Social Network Analysis: Lab 1

Mikael Brunila

2017-02-20

Contents

Introduction

```
relationships <- read_csv("SocialEvolution/RelationshipsFromSurveys.csv", col_names = TRUE)
health <- read_csv("SocialEvolution/Health.csv", col_names = TRUE)
flu <- read_csv("SocialEvolution/FluSymptoms.csv", col_names = TRUE)
```

For my first lab, I decided to use the Social Evolution dataset from the MIT Media Lab Reality Commons. The data was gathered during an ambitious experiment,

> to closely track the everyday life of a whole undergraduate dormitory with mobile phones, so that social scientists can validate their models against the spatio-temporal patterns and behavior-network co-evolution as contained in this data. The Social Evolution experiment covered the locations, proximities, and phone calls of more than 80% of residents who lived in the dormitory used in the Social Evolution experiment, as captured by their cell phones from October 2008 to May 2009.

The dormitory had a population of approximately 30 freshmen, 20 sophomores, 10 juniors, 10 seniors and 10 graduate student tutors. The survey data for the study was gathered on a weekly, monthly or even daily basis to help facilitate an understanding of how different variables were adapted. Out of these, I will focus on the ones that describe depression and stress among the students. Because a timeseries analysis if beyond the scope of this lab, I used different methods to capture an average over time, to be detailed below.

1&2) Hypothesis and explanation

Develop a hypothesis about how some ego-network measure (e.g., degree/size, density, diversity, average-level of alters, homophily, structural holes, or brokerage) may be related to some other variable of interest.

My first research hypothesis is the following:

> The in-degree of the ego-network of a student is correlated to her level of depression in a positive negative way.

Explain why you think these two variables should be related.

Degree is a measure of the amount of connections a node has to other nodes. In a directed graph, in-degree shows how many alters are connected to the ego (but not necessarily the other way around). As such, degree offers in this case a measure of how socially connected someone in the dormitory is. A student with a high in-degree, has been referenced by many other students as a person they have had contact with.

3) Variables

Tell me about your variables. What is your dependent variable? What are your independent variables? How are they coded? How are they recoded? How are the calculated, if appropriate?

In order to perform my analysis, I had to recode my data a number of ways. The Social Evolution dataset offered five different measures of closeness:

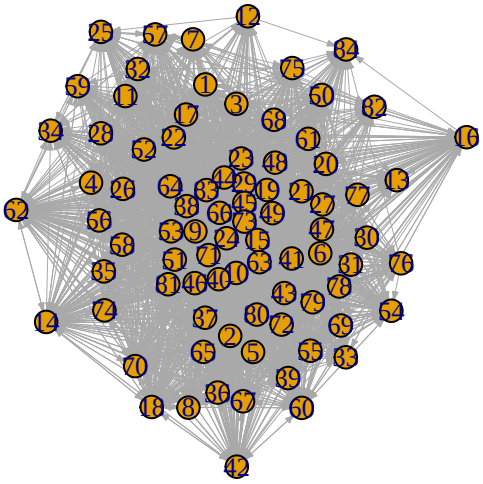
- CloseFriend,
- SocializeTwicePerWeek,
- PoliticalDiscussant,
- FacebookAllTaggedPhotos,
- BlogLiveJournalTwitter

I assumed these were ordered and recoded them as numbers, to then get a mean over time of the different ties that students reported in the surveys. I then used this to produce a graph object and visualization of the relationships in the dataset. In addition, I used only the CloseFriend value to get a graph of all the friendships that were reported in the data.

