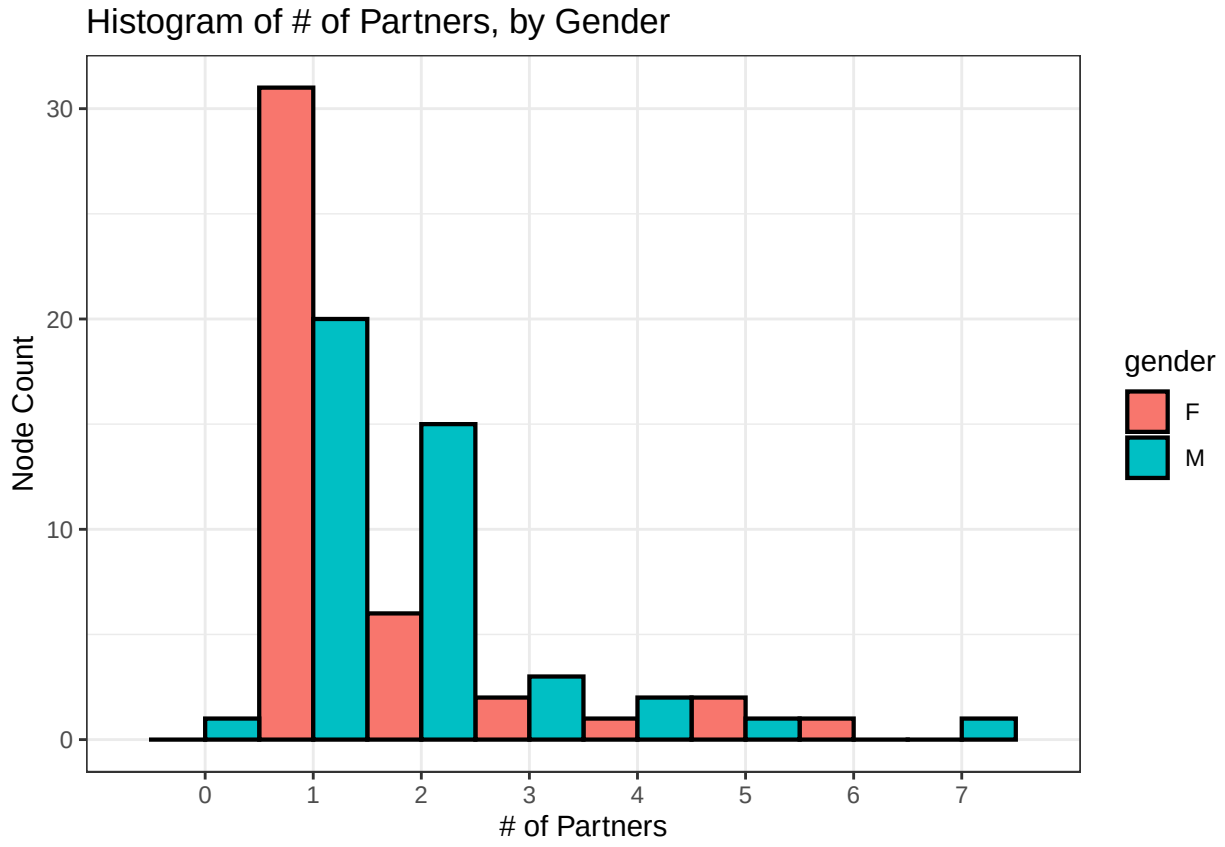


Figure 3: Histogram of partners, by gender.

In this case, we are seeing difference in the two distributions of men and women; there are many more women (10+) who had only one sexual partner in this network, with the remaining numbers between comparatively small. Conversely the majority of the men are split between either one or two partners. Overall the distributions for 3 or more partners are similar across genders.

