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.

Methods

Results

Discussion

Conclusions

```
par(mfrow = c(1,1))
set.seed(10)
org_coord <-
  gplot(gonnet, vertex.col = gonnet_df$col,
      vertex.sides = gonnet_df$gender_lty,
      displaylabels = T, label.cex = 0.6,
      boxed.labels = F, pad = 2, usearrows = T,
      vertex.cex = 1.25)
# legend for gender
legend("topleft",
       legend = c("yes", "no"),
       col = c("tomato1", "grey"),
       fill = F, border = "white", pch = 19,
       title = "Bar Attendance", bty = "n")
legend("topright",
       legend = c("m", "f"),
       col = c("black"),
       fill = F, border = "white", pch = c(0, 2),
       title = "Gender", bty = "n")
# legend for bar attendance
```

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 outgoing 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.

The first approach in the analysis was to look at the *centrality measures* of each of the nodes.

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.

Among the centrality measures used in this analysis were that of *degree centrality*, *eigenvector centrality*, *load centrality*, and *information centrality*.

Degree Centrality

Degree centrality can be measured in multiple ways; the first is *indegree* which is a count of the number of incoming ties that a node has. The second is *outdegree* which conversely is the count of the number of outgoing ties that a node has.

The vast majority of nodes in the directed network have an indegree of 1, i.e. only one individual nominated that person as a sexual partner when interviewed. The outdegree was *at least* 1 for most nodes in the network, but could be zero for nodes on the very outside of the network who did not name any sexual partners.

The overall degree of a node can be taken as the total amount of ties a node has, either incoming or outgoing, but in the case of an undirected graph, these are one and the same.

```r
gonnet_df2 <- cbind(gonnet_df, cen) %>% select(-id)

# boxplot comparing outdegree
outdegree_gender <- gonnet_df2 %>%
  select(nodes, gender, outdegree) %>%
  drop_na()

# boxplot of outdegree by gender
ggplot(outdegree_gender, aes(x=gender, y=outdegree)) +
  geom_boxplot()
```

Outdegree Centrality The analysis gave that the overall average outdegree 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 outdegree 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 outdegree 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.

```r
# Student's t-test comparing the mean outdegree of men and women
# H0: mu_x - mu_y = 0
t.test(x = outdegree_m,
       y = outdegree_f,
       var.equal = T, alternative = "two.sided")
# p-value = 0.5232 > 0.05, we fail to reject H0
```

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.

Load Centrality

Information Centrality

```r
# boxplot comparing distribution of outdegree between bar (not)-attended
ggplot(gonnet_df2 %>% filter(!(nodes %in% c("b", "x", "x2"))),
       aes(x = attended_bar, outdegree)) +
  geom_boxplot() + theme_bw()

# average outdegree among bar attendees
outdegree_bar <- gonnet_df2 %>%
  select(nodes, attended_bar, outdegree) %>%
  filter(!(nodes %in% c("b", "x2", "x"))) %>%
  filter(attended_bar == TRUE)

# average outdegree among non-bar attendees
outdegree_nobar <- gonnet_df2 %>%
  select(nodes, attended_bar, outdegree) %>%
  filter(!(nodes %in% c("b", "x2", "x"))) %>%
  filter(attended_bar == FALSE)

# sample sizes are unequal, so we cannot assume equal variance
# try two-sided Welch's t.test
# H0: mu_x - mu_y = 0
t.test(x = outdegree_bar$outdegree,
       y = outdegree_nobar$outdegree,
       var.equal = FALSE, alternative = "two.sided")
# p-value < 0.05, reject H0
# conclude the difference in mean outdegree is not equal to zero
```

Principal Component Analysis

```r
# extracting the centralities that were important based on the PCA
centrality_eigen <- centralities$`eigenvector centralities`
centrality_load <- centralities$`Load Centrality`
centrality_degree <- centralities$`Degree Centrality`
centrality_geodesic <- centralities$`Geodesic K-Path Centrality`
centrality_shortest <- centralities$`Shortest-Paths Betweenness Centrality`
centrality_info <- centralities$`Information Centrality`
```

```r
km <- kmeans(gonnet, centers = 21, nstart = 25)

# fviz_cluster(km, data = gonnet_nob)

# library(netdiffuseR) is loaded
gonnet_edgelist <- adjmat_to_edgelist(gonnet, undirected = F)

# cluster_membership
km_cluster_mem <- km$cluster %>% as.data.frame() %>%
  tibble::rownames_to_column() %>%
  rename(node = 'rowname', cluster = '.')
```

```r
fit <-
  pvclust(gonnet,
  method.hclust = "single",
  method.dist = "euclidean",
  iseed = 10, # to get same results
  parallel = T, # to use all but one CPU thread
  nboot = 1000)
```

```
fit_hclust <- fit$hclust
fit_hclust %>% cutreeDynamicTree(deepSplit = F)
```

```
par(mfrow = c(1,1))
set.seed(10)
edgelist <- as.edgelist(gonnet, n = dim(gonnet)[1])
plot_kcores(edgelist, sym = F, mode = "digraph",
            coord = org_coord,
            cmode = "outdegree")
```

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.

```r
# library(ergm) is loaded
edges <- rep(1, 89)
ergm_model <- ergm(gonnet ~ edges)
summary(ergm_model)

exp1 <- coef(ergm_model) %>% exp()

exp1 / (1 + exp1)

# nodes that went to the bar
bar_attendees <-
  gonnet["b",] %>% as.data.frame() %>%
  dplyr::rename(bar = ".") %>%
  filter(bar == 1) %>% row.names()

ergm_model2 <- ergm(gonnet ~ edges + nodematch(attr = ""))
summary(ergm_model2)
```

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