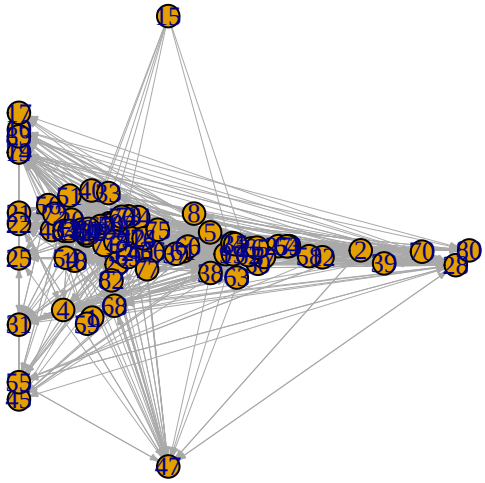
In both cases, I noticed that some ties were reported in an odd fashion, with a node being connected to itself. I assumed these were coding errors and removed them from the data.

```
# Courtesy of http://realitycommons.media.mit.edu/socialevolution4.html
relationships$id.A <- factor(relationships$id.A, levels = 1:84)
relationships$id.B <- factor(relationships$id.B, levels = 1:84)
relationships$relationship <- factor(relationships$relationship,
                                     levels =
                                       c("SocializeTwicePerWeek",
                                         "CloseFriend",
                                         "PoliticalDiscussant",
                                         "FacebookAllTaggedPhotos",
                                         "BlogLivejournalTwitter"))

# Own code again
namevector <- c("value")
relationships[ , namevector] <- NA

# Recode factors as numbers
for(row in 1:nrow(relationships)) {

  if(relationships[row, 3] == "CloseFriend"){
    relationships[row, 5] <- 5
  }
  if(relationships[row, 3] == "SocializeTwicePerWeek"){
    relationships[row, 5] <- 4
  }
  if(relationships[row, 3] == "PoliticalDiscussant"){
    relationships[row, 5] <- 3
  }
  if(relationships[row, 3] == "FacebookAllTaggedPhotos"){
    relationships[row, 5] <- 2
  }
  if(relationships[row, 3] == "BlogLivejournalTwitter"){
    relationships[row, 5] <- 1
  }
```

```
}

# Averaging all ties over time
relationships_weighted <- relationships %>%
  group_by(id.A, id.B, survey.date) %>%
  tally(wt = value) %>%
  group_by(id.A, id.B) %>%
  summarise(mean = mean(n))

relationships_weighted$mean <- round(as.double(relationships_weighted$mean), 1)

for(row in 1:nrow(relationships_weighted)) {
  if(as.integer(relationships_weighted[row, 1]) ==
     as.integer(relationships_weighted[row, 2])) {
    relationships_weighted[row, 3] <- "0"
  }
}

row_sub <- apply(relationships_weighted, 1, function(row) all(row != "0"))
relationships_weighted <- relationships_weighted[row_sub,]

# Averaging CloseFriends ties over time
friendships <- relationships[relationships$relationship == "CloseFriend", ]
friendships <- friendships %>%
  group_by(id.A, id.B, survey.date) %>%
  tally(wt = value) %>%
  group_by(id.A, id.B) %>%
  summarise(value = mean(n))

for(row in 1:nrow(friendships)) {
  if(friendships[row, 1] == friendships[row, 2]) {
    friendships[row, 3] <- 0
  }
}

row_sub_f <- apply(friendships, 1, function(row) all(row != 0))
friendships <- friendships[row_sub_f,]

# Graph of all ties averaged over time
g <- graph.data.frame(relationships_weighted)

e <- get.edgelist(g)
e <- cbind(as.integer(e[,1]), as.integer(e[,2]))
l <- qgraph.layout.fruchtermanreingold(e, vcount = gorder(g))
V(g)$label.cex = 0.8
plot(g, layout = l, edge.width = 0.1, edge.arrow.size=0.2, vertex.size = 10)
```



```
# Graph of ties formed by CloseFriends averaged over time
g_friendships <- graph.data.frame(friendships)

e_friendships <- get.edgelist(g_friendships)
e_friendships <- cbind(as.integer(e_friendships[,1]), as.integer(e_friendships[,2]))
l <- qgraph.layout.fruchtermanreingold(e_friendships, vcount = gorder(g_friendships))
V(g_friendships)$label.cex = 0.8
plot(g_friendships, layout = l, edge.width = 0.1, edge.arrow.size=0.2, vertex.size = 10)
```



In addition to the actual vertices, I calculated a number of attributes, using both network measures and variables about mental health. The latter were stored in a file on flu syndroms, and measured depression and stress as a binary (0 or 1) dummy variable. I measured the mean for each respondent over time for both of these. For network measures I calculated indegree and density, but ended up using only the former.

