Lab 2

Arsalan Khan

3/6/2021

Contents

1.

Describe the social network(s) to me, in terms of how it was collected, what it represents and so forth. Also give me basic topography of the network: the nature of the ties; direction of ties; overall density; and if attributes are with the network, the distribution of the categories and variables of those attributes.

This social network is about college students in an undergraduate course on social networks. Data on ties between the students was collected in four waves at different intervals over the 11 weeks of the course. Attribute data about the nodes in the network was also collected by having the students report it themselves and by using the instructor's records on the students. There are a total of 75 nodes/students in the network. The student's were asked to pick from a list of all the other students in the class who they would consider close enough to be able to ask for a small favor like borrowing class notes. Ties in the network are measured as asymmetric and binary.

```
edge_density(classroom_net_graph)
```

```
## [1] 0.1612613
```

We see that the overall density of this network is about 0.16.

```
table(classroom_att_data$Gender)
```

```
##
##  2  1
## 48 27
```

We see that the class had more women than men.

```
table(factor(classroom_att_data$Ethnicity, levels = 1:4, labels = c('White', 'Hispanic', 'Asian', 'Afri
```

```
##
##                        White                     Hispanic                        Asian
##                           17                           20                           32
## African American or Other
##                            6
```

Above we see the breakdown of the nodes based on ethnicity.

2. Calculate degree centrality (in- and out-degree, too, if you have such data); closeness centrality; betweenness centrality; and eigenvector centrality. Correlate those measures of centrality. Highlight which nodes are most central and least central, along different dimensions.

```
classroom_att_data <- merge(classroom_att_data,  #  Merge classroom_data
                            data.frame(  # With a new data.frame
                              ID = classroom_att_data$ID,
                              in.deg= degree(classroom_net_graph, mode = c("in"), loops = FALSE, normali
                              out.deg= degree(classroom_net_graph, mode = c("out"), loops = FALSE, norma
                              btwn= betweenness(classroom_net_graph, directed = F),
                              close = closeness(classroom_net_graph, mode = c("all")),
                              eigen <- evcent(classroom_net_graph),
                              bon <- bonpow(classroom_net_graph)
                            ),
                            by = 'ID')
```

```
## Warning in closeness(classroom_net_graph, mode = c("all")): At centrality.c:
## 2784 :closeness centrality is not well-defined for disconnected graphs
```

```
classroom_att_data <- classroom_att_data[,c(1:17, 39)]

names(classroom_att_data)[names(classroom_att_data)=="bon....bonpow.classroom_net_graph."] <- "bon"
```

Using in-degree centrality measure.

```
head(classroom_att_data %>% arrange(desc(in.deg))) %>% select(c(ID:in.deg))
```

```
##   ID Ethnicity Gender Group Attend1 Attend2 Attend3 E1 E2 E3 Partic Paper
## 1 MA         4      2     6     100     100      75 60 56 74    100    87
## 2 CM         2      1     8     100     100     100 70 77 67     92    83
## 3 FD         1      1     3      75     100     100 78 62 73    100    93
## 4 LC         2      1     1     100     100     100 83 69 80     89    80
## 5 RJ         2      1     4     100     100     100 56 76 73     96    72
## 6 SA         1      2     8      75     100     100 80 63 68     92    83
##   in.deg
## 1     22
## 2     20
## 3     20
## 4     20
## 5     20
## 6     19
```

Using out.deg centrality measure.

```
head(classroom_att_data %>% arrange(desc(out.deg)) %>% select(c(ID:Paper, out.deg)))
```

```
##   ID Ethnicity Gender Group Attend1 Attend2 Attend3 E1 E2 E3 Partic Paper
## 1 MA         4      2     6     100     100      75 60 56 74    100    87
## 2 CM         2      1     8     100     100     100 70 77 67     92    83
## 3 FD         1      1     3      75     100     100 78 62 73    100    93
## 4 LC         2      1     1     100     100     100 83 69 80     89    80
## 5 RJ         2      1     4     100     100     100 56 76 73     96    72
## 6 SA         1      2     8      75     100     100 80 63 68     92    83
##   out.deg
## 1      22
## 2      20
## 3      20
## 4      20
## 5      20
## 6      19
```

3a. If you have a network with attribute data, then state some hypothesis about how an attribute may be related to some (or all of the) measures of centrality. Explains why you think these two variables should be related.

My first hypothesis is that the variation in a student's score on the last exam of the class can be explained by their out.degree measure. Also if a student has a high out.deg score then it's likely that they will have a high score on the last exam since they'll be more actively seeking out help from others in the class. The reason for using the last exam score as a dependent variable is that by that time during the semester students network positions have somewhat stabilized and are thus a more accurate representation of their final network position.

```
summary(lm(E3 ~ out.deg, data = classroom_att_data))
```

```
##
```

```
## Call:
## lm(formula = E3 ~ out.deg, data = classroom_att_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -65.131  -7.175   1.380   8.314  28.912
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  61.0436     4.1025  14.880   <2e-16 ***
## out.deg       0.5110     0.3224   1.585    0.117
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.2 on 73 degrees of freedom
## Multiple R-squared:  0.03327,    Adjusted R-squared:  0.02002
## F-statistic: 2.512 on 1 and 73 DF,  p-value: 0.1173
```

From this simple regression model we see that there is indeed a positive relationship between out.degree centrality and a student's performance on the last exam. However the coefficient on out.degree centrality is not statistically significant and so we can't be sure of the this effect on the exam score.

