Analyzing centralitty measures for a sexual network of gonorrhea transmission

SOC 280 Fall 2021

J Steven Raquel

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 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 was sourced from an earlier 2001 study by the same authors (De, et al. 2001) in which they formulated a plan to address the outbreak.

Background

On 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. In Canada, reported cases increased by 38.9% between 2003 and 2012, with rates highest in the 20-24 year old age group. It is the second most commonly reported STI in Canada (Totten, et al. 2015).

On First Nations in Alberta, Canada

According to the 2016 Canadian census, approximately 7% of Albertans identify as First Nations, one of the indigenous groups native to Canada, 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 resoruces such as those in more urbanized areas and more populated by non-indigenous people.

Data

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.

The design suffers from incomplete-network bias when partners cannot be traced or recruited for a variety of reasons (Doherty 2005). 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 data is 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. Societal attitudes towards sex and sexual health both in a Wester/Canadian context and also in an indigenous/Aboriginal context can and should be kept in mind when drawing conclusions from this data.

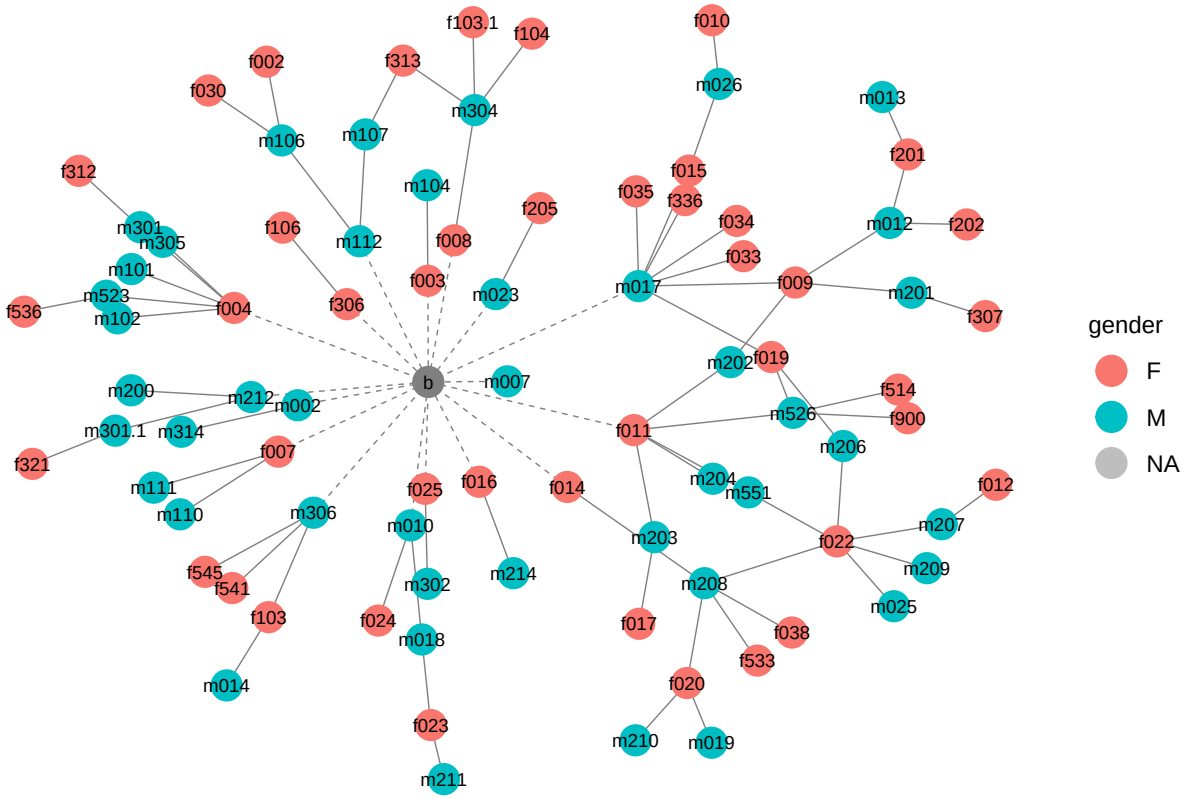
The contact tracing of sexual partners is relegated to those who have a positive test result, and since such individuals may have markedly different positions within a network, tracing a network consisting solely of STI-positive partners has an inherent bias. In our case, since the STD clinical reports don't contain complete sexual contact information, the original index case could not be identified. Questions about sexual risk factors and drug use were also omitted due to a lack of respondents (De 2001), and the lack of information on sexual practices e.g. condom usage, drug use could be considered a blind spot in this data.