I did both operations on both graph objects.

```
g_attributes <- flu %>%
  group_by(user_id) %>%
  summarise(depression = mean(as.double(sad.depressed)), stress = mean(as.double(open.stressed)))

# Idea in the code below are implemented based on class slides, adjusted for my data
V(g)$depression <- g_attributes$depression[match(V(g)$name, as.character(g_attributes$user_id))]
V(g)$stress <- g_attributes$stress[match(V(g)$name, as.character(g_attributes$user_id))]
V(g)$degree <- igraph::degree(g, mode = c("in"), normalized = TRUE)
```

```r
dens <- data.frame(transitivity = transitivity(g, type="local"))

all <- relationships_weighted

all <- merge(all, g_attributes, by.x = c("id.A"), by.y = c("user_id"))
all <- select(all, c(ego = id.A,
                     alter = id.B,
                     tie_strength = mean,
                     depression_ego = depression,
                     stress_ego = stress))
all <- merge(all, g_attributes, by.x = c("alter"), by.y = c("user_id"))
all <- select(all, c(ego, alter, tie_strength, depression_ego, stress_ego,
                     depression_alter = depression, stress_alter = stress))

dd <- data.frame(ID1 = V(g)$name)

cb1 <- cbind(dd, V(g)$degree)
cb2 <- cbind(dd, dens)

all <- merge(all, cb1, by.x = "ego", by.y = c("ID1"))
all <- merge(all, cb2, by.x = "ego", by.y = c("ID1"))
colnames(all)[8] <- "degree"
colnames(all)[9] <- "density"
```

```r
# Idea in the code below are implemented based on class slides, adjusted for my data
V(g_friendships)$depression <- g_attributes$depression[match(V(g)$name, as.character(g_attributes$user_
V(g_friendships)$stress <- g_attributes$stress[match(V(g)$name, as.character(g_attributes$user_id))]
V(g_friendships)$degree <- igraph::degree(g_friendships, mode = c("in"), normalized = TRUE)
dens <- data.frame(transitivity = transitivity(g_friendships, type="local"))

all_friendships <- friendships

all_friendships <- merge(all_friendships, g_attributes, by.x = c("id.A"), by.y = c("user_id"))
all_friendships <- select(all_friendships, c(ego = id.A,
                     alter = id.B,
                     tie_strength = value,
                     depression_ego = depression,
                     stress_ego = stress))
all_friendships <- merge(all_friendships, g_attributes, by.x = c("alter"), by.y = c("user_id"))
all_friendships <- select(all_friendships,
                          c(ego, alter, tie_strength, depression_ego, stress_ego,
                            depression_alter = depression, stress_alter = stress))

dd <- data.frame(ID1 = V(g_friendships)$name)

cb1 <- cbind(dd, V(g_friendships)$degree)
cb2 <- cbind(dd, dens)

all_friendships <- merge(all_friendships, cb1, by.x = "ego", by.y = c("ID1"))
all_friendships <- merge(all_friendships, cb2, by.x = "ego", by.y = c("ID1"))
colnames(all_friendships)[8] <- "degree"
colnames(all_friendships)[9] <- "density"
```