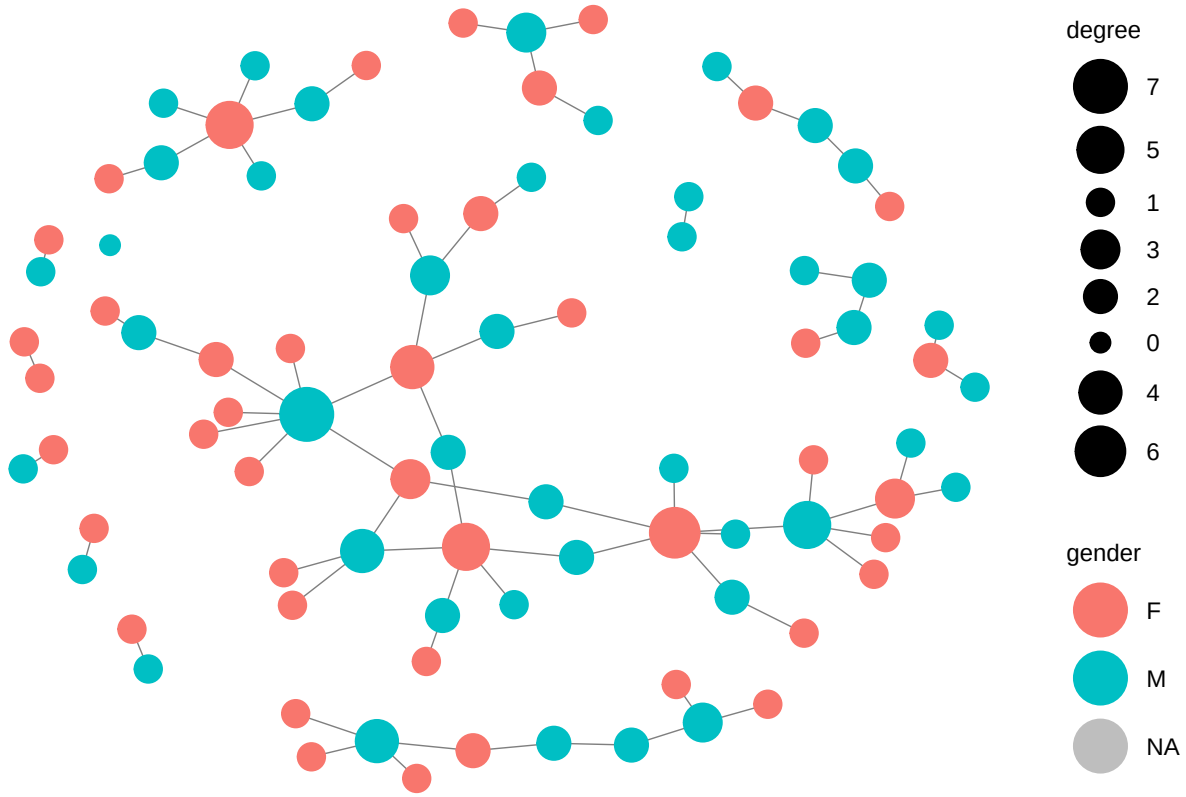
The analysis gave that the overall average degree for the directed graph is approximately 1, implying that on average individuals named one sexual partner. It's important to note that many outer edges have an degree of 0, which skews down the mean calculation somewhat.

The implication of this observation is two-fold. For one, rather than most individuals in the network having many sexual partners, the implication of how degree is distributed in this network is that most individuals have as few as one, but there is a minority of individuals (of both genders) who named more.

