INFX 576: Problem Set 1 - Basic Concepts and Network Data*

Harkar Talwar

Due: Friday, April 6, 2018

Collaborators: Prateek Tripathi

Instructions:

Before beginning this assignment, please ensure you have access to R and RStudio.

1. Download the `problemset1.Rmd` file from Canvas. Open `problemset1.Rmd` in RStudio and supply your solutions to the assignment by editing `problemset1.Rmd`. You will also need the `problemset1_data.Rdata` file which contains the network datasets needed for this assignment.
2. Replace the "Insert Your Name Here" text in the `author:` field with your own full name. Any collaborators must be listed on the top of your assignment.
3. Be sure to include well-documented (e.g. commented) code chucks, figures and clearly written text chunk explanations as necessary. Any figures should be clearly labeled and appropriately referenced within the text.
4. Collaboration on problem sets is acceptable, and even encouraged, but each student must turn in an individual write-up in his or her own words and his or her own work. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.
5. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit PDF`, rename the R Markdown file to `YourLastName_YourFirstName_ps1.Rmd`, knit a PDF and submit the PDF file on Canvas.

Setup:

In this problem set you will need, at minimum, the following R packages.

```r
# Load standard libraries
library(statnet)

# Load data
load("problemset1_data.Rdata")
ls() # Print objects in workspace to see what is available
```

```
## [1] "sw.incidence"
```

Problem 1: Manipulating Network Data

Consider the follow question for eliciting social relationships: "Who have you talked to about important matters within the last 48 hours." Following the steps outlined below, import this data into R and make a visualization of your egocentric network.

- Write down a response to this question.
  I've talked to 5 persons about important matters within the last 48 hours. They belong to two categories,

*Problems originally written by C.T. Butts (2009) *friends and family. Among my friends are two individuals, Sahil and Prateek. In my family, I spoke with my dad, mom and sister.*

- Consider the induced network among the people who have talked to about important matters. Estimate (use your best guess) the relationships between each pair of individuals.
  I've estimated several relationships between my alters. In my family, my mom, dad and sister, all speak to each other on important matters. Further, as per my knowledge, both my friends, Sahil and Prateek also talk to each other on important matters. There are however no ties between my family and friend sub-groups.

- Organize your data into a spreadsheet using your favorite editor. Save your network data as a CSV. Import the data into R and make a visualization of your network.

```
# Read the adjacency matrix of the egocentric network from a CSV file
csv.matrix <- read.csv("EgoCentricNetwork.csv", header = TRUE, row.names = 1, check.names = FALSE)
print(csv.matrix)
```

```
##         Harkar Friend_Sahil Friend_Prateek Family_ Dad Family_Mom
## Harkar       0            1              1           1          1
## Sahil        1            0              1           0          0
## Prateek      1            1              0           0          0
## Dad          1            0              0           0          1
## Mom          1            0              0           1          0
## Sister       1            0              0           1          1
##         Family_Sister
## Harkar              1
## Sahil               0
## Prateek             0
## Dad                 1
## Mom                 1
## Sister              0
```

Figure 1a. Harkar's Egocentric Network as an Adjacency Matrix

```
# Convert into an R matrix for the purpose of our analysis
analysis.matrix <- as.matrix(csv.matrix)
# Create a network object from the matrix
egocentric.network <- network(analysis.matrix, matrix.type = "adjacency", directed = FALSE)
# Create a custom 'type' attribute for the vertices in the network
egocentric.network %v% "type" <- c("Self", "Friend", "Friend", "Family", "Family", "Family")
# Set the color for nodes of each type
vertex_colors <- rep("", network.size(egocentric.network))
for (i in 1 : length(vertex_colors)) {
    if (get.node.attr(egocentric.network, "type")[i] == "Self") {
        vertex_colors[i] <- "brown"
    } else if (get.node.attr(egocentric.network, "type")[i] == "Friend") {
        vertex_colors[i] <- "darkolivegreen1"
    } else {
        vertex_colors[i] <- "deepskyblue"
    }
}
# Plot/visualize the network
gplot(egocentric.network, gmode = "graph", label = network.vertex.names(egocentric.network), vertex.col
```