4) Initial results

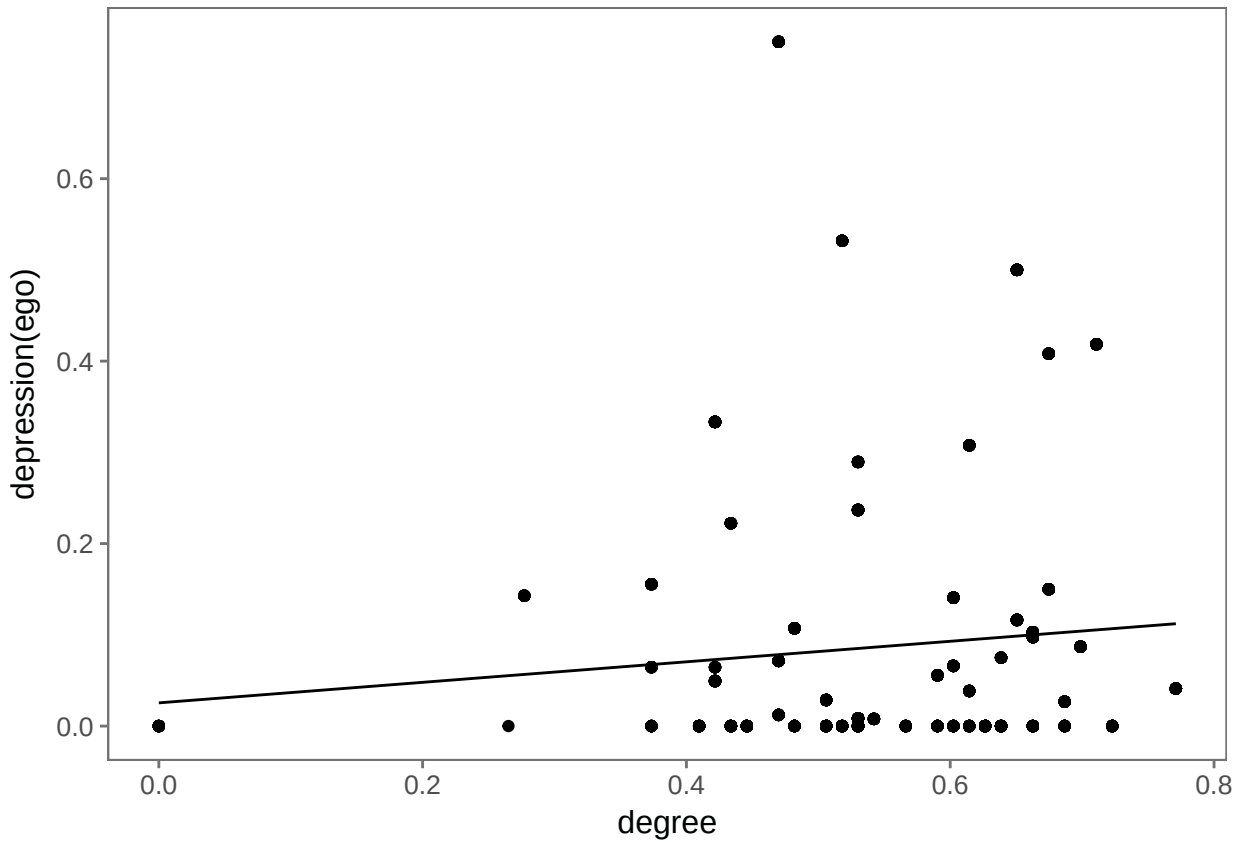
Present your initial results from your first few models. What do they indicated about your hypothesis?

The results from my first model were statistically significant, but contrary to my hypothesis. Indegree is indeed correlated to depression, but in a positive manner. The more people say they are connected to you, the more depressed you are! The coefficient for depression was 0.07.

```
model1 <- lm(depression_ego ~ degree, all)
summary(model1)
```

```
##
## Call:
## lm(formula = depression_ego ~ degree, data = all)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.10675 -0.08506 -0.07107  0.01767  0.67172
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.02542    0.01427   1.782   0.0749 .
## degree       0.11250    0.02487   4.524 6.37e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1406 on 2400 degrees of freedom
## Multiple R-squared:  0.008455,   Adjusted R-squared:  0.008042
## F-statistic: 20.46 on 1 and 2400 DF,  p-value: 6.37e-06
```

```
ggplot(all, aes(degree, depression_ego)) +
  geom_point() +
  geom_line(aes(y = fitted(model1))) +
  labs(y = "depression(ego)", x = "degree") +
  theme_few()
```



5) Interactions

Consider alternate specifications of your variables (i.e., recodings of various kinds). Consider interactions among your variables.