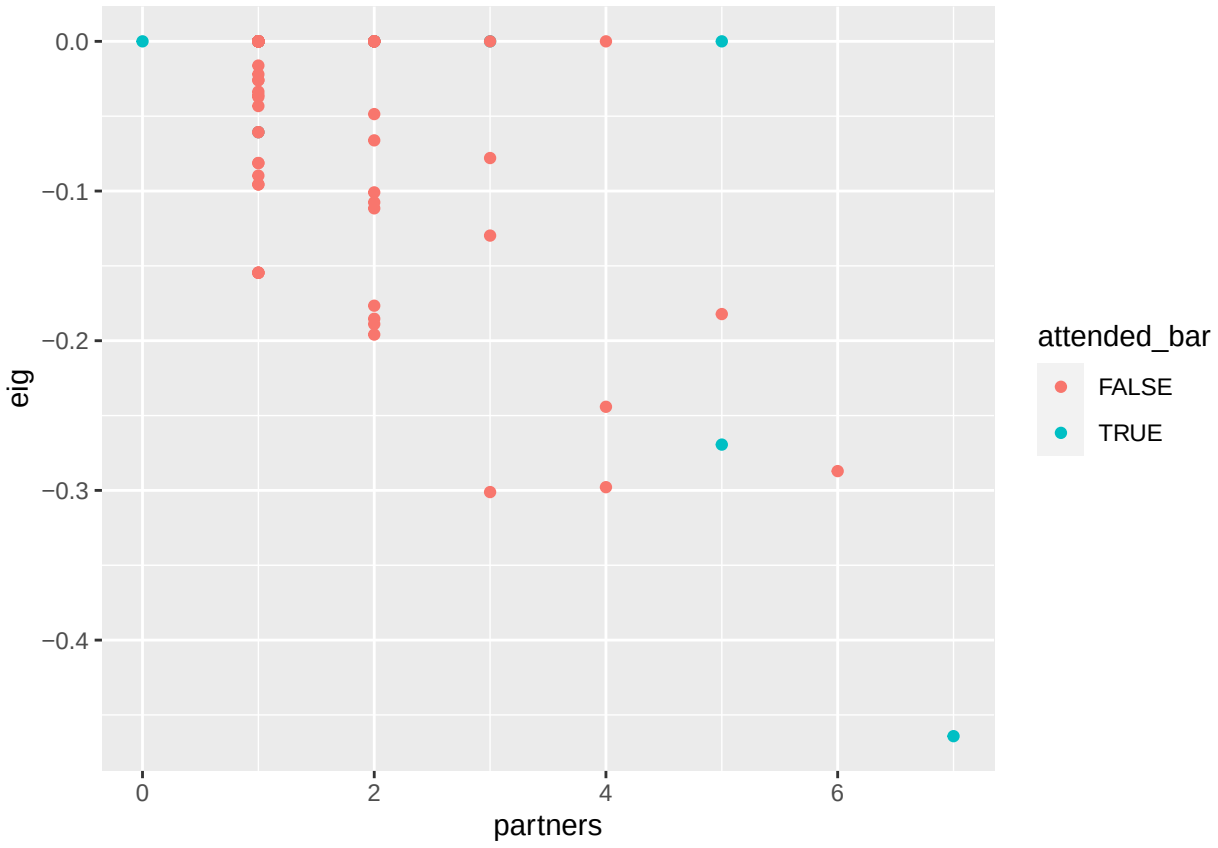
Eigenvector Centrality

Eigenvector centrality is calculated both as a function of a node's degree but also as a function of the degree of the nodes it is connected to. In other words, a node with a high eigenvector centrality is well-connected to nodes that are themselves well-connected. Contextually, a node in this setting would have a high number of partners who also have a high number of partners.

```
eigenvectors <- eigen(gonnet_net_nobar)
eig <- eigenvectors$vectors[,1] %>% as.numeric()

# adding eigenvector centrality as a vertex attribute to the networkr
network::set.vertex.attribute(gonnet_net_nobar, 'eig', eig)
network::set.vertex.attribute(gonnet_net_nobar, 'bar', attended_bar[-1])
# plotting the network, sized by eigenvector centrality
ggnet2(gonnet_net_nobar,
  color = "gender",
  palette = c("F" = "#F8766D", "M" = "#00BFC4", "NA" = "grey"),
  mode = "fruchtermanreingold",
  label = F,
  size = "degree")
```




```
gonnet_df_nobar <- gonnet_df[-1,]

gonnet_df_nobar <- cbind(gonnet_df_nobar, eig)

gonnet_df_nobar %>%

  ggplot(aes(x = partners, y = eig, col = attended_bar)) +

  geom_point()
```



Katz Centrality

Average Distance

Exponential Random Graph Model (ERGM)

Exponential random graph models (ERGMs) are a family of statistical models for social networks that permit inference about prominent patterns in the data, given the presence of other network structures (Carrington and Scott, 2011). For a given set of n actors, an ERGM models an observed network x by assigning a probability to every network of n actors, and the form of such a model is as follows

$$\Pr(X = x) = \frac{1}{k} \exp\{\sum_{A} \eta_A g_A(x)\}$$

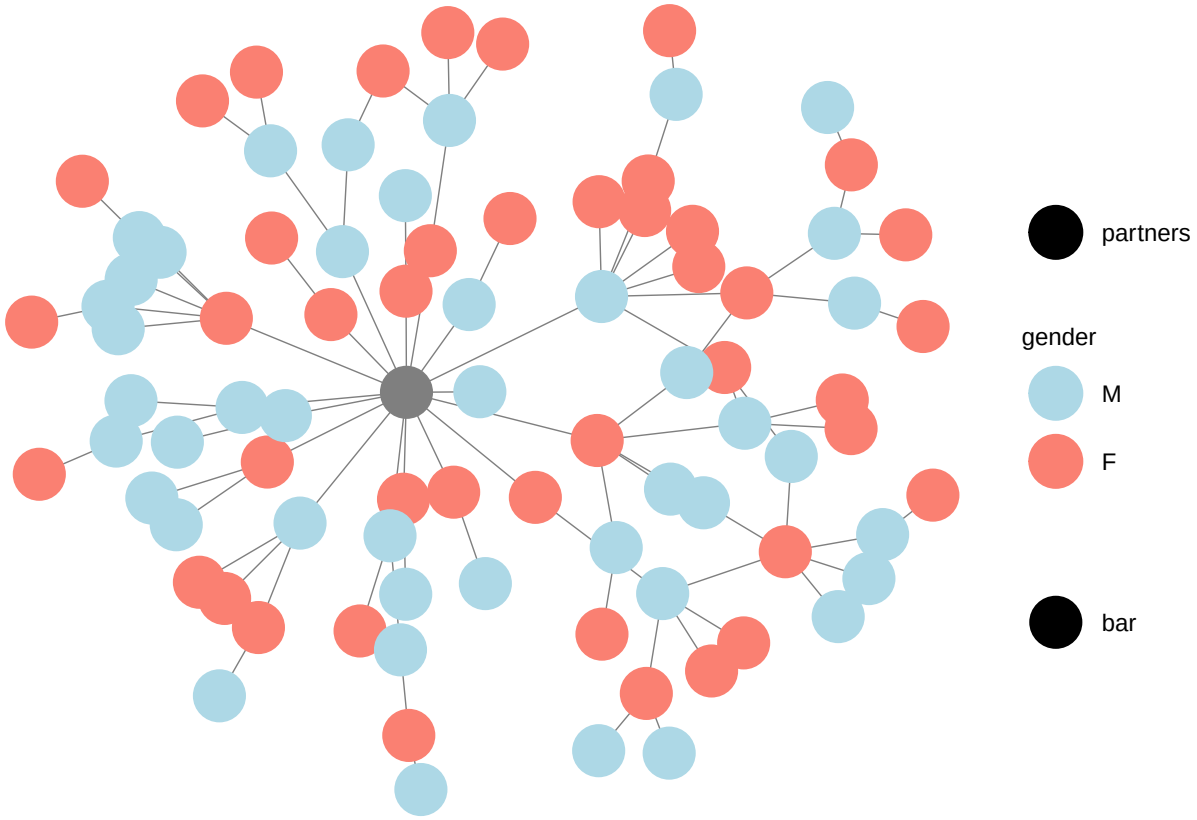


Figure 4: Gonorrhea network, sized by degree.