Data Structure

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. These nodes with missing data were dropped on account of their missing information and it is thought that their exclusion is mostly inconsequential on account of their both having only one tie.

Methods



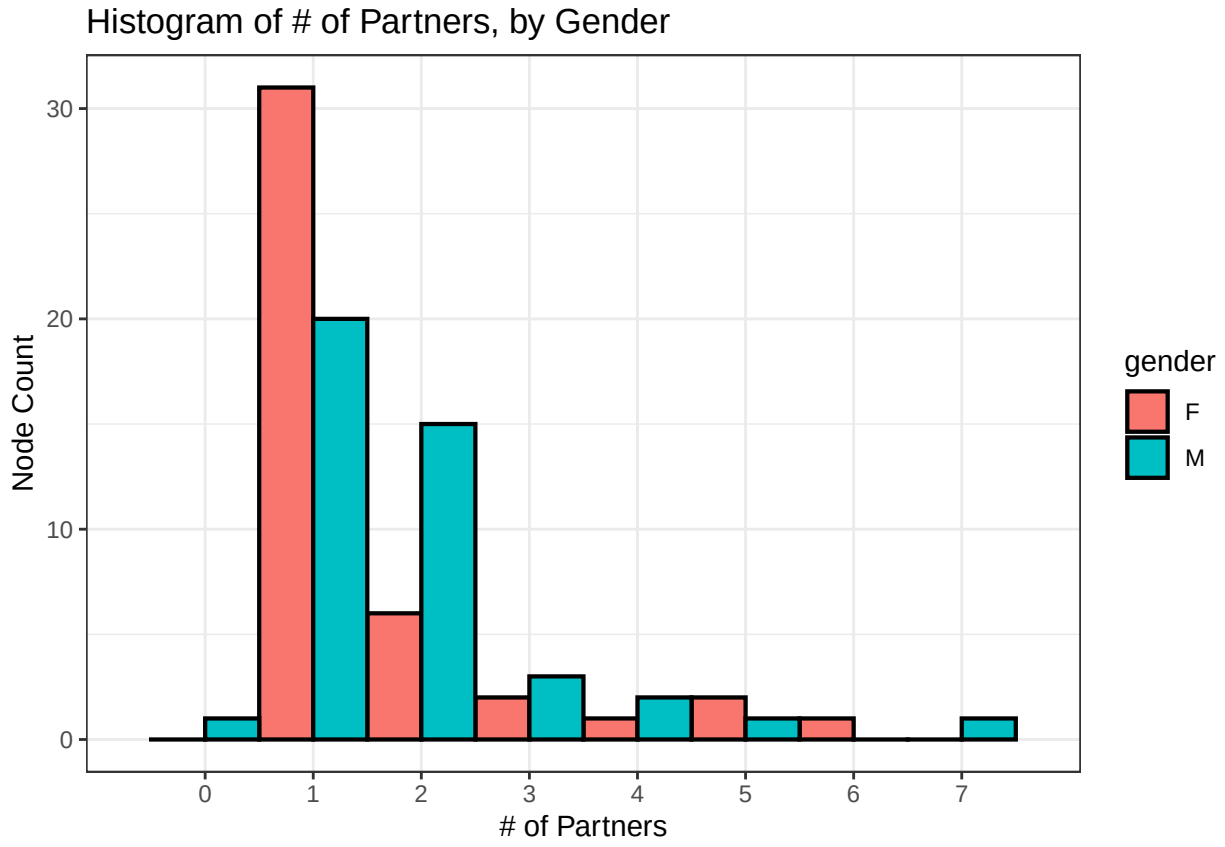
Looking at this initial sociogram, where the bar node is firmly in the center, we see that it has 17 ties to other individual nodes, which themselves have connections to at least one other node in the rest of the network. The remaining 74 edges in the dataset are between individuals. Of these 74 edges, approximately 90.5% of them are between male and female nodes, and approximately 48.6% of them are between bar-attendees and non-attendees.