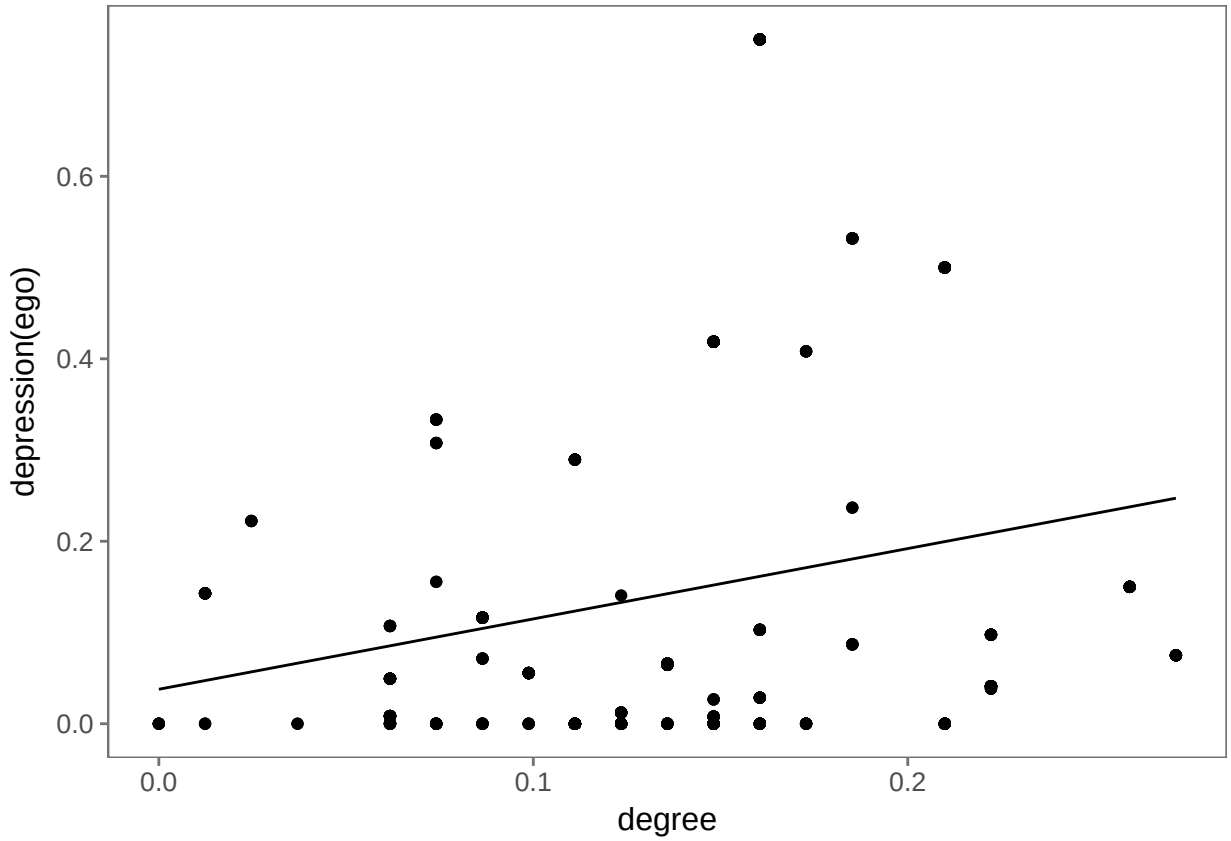
Moving on from my first model, I wanted to see whether fitting a similar model on the friendship graph would produce similar results. My thinking here was that a high in-degree might be related to stress and social pressures. Maybe ties that were based on friendship would give a different result, as they capture a different type of relationship than the other measures of connection? The reults contradicted this expectation, as it was in line with the results from my first model. Again, indegree was correlated to depression, and this time with a very high (0.77) and statistically significant estimate.

```
model1f <- lm(depression_ego ~ degree, all_friendships)
summary(model1f)
```

```
##
## Call:
## lm(formula = depression_ego ~ degree, data = all_friendships)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.19960 -0.13296 -0.08538  0.09555  0.58848
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.03778    0.02050   1.843   0.0658 .
## degree       0.77098    0.13835   5.573 3.89e-08 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.194 on 564 degrees of freedom
## Multiple R-squared:  0.05219,    Adjusted R-squared:  0.05051
## F-statistic: 31.06 on 1 and 564 DF,  p-value: 3.887e-08
```

```
ggplot(all_friendships, aes(degree, depression_ego)) +
  geom_point() +
  geom_line(aes(y = fitted(model1f))) +
  labs(y = "depression(ego)", x = "degree") +
  theme_few()
```