where the sum is over all configuration types A;

- η_A is a parameter correpsonding to configuration type A;
- $g_A(x)$ is the *network statistic* for A and is the number of configurations A observed in x
- k normalizes this to be a proper probability distribution.

This equation implies that there is a probability distribution of all possible networks with n nodes, with each such network having their own distinct probability.

The ERGM model requires a one-mode network, so we can create one by dropping the bar node from the data (while still retaining the information of bar attendance as a vertex attribute).

Figure 3 represents the network in which the bar node is not a part of the graph; instead bar attendance is regarded as a vertex attribute in the data, as is gender. There are 7 sub-components of size greater than 2. Most interestingly is the component that comprises almost half (39) of all of the nodes in the dataset, despite containing as few as 3 bar attendees.

```
## Starting maximum pseudolikelihood estimation (MPLE):

## Evaluating the predictor and response matrix.

## Maximizing the pseudolikelihood.

## Finished MPLE.

## Starting Monte Carlo maximum likelihood estimation (MCMLE):

## Iteration 1 of at most 60:

## Optimizing with step length 0.4429.

## The log-likelihood improved by 2.7651.
```

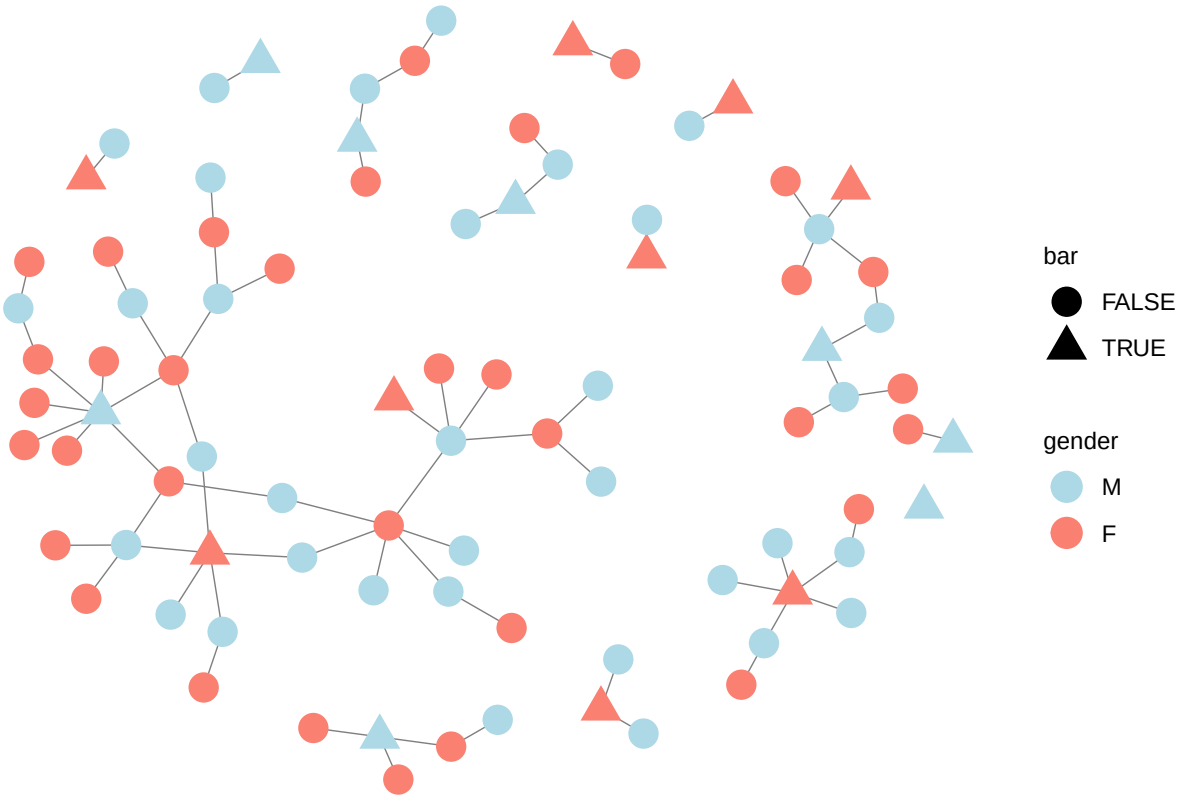
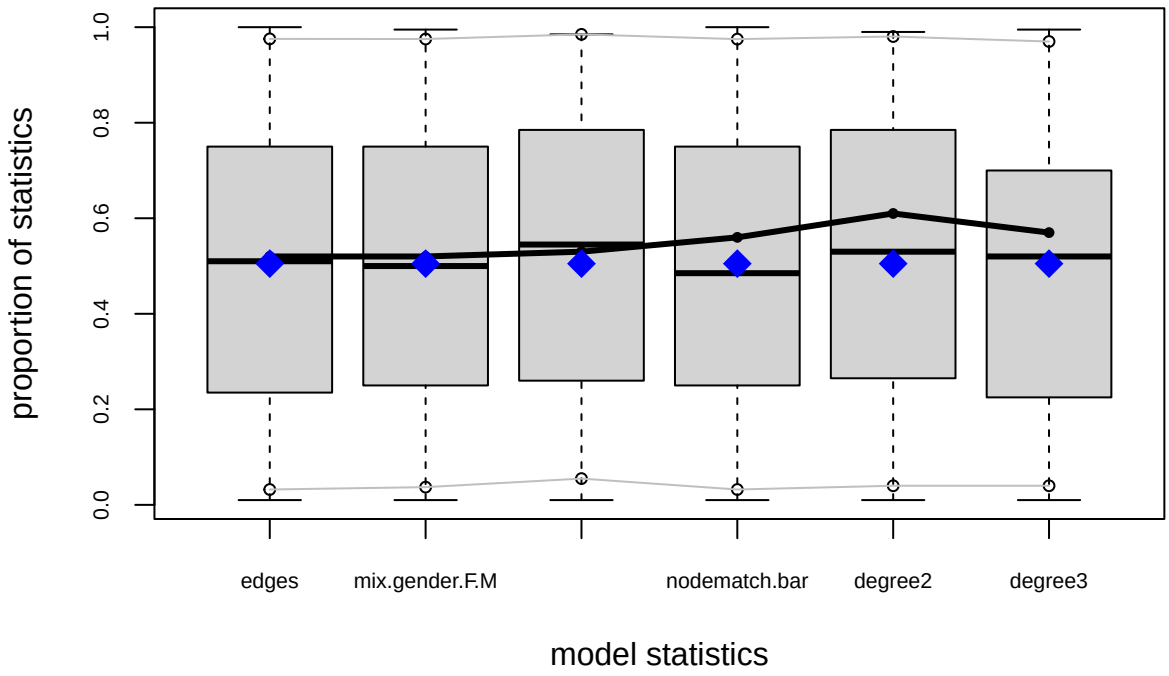