Below is a model that regresses students exams scores on the last exam against all measures of centrality.

```
summary(lm(E3 ~ out.deg + in.deg + btwn + close + vector + bon, data = classroom_att_data))
```

```
##
## Call:
## lm(formula = E3 ~ out.deg + in.deg + btwn + close + vector +
##     bon, data = classroom_att_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.390  -5.721   0.920   7.470  28.368
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   63.28107   13.61209   4.649 1.58e-05 ***
## out.deg        0.09242    1.38872   0.067    0.947
## in.deg         0.13144    1.40959   0.093    0.926
## btwn           0.11502    0.12918   0.890    0.376
## close         61.69037 3841.35921   0.016    0.987
## vector        -6.93365   36.75680  -0.189    0.851
## bon           -0.67970    1.76161  -0.386    0.701
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.5 on 68 degrees of freedom
## Multiple R-squared:  0.05925,    Adjusted R-squared:  -0.02376
## F-statistic: 0.7138 on 6 and 68 DF,  p-value: 0.6397
```

We see that none of the centrality measures seem to be statistically significant at any reasonable level of significance which is a bit strange.

4. In either case, when you are done above, then consider alternate specifications of your variables and codings and decisions and models. What would you want to consider changing and why. If you can, report on what are the consequences of those changes?

Perhaps it might make more sense to include other control variables such as a student's gender and their ethnicity. We can then study the effects of different centrality measures net of their gender and ethnicity.

```
summary(lm(E3 ~ out.deg + in.deg + btwn + close + vector + bon + Ethnicity + Gender, data = classroom_a
```

```
##
## Call:
## lm(formula = E3 ~ out.deg + in.deg + btwn + close + vector +
##     bon + Ethnicity + Gender, data = classroom_att_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.923  -5.043   1.109   6.968  21.790
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   61.04598   14.22288   4.292 6.11e-05 ***
## out.deg        0.62270    1.32226   0.471   0.6393
## in.deg         0.32998    1.36092   0.242   0.8092
## btwn           0.01428    0.12702   0.112   0.9108
## close       1247.99107 3766.25387   0.331   0.7415
## vector       -17.33542   35.73151  -0.485   0.6292
## bon            0.27035    1.68493   0.160   0.8730
## Ethnicity2    -3.79696    4.45126  -0.853   0.3968
## Ethnicity3    -7.36955    4.34195  -1.697   0.0945 .
## Ethnicity4   -13.59129    6.48320  -2.096   0.0400 *
## Gender1        7.66919    3.39666   2.258   0.0274 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.65 on 64 degrees of freedom
## Multiple R-squared:  0.2218, Adjusted R-squared:  0.1002
## F-statistic: 1.824 on 10 and 64 DF,  p-value: 0.07395
```

From the above model summary we see that if a student is a male there exam score goes up by about 7.7 points. Also being any race other than white has a negative impact on one's exam scores.

It might make more sense to aggregate a student's score across all three exams and use that as a dependent variable.

```
classroom_att_data <- cbind(classroom_att_data, "aggregate" = rowMeans(cbind(classroom_att_data$E1, clas

summary(lm(aggregate ~ out.deg + in.deg + btwn + close + vector + bon + Ethnicity + Gender, data = class
```

```
##
## Call:
## lm(formula = aggregate ~ out.deg + in.deg + btwn + close + vector +
##     bon + Ethnicity + Gender, data = classroom_att_data)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.722  -4.450   1.080   5.022  24.577
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.920e+01  1.103e+01   6.275 3.42e-08 ***
## out.deg     -1.490e-01  1.025e+00  -0.145   0.8849
## in.deg       1.186e-01  1.055e+00   0.112   0.9109
## btwn         5.518e-03  9.849e-02   0.056   0.9555
## close        7.471e+02  2.920e+03   0.256   0.7989
## vector      -3.964e+00  2.771e+01  -0.143   0.8867
## bon         -9.878e-01  1.307e+00  -0.756   0.4524
## Ethnicity2  -2.100e+00  3.452e+00  -0.608   0.5450
## Ethnicity3  -3.722e+00  3.367e+00  -1.106   0.2731
## Ethnicity4  -9.491e+00  5.027e+00  -1.888   0.0636 .
## Gender1      5.638e+00  2.634e+00   2.141   0.0361 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.811 on 64 degrees of freedom
## Multiple R-squared:  0.1774, Adjusted R-squared:  0.04891
## F-statistic: 1.381 on 10 and 64 DF,  p-value: 0.2096
```

We see that using an aggregate score doesn't really impact the coefficients in any significant way. Suggesting that perhaps the effect of network position is seen more on the last exam than the first 2.

5. Lastly, give your best conclusion as to what you learned from your analysis. Did it make sense, given your initial expectations? Why? Why not.

My initial expectation was that a central position in one's ego network might have a very strong impact on a student's performance across the semester. While there is a positive impact of network centrality on student performance it doesn't seem to be too important for determining a student's performance (at least in this class).