While the majority of ties are between individuals of the opposite sex, e.g. male-to-female or female-to-male, 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.

Distribution of ties



Note that the distribution in number of partners is more right skewed for women than it is for men, owing to the fact that about a third of more women in this network had 1 partner compared to men who had 1 partner. Generally men's number of partners is more spread between 1-2 partners. The distribution is similar for both genders when looking at those

having more than 2 partners.

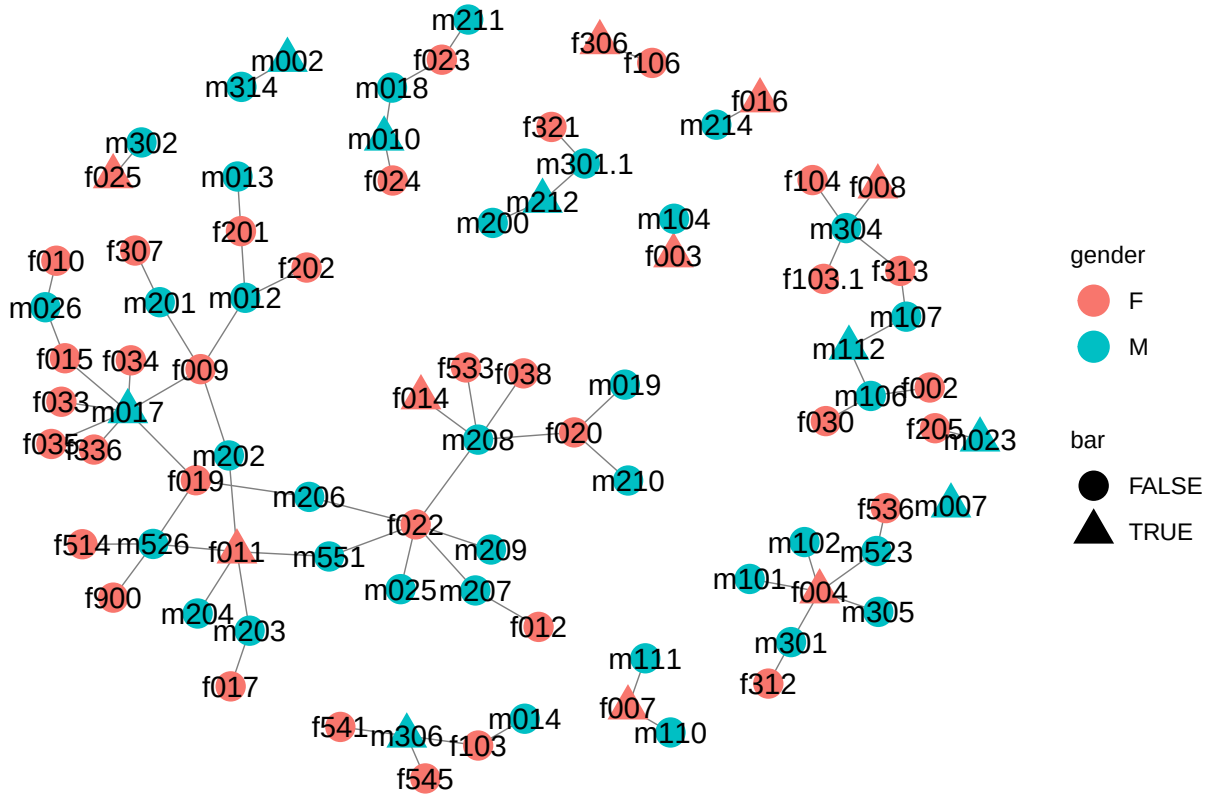
This histogram, which depicts a histogram of number of partners, colored by bar attendance, shows that a large number of non-bar patrons only have one partner; this is mostly bolstered by the fact that most of the members of the network do not frequent the bar, and also there are a sizeable number of individuals in the network who were simply alters of an ego who did attend the bar and do not have any other alters themselves.

Centrality Measures

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.

Network centrality was covered in De, at al (2004) but in the context of the connected network, and we want to expand on this by looking at centrality within the disconnected network (which does not contain the bar node).