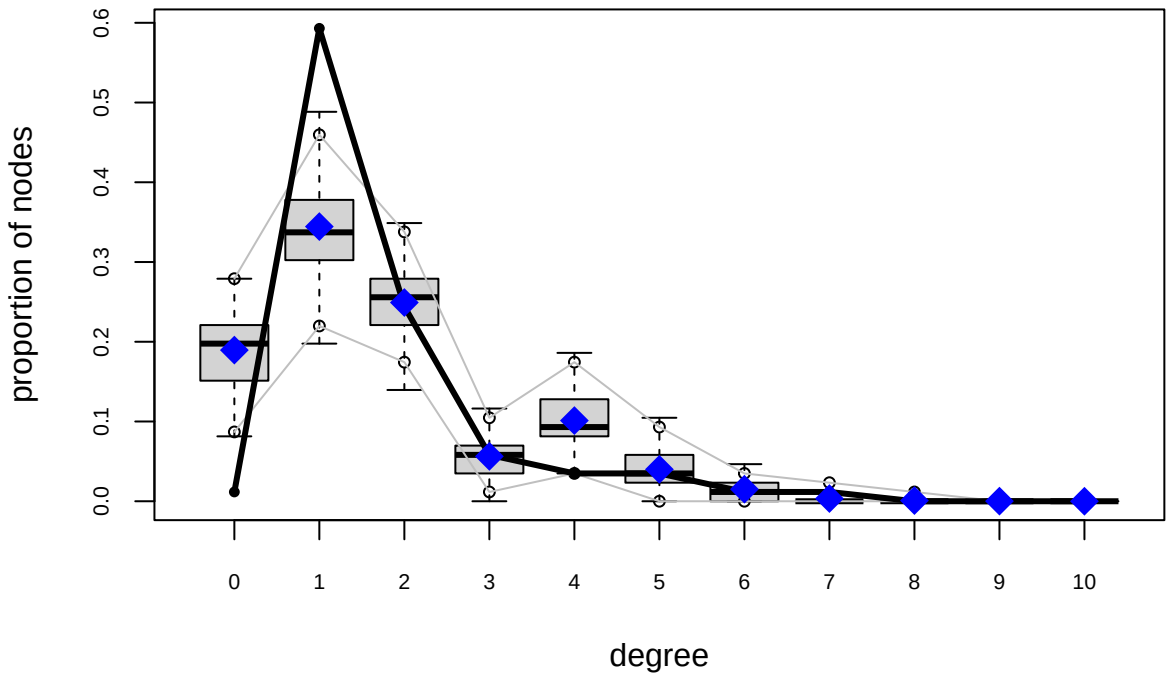
Figure 5: The network without the bar as a node.