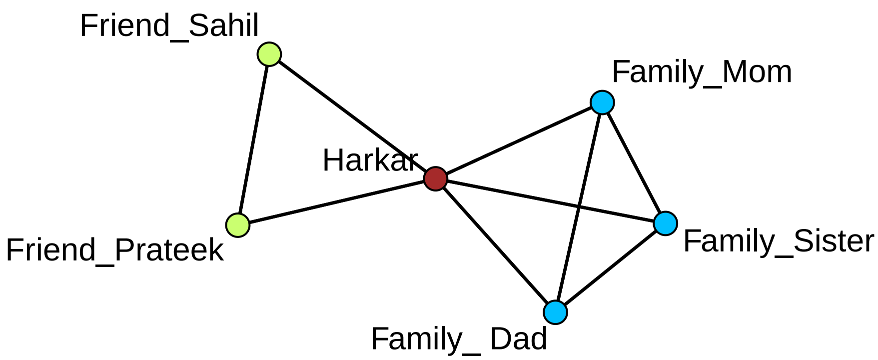

Figure 1b. Plot of Harkar's Egocentric Network

- Comment on what you see.
  The plot in Figure 1b. helps visualize the relations based on "free recall", which I had recorded in the CSV file. I, being the focal actor, am connected to each of the alters. The sub-groups are represented by different colors as coded above. Relationships with the "friends" and "family" sub-groups as represented in the matrix (Figure 1a.) are depicted in the plot (Figure 1b.), with both sub-groups tied to the focal actor and no ties between the sub-groups themselves.

Problem 2: Two-Mode Network Data

After loading the data for this problem set, you can use the `ls()` command to reveal the object `sw.incidence` This is the incidence matrix for the famous "Southern Women" dataset from Davis, Gardner, and Gardner's 1941 study of class and social interaction in the Deep South[1]. The matrix shows the attendance of 18 women at 14 informal social events during a nine-month observation period, based on various data sources such as interviews, guest lists, and participant observation. This is clearly two-mode data, with individuals as the "row vertices" and events as the "column vertices".

(a) Exploring Network Data

Begin by printing the matrix, and plotting it using `plot.sociomatrix`. Who seems to be the most active? Are all the women active in the same events? Describe what you observe.

```r
print(sw.incidence)
```

```
##           E1 E2 E3 E4 E5 E6 E7 E8 E9 E10 E11 E12 E13 E14
## EVELYN     1  1  1  1  1  1  0  1  1   0   0   0   0   0
```

[1] Davis, Gardner, and Gardner. (1941) *Deep South*. Chicago: The University of Chicago Press.

```
## LAURA       1  1  1  0  1  1  1  1  0   0   0   0   0   0
## THERESA     0  1  1  1  1  1  1  1  1   0   0   0   0   0
## BRENDA      1  0  1  1  1  1  1  1  0   0   0   0   0   0
## CHARLOTTE   0  0  1  1  1  0  1  0  0   0   0   0   0   0
## FRANCES     0  0  1  0  1  1  0  1  0   0   0   0   0   0
## ELEANOR     0  0  0  0  1  1  1  1  0   0   0   0   0   0
## PEARL       0  0  0  0  0  1  0  1  1   0   0   0   0   0
## RUTH        0  0  0  0  1  0  1  1  1   0   0   0   0   0
## VERNE       0  0  0  0  0  0  1  1  1   0   0   1   0   0
## MYRNA       0  0  0  0  0  0  0  1  1   1   0   1   0   0
## KATHERINE   0  0  0  0  0  0  0  1  1   1   0   1   1   1
## SYLVIA      0  0  0  0  0  0  1  1  1   1   0   1   1   1
## NORA        0  0  0  0  0  1  1  0  1   1   1   1   1   1
## HELEN       0  0  0  0  0  0  1  1  0   1   1   1   1   1
## DOROTHY     0  0  0  0  0  0  0  1  1   1   0   1   0   0
## OLIVIA      0  0  0  0  0  0  0  0  1   0   1   0   0   0
## FLORA       0  0  0  0  0  0  0  0  1   0   1   0   0   0
```