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.

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 node i is equal to

Gender	Mean Number of Partners
F	1.605
M	1.837

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

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.

As noted earlier in the figure with the disconnected graph, 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 degree of 5, but all of its alters have a degree of only 1 or 2. In this literature "degree" will be used interchangably with "number of (sexual) partners" as these are the same thing in this context.

We observe from the data that on average, men have a slightly higher number of sexual partners than women in this sample. Note that we cannot compare the mean number of partners via a t-test in this case, because in this case, the two groups influence each others' degree.

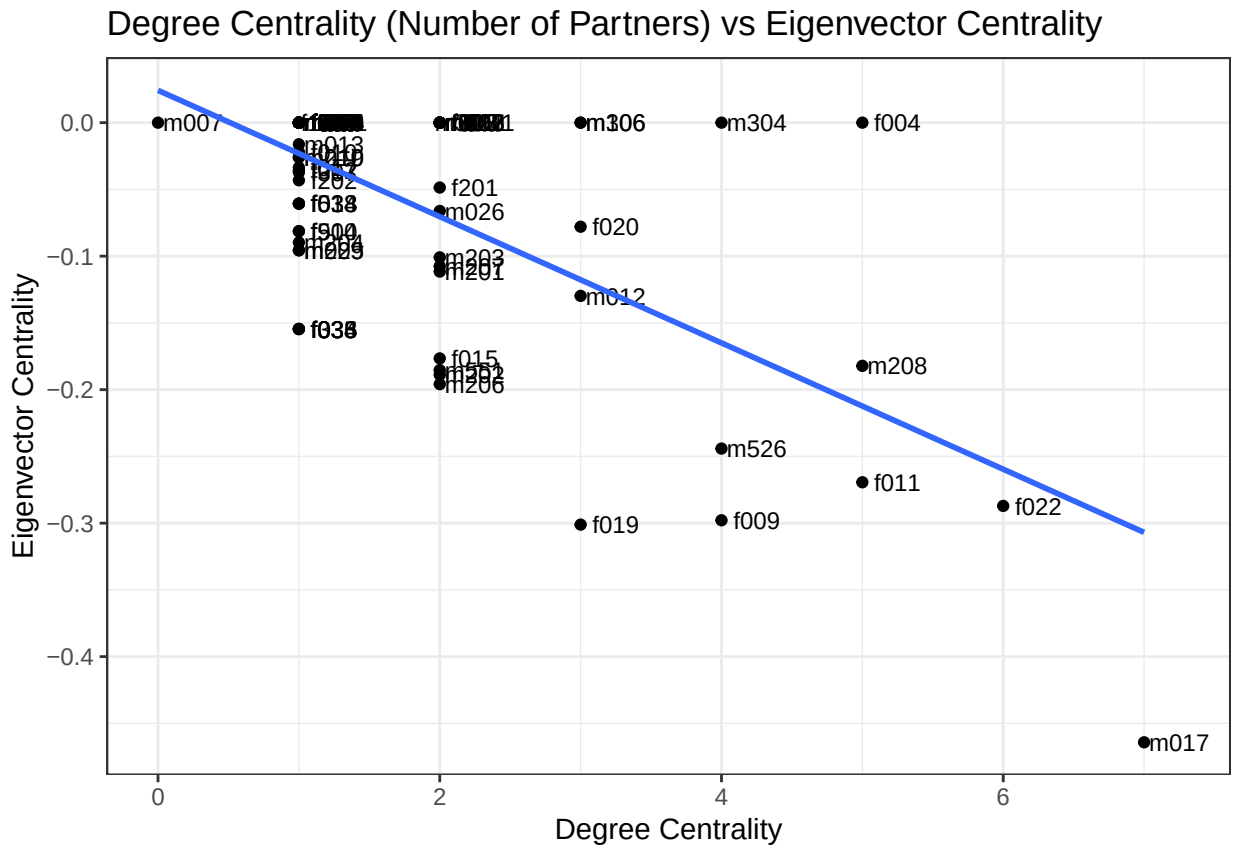
We further observe that those who attend the bar end up having an average of about 1.6 sexual partners compared to as many as 2.1 partners from those who did. Again, these two

Attended Bar	Mean Number of Partners
FALSE	1.623
TRUE	2.118

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.