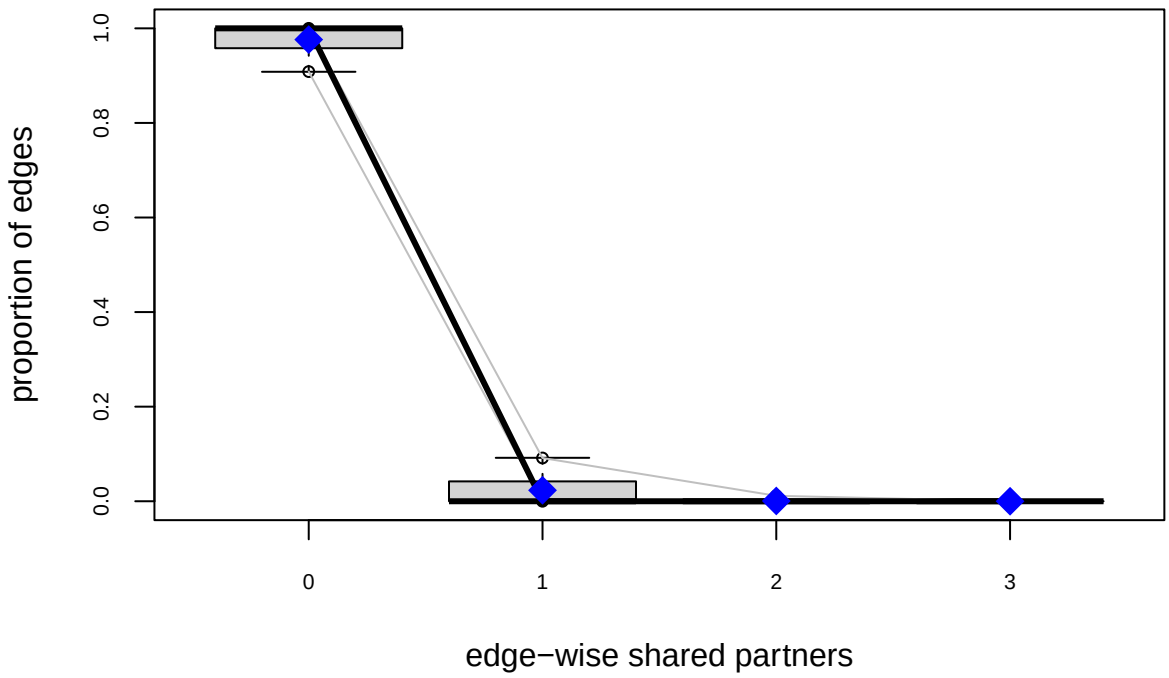
```
## Estimating equations are not within tolerance region.

## Iteration 2 of at most 60:

## Optimizing with step length 0.7617.

## The log-likelihood improved by 2.1098.

## Estimating equations are not within tolerance region.

## Iteration 3 of at most 60:

## Optimizing with step length 1.0000.

## The log-likelihood improved by 1.3038.

## Estimating equations are not within tolerance region.

## Iteration 4 of at most 60:

## Optimizing with step length 1.0000.

## The log-likelihood improved by 0.1618.

## Estimating equations are not within tolerance region.

## Iteration 5 of at most 60:

## Optimizing with step length 1.0000.

## The log-likelihood improved by 0.1832.

## Estimating equations are not within tolerance region.

## Iteration 6 of at most 60:

## Optimizing with step length 1.0000.

## The log-likelihood improved by 0.7846.
```