Figure 2a. Southern Women Events Affiliation Matrix

```r
plot.sociomatrix(sw.incidence, cex.lab = 0.35, xlab = 'Figure 2b. Southern Women Events Affilation - So
```

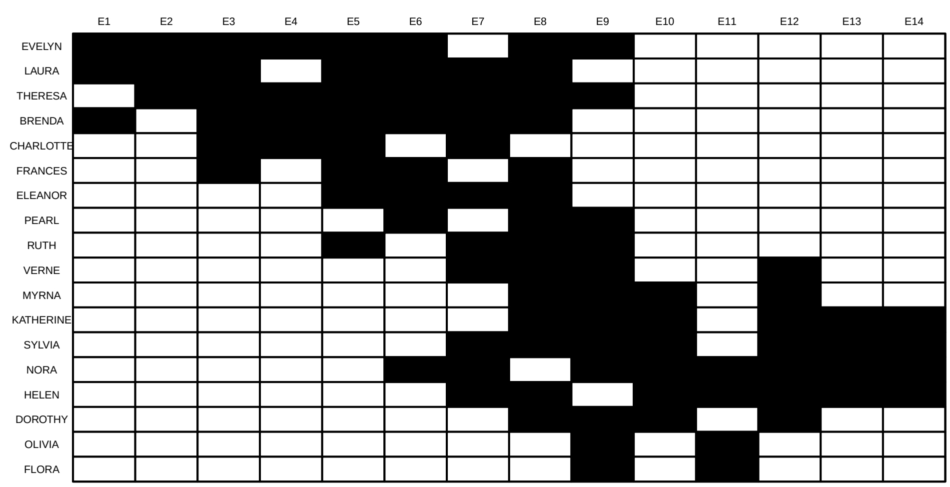

Figure 2b. Southern Women Events Affilation − Sociomatrix Form

As per Figure 2a, 2b, Evelyn, Theresa and Nora seem to be the most active women, each having ties to 8 events. All women are not active in the same events, since the rows in the adjacency matrix printed above are not identical for all of them. (Olivia, Flora) and (Dorothy, Myrna) are pairs of women who are active in the same events (Figure 2b). Moreover, there seem to be almost two, although not clearly defined sub-groups of the women, which are discussed in greater detail later. Lastly, it can be observed that the event E8 has the *most number of active women (14).*

(b) One-Mode Projections

Consider how these women are connected through events. To do this, form the (valued) row projection of `sw.incidence` and say it as `sw.p2p`. You might find it helpfull to know that `%*%` is R's inner product operator, and `t()` is a function to transpose a matrix. `sw.p2p[i,j]` should now be the number of events that i and j have in common. Plot this matrix as in part (a) and answer the following:

```r
sw.p2p <- sw.incidence %*% t(sw.incidence)
sw.p2p
```

```
##           EVELYN LAURA THERESA BRENDA CHARLOTTE FRANCES ELEANOR PEARL RUTH
## EVELYN         8     6       7      6         3       4       3     3    3
## LAURA          6     7       6      6         3       4       4     2    3
## THERESA        7     6       8      6         4       4       4     3    4
## BRENDA         6     6       6      7         4       4       4     2    3
## CHARLOTTE      3     3       4      4         4       2       2     0    2
## FRANCES        4     4       4      4         2       4       3     2    2
## ELEANOR        3     4       4      4         2       3       4     2    3
## PEARL          3     2       3      2         0       2       2     3    2
## RUTH           3     3       4      3         2       2       3     2    4
## VERNE          2     2       3      2         1       1       2     2    3
## MYRNA          2     1       2      1         0       1       1     2    2
## KATHERINE      2     1       2      1         0       1       1     2    2
## SYLVIA         2     2       3      2         1       1       2     2    3
## NORA           2     2       3      2         1       1       2     2    2
## HELEN          1     2       2      2         1       1       2     1    2
## DOROTHY        2     1       2      1         0       1       1     2    2
## OLIVIA         1     