I also ran two models with density as a control variable and interaction variable and the results were among similar lines to my initial model, with indegree remaininsh a statistifically significant predictor for depression.

```
model_control1 <- lm(depression_ego ~ degree + density, all)
summary(model_control1)
```

```
##
## Call:
## lm(formula = depression_ego ~ degree + density, data = all)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.10673 -0.08653 -0.06739  0.01886  0.67035
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -0.02112    0.03531  -0.598     0.55
## degree       0.15407    0.03808   4.046 5.39e-05 ***
## density      0.07058    0.04898   1.441     0.15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1406 on 2399 degrees of freedom
## Multiple R-squared:  0.009312,   Adjusted R-squared:  0.008486
## F-statistic: 11.28 on 2 and 2399 DF,  p-value: 1.337e-05
```

```
model_interactions1 <- lm(depression_ego ~degree*density, all)
summary(model_interactions1)
```

```
##
## Call:
## lm(formula = depression_ego ~ degree * density, data = all)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.10726 -0.08649 -0.06756  0.01916  0.66996
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.01878    0.03734  -0.503   0.6151
## degree          0.14701    0.05279   2.785   0.0054 **
## density         0.06320    0.06214   1.017   0.3092
## degree:density  0.02291    0.11860   0.193   0.8468
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1406 on 2398 degrees of freedom
## Multiple R-squared:  0.009328,   Adjusted R-squared:  0.008088
## F-statistic: 7.526 on 3 and 2398 DF,  p-value: 5.188e-05
```

Finally, I wanted to see if degree works as a predictor for stress, which it does. This is easier to make sense of: a person who is very busy socially, might also be more stressed than a more reclusive student.

```
model_interactions1 <- lm(stress_ego ~ degree, all)
summary(model_interactions1)
```

```
##
## Call:
## lm(formula = stress_ego ~ degree, data = all)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.2609 -0.1952 -0.1314  0.1399  0.7768
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.009446   0.028057   0.337    0.736
## degree      0.366203   0.048905   7.488 9.77e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2765 on 2400 degrees of freedom
```

```
## Multiple R-squared:  0.02283,    Adjusted R-squared:  0.02242
## F-statistic: 56.07 on 1 and 2400 DF,  p-value: 9.766e-14
```

6) Conclusions

And give your best conclusion as to whether your initial hypothesis held up - and if not, why not.

My initial hypothesis did not hold up. Although there was indeed a relationship between the two variables measured (degree and depression), it was opposite to what I had expected. It is hard to find immediate reasons for this relationship. The most obvious reason would of course be that I am operating on false assumptions about how social life and depression are related. Another, also seemingly likely, reason would be that something went wrong when recoding the variables, that threw of my regression models.