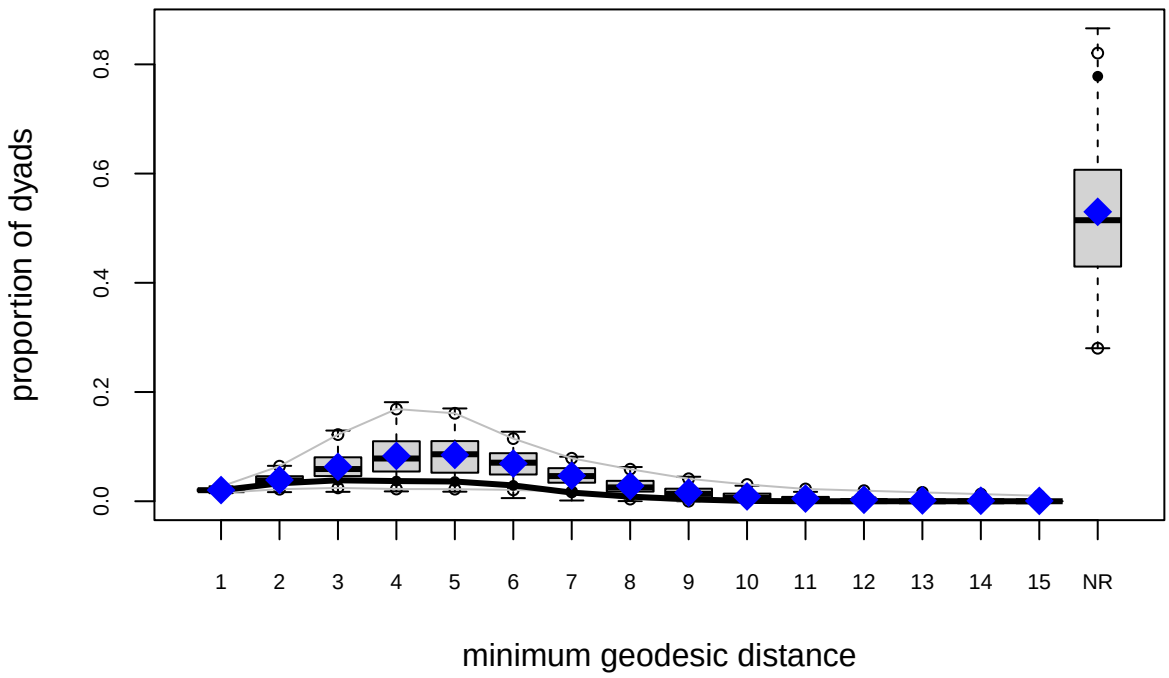
```
## Estimating equations are not within tolerance region.

## Iteration 7 of at most 60:

## Optimizing with step length 1.0000.

## The log-likelihood improved by 0.6807.

## Estimating equations are not within tolerance region.

## Iteration 8 of at most 60:

## Optimizing with step length 1.0000.

## The log-likelihood improved by 0.2494.

## Estimating equations are not within tolerance region.

## Iteration 9 of at most 60:

## Optimizing with step length 1.0000.

## The log-likelihood improved by 0.2444.

## Estimating equations are not within tolerance region.

## Iteration 10 of at most 60:

## Optimizing with step length 1.0000.

## The log-likelihood improved by 0.0255.

## Convergence test p-value: 0.8199. Not converged with 99% confidence; increasing sampl
## Iteration 11 of at most 60:
## Optimizing with step length 1.0000.
## The log-likelihood improved by 0.0140.
## Convergence test p-value: 0.8548. Not converged with 99% confidence; increasing sampl
## Iteration 12 of at most 60:
## Optimizing with step length 1.0000.
## The log-likelihood improved by 0.0661.
## Convergence test p-value: 0.0679. Not converged with 99% confidence; increasing sampl
```

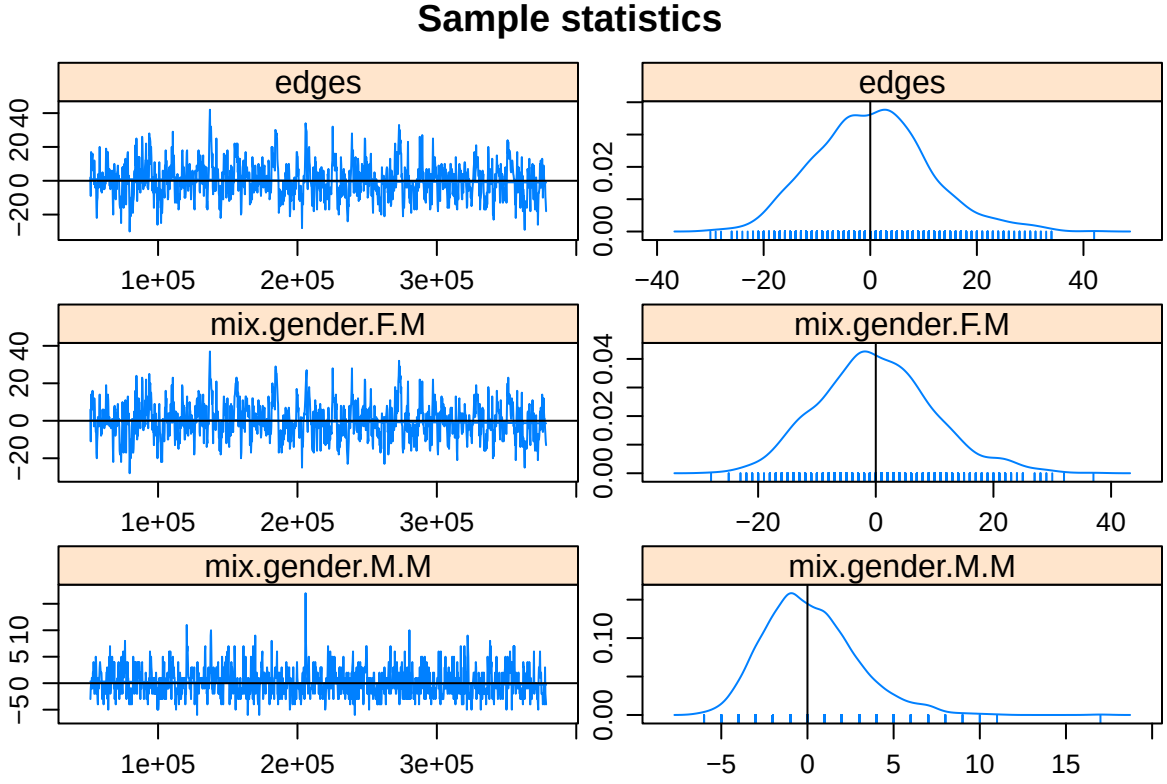
```
## Iteration 13 of at most 60:
## Optimizing with step length 1.0000.
## The log-likelihood improved by 0.0120.
## Convergence test p-value: 0.0032. Converged with 99% confidence.
## Finished MCMLE.
## Evaluating log-likelihood at the estimate. Fitting the dyad-independent submodel...
## Bridging between the dyad-independent submodel and the full model...
## Setting up bridge sampling...
## Using 16 bridges: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 .
## Bridging finished.
## This model was fit using MCMC.  To examine model diagnostics and check
## for degeneracy, use the mcmc.diagnostics() function.
```