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, an ego in this setting would have a high number of partners who themselves also have a high number of partners.

```
## `geom_smooth()` using formula 'y ~ x'
```



As we can see, all of the values of eigenvector centrality are at most 0, with the rest negative. There seems to be a negative correlation in general.

Closeness Centrality Closeness centrality is calculated as the reciprocal of the sum of the length of shortest paths between the node and all other nodes in the graph. In its normalized form, it is multiplied by the number of nodes in the graph N, as seen here

$$C(i) = \frac{1}{\sum_y d(j,i)} \times N$$

where $d(j,i)$ is the distance between nodes i and j.

Cutpoints and bridges

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.

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

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

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). Many of these models, although not all, were generated with Markov Chain Monte Carlo (MCMC) inference.

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, i.e.

$$\Pr(X_{t+1} = x | X_t = x_t)$$

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, which is used in the `ergm()` function to get paramter estimates. In

this algorithm, 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 ν, 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 ν 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. After some given number of samples (some of which may be 'thrown out' due to burn-in or thinning), we arrive at some estimate for the parameter, which in this case, would be the estimate for the coefficient η_A corresponding to the effect A in the ERGM, where A can be any of the many effects possible for an ERGM model e.g. homo/heterophily.

Model Fitting

Having delineated our techniques for estimating our parameters, we then opted to fit several different ERGM models. When fitting these models, 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.

$$AIC = -2\ln(\mathcal{L}) + 2k$$

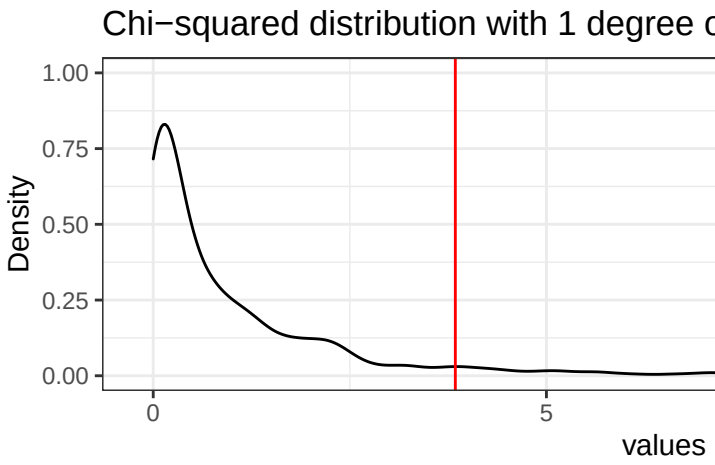
Where L is the likelihood of the model, and k is the number of parameters.

Here we have a table depicting the various AICs for each individual ERGM model. In model selection we start with the most simple intercept model, and then gradually add effects to it. In this case, we notice that among these effects, homophily based on attending the bar (`nodematch("bar")`), degree (`degree(d = c(1:3))`), and homo/heterophily for gender (`nodemix('gender')`) seemed to produce the greatest decreases in AIC. We then proceeded to fit models with these effects. One might notice that homophily based on bar attendance has something of a meager effect on decreasing the AIC. We can fit two models that both include and exclude this effect (a full and a nested model), and compare them with a likelihood ratio test.

```
set.seed(55)

ergm_model7 <- quietly(ergm)(gonnet_net_nobar ~
                      edges +
                      degree(d = c(1:3)) +
                      nodemix("gender", levels = c("M", "F")))$result


ergm_model8 <- quietly(ergm)(gonnet_net_nobar ~
                      edges +
                      degree(d = c(1:3)) +
                      nodemix("gender", levels = c("M", "F")) +
                      nodematch("bar"))$result
```