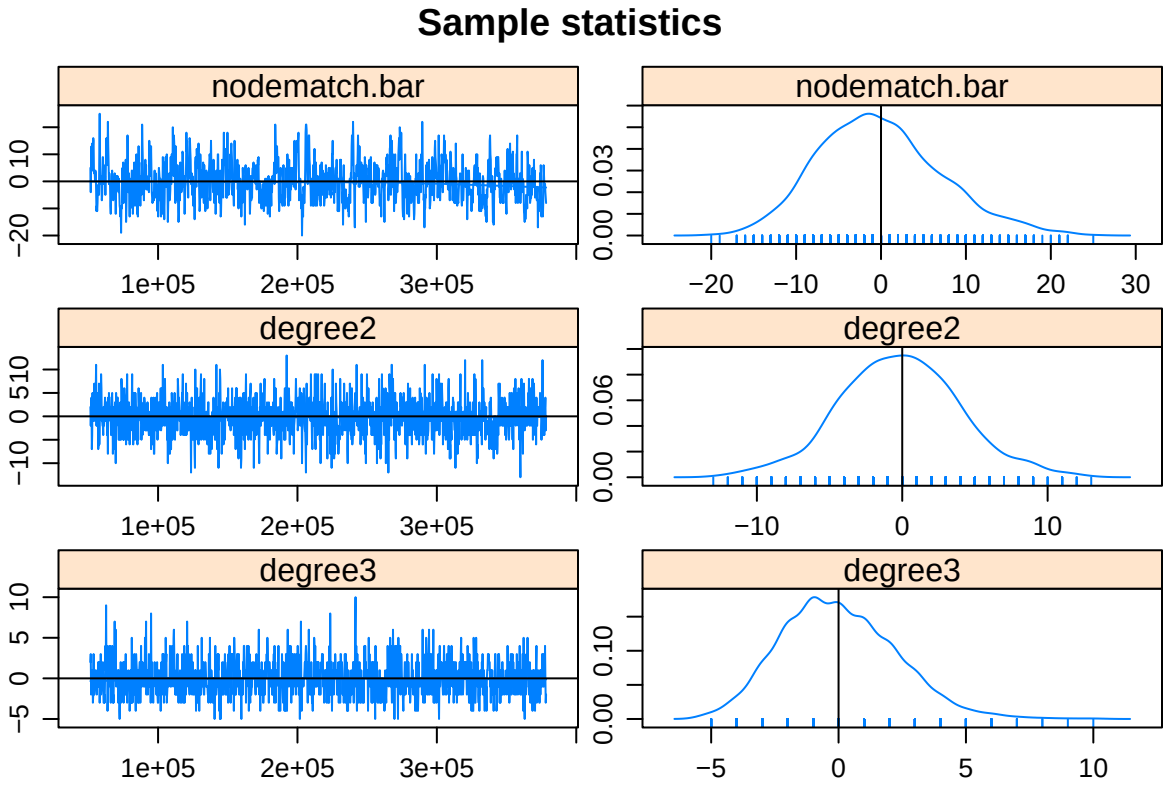



proportion of nodes
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
1
2
3
4
5
6
7
8
9
10
degree


proportion of edges
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
edge−wise shared partners


Goodness−of−fit diagnostics
proportion of dyads
0.0
0.2
0.4
0.6
0.8
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
NR
minimum geodesic distance


Sample statistics
edges
edges
mix.gender.F.M
mix.gender.F.M
mix.gender.M.M
mix.gender.M.M
1e+05
2e+05
3e+05
−40
−20
0
20
40
−5
5
10
15
0.00
0.02
0.04
0.10




```
lmtest::lrtest(ergm_model8, ergm_model7)
```

```
## Likelihood ratio test
##
## Model 1: gonnet_net_nobar ~ edges + nodemix("gender") + nodematch("bar",
##     levels = 1) + degree(d = c(2:3))
## Model 2: gonnet_net_nobar ~ edges + nodemix("gender") + nodematch("bar")
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -324.52
## 2   4 -327.19 -2 5.3213     0.0699 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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.

Appendix