Likelihood Ratio Testing for Model Selection

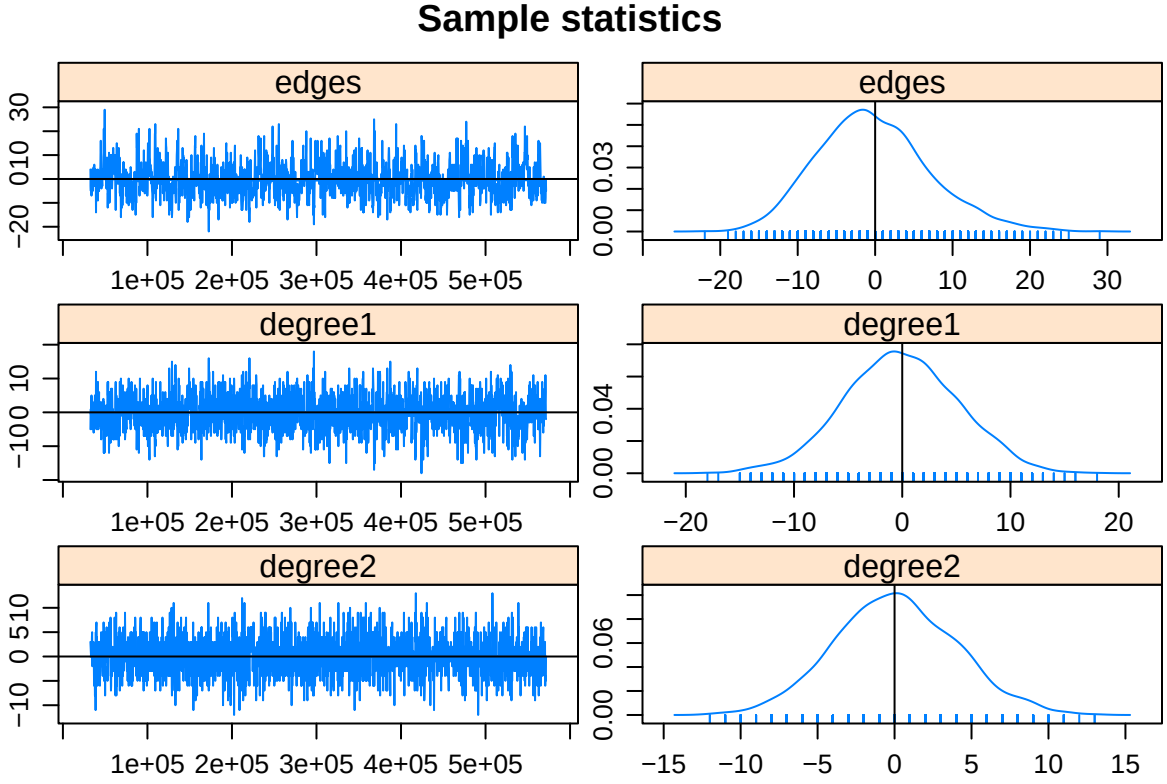
This curve represents the chi-squared distribution with 1 degree of freedom; which is the distribution of our likelihood ratio test statistic under the null hypothesis H_0, which states that the nested model fits the data as well as the full model, i.e. the nested model is preferred. Should we reject this null hypothesis, we would opt for the full model with the added effect instead. Since the full model only adds one more term (`nodematch`) to the nested model, the distribution of the test statistic under the null hypothesis has only 1 degree of freedom.

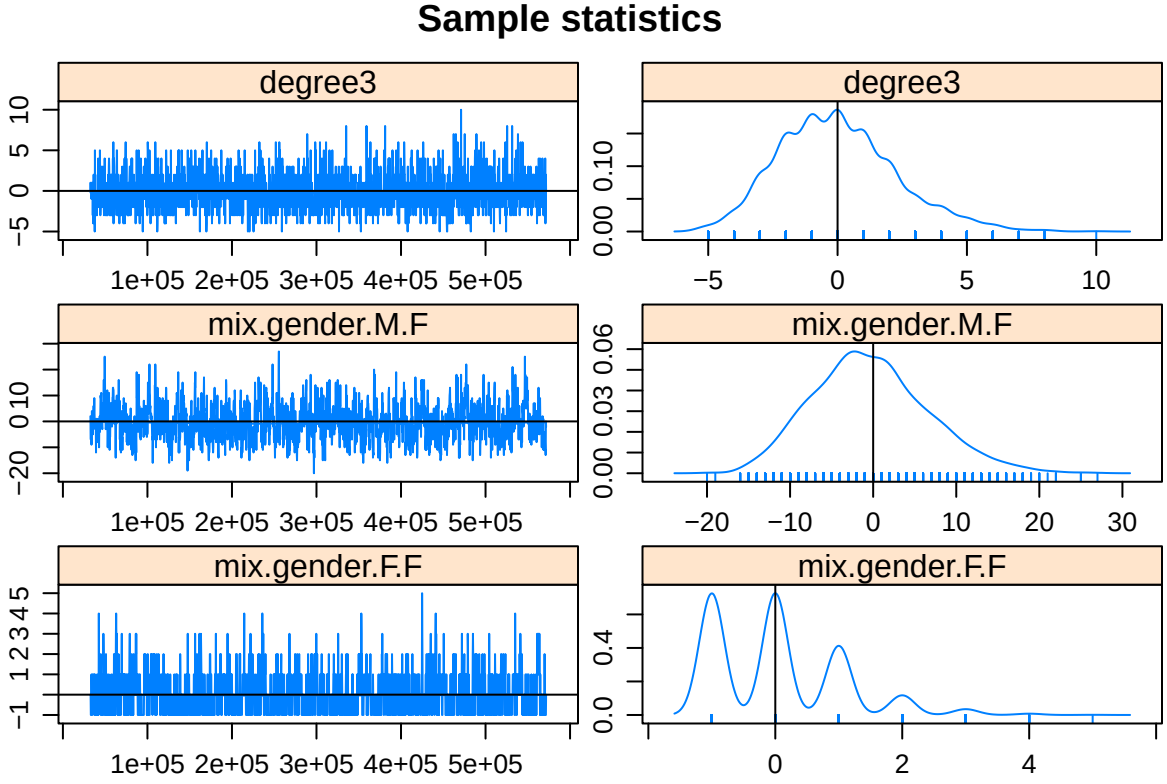
The red line in this curve represents the 95% critical value of the distribution, the point at which we would reject the null hypothesis. Note that our test statistic LRT, denoted by the dashed blue line, falls beyond this critical value, hence we have enough evidence at a 0.05 significance level to reject H_0.

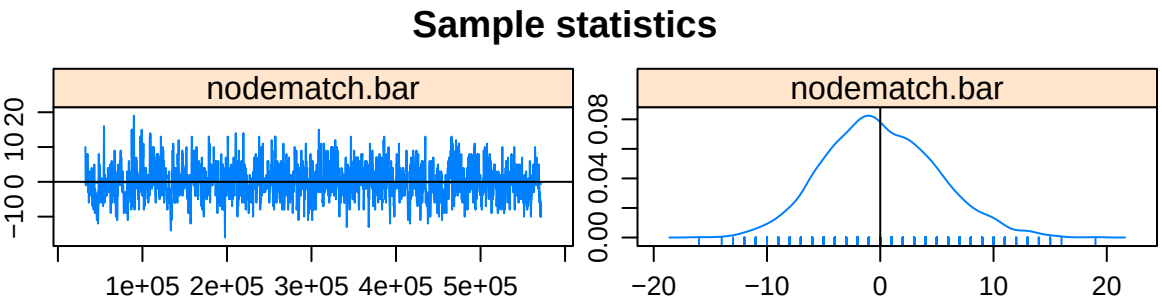
Thus, we conclude that the nested model does *not* fit the data as well as the full model, and opt to include the `nodematch` term for bar attendance in our final model.

Model 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.


Sample statistics
edges
edges
degree1
degree1
degree2
degree2
1e+05 2e+05 3e+05 4e+05 5e+05
−20 −10 0 10 20 30
0.00 0.03
−20 010 30
−100 10
0.00 0.04
−20 −10 0 10 20
−10 0 510
0.00 0.06
−15 −10 −5 0 5 10 15


Sample statistics
degree3
degree3
mix.gender.M.F
mix.gender.M.F
mix.gender.F.F
mix.gender.F.F
1e+05 2e+05 3e+05 4e+05 5e+05




```
##
## MCMC diagnostics shown here are from the last round of simulation, prior to computati
```

Interpretation

Recall the equation for the ERGM model

$$\log\left[\frac{\Pr(Y_{ij}=1|y_{ij}^{C})}{\Pr(Y_{ij}=0|y_{ij}^{C})}\right] = \sum_{A(Y_{ij})\eta_A d_A(y)}$$

Hence we have,

$$\log\left[\frac{\Pr(Y_{ij}=1|y_{ij}^{C})}{\Pr(Y_{ij}=0|y_{ij}^{C})}\right] = -5.1948+2.7430\text{Degree}_1+1.4678\text{Degree}_2+(3.4534\times\text{Gender}_{FM})+(1.7138\times\text{Gender}_{M}$$

```
##
## Call:
## .f(formula = ..1)
##
## Last MCMC sample of size 2107 based on:
##          edges         degree1         degree2         degree3  mix.gender.M.F
##       -3.44688         2.73464         1.46501         0.03515         1.69350
## mix.gender.F.F   nodematch.bar
##       -1.89839        -0.72321
##
## Monte Carlo Maximum Likelihood Coefficients:
##          edges         degree1         degree2         degree3  mix.gender.M.F
##       -3.43388         2.79904         1.50643         0.06844         1.71635
## mix.gender.F.F   nodematch.bar
##       -1.90945        -0.72712
```

Note that the MCMC estimates are relatively close to the estimates from Monte Carlo Maximum Likelihood Estimation (MCMLE), of which the `ergm()` function does both. Our interpretations will opt to utilze the estimates from the former.

Discussion

Our goal in this paper was to determine whether the attributes of gender and bar attendance were meaningful in the formation of ties in this sexual network, and quantifying these relationships using a model created under the exponential random graph model (ERGM) framework. We were able to create a model using Markov Chain Monte Carlo techniques, to achieve this.

In an epidemiological context, these results demonstrate that ERGMs can be applied to a sexual network, including one in which geographical information is a component. There are several different determinants to sexual network formation and a dataset that was more complete with this kind of information, be it demographical (e.g. ethnicity, race, religion, class), or geographical (e.g. city, state), would be able to apply these techniques in a more advanced and critically applicable way. It is crucial that epidemiologists and sociologists delve into techniques that can accurately model sexual behavior, as typical approaches to epidemic modeling, e.g. the SIR model (in which a population is placed into three components based on being either susceptible, infectious, or recovered), are short-sighted with regard their application in STI epidemiology, as their parameterizations make unrealistic assumptions such as "random mixing".

One of the critical advantages that network analysis has over traditional epidemiological approaches in a sexual network context is that it is at least somewhat robust to the missingness of data that is typical of egocentric network designs. While missingness of data can lead to issues such as computational intractibility when estimating parameters, network measures such as degree centrality are fairly robust to these issues (although this varies depending on the metric). Moreover, the flexibility that Markov Chain Monte Carlo brings to otherwise computationally intractable situations is also a benefit.

Techniques around intervention can also be devised based on the results of modeling an exponential random graph model. For example, consider our quantification of the effect that attending this local bar has in the area, on the dissemination of an STI in a community. This only quantified the effect of one location, but suppose the dataset contained information on *multiple* venues of contact, then we would be able to specifically model the effect each individual venue has on the probability of a tie forming, and hence, the odds of an outbreak taking place. From there, public health officials can then strategically determine which among these locations would be the best to actively intervene or extend outreach to. This is especially critical in the case of indigenous communities such as First Nations.

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 MBBCh, MSc, FRCPC‡ Outbreak of Neisseria gonorrhoeae in Northern Alberta, Canada, Sexually Transmitted Diseases: June 2003 - Volume 30 - Issue 6 - p 497-501

Valente TW. Social networks and health: models, methods, and applications. New York: Oxford University Press, 2010.

Public Health Agency of Canada. Report on sexually transmitted infections in Canada: 2011. Centre for Communicable Diseases and Infection Control, Infectious Disease Prevention and Control Branch, Public Health Agency of Canada; 2014. 2011. http://publications.gc.ca/site/eng/469949/publication.html.

Appendix

```
##                obs min  mean max MC p-value
## edges           74  60 74.24  92       0.92
## degree1         51  38 51.80  64       0.92
## degree2         21  11 20.36  30       0.90
## degree3          5   0  5.02  12       1.00
## mix.gender.M.F  67  54 67.22  83       1.00
## mix.gender.F.F   1   0  1.04   4       1.00
## nodematch.bar   38  25 37.40  49       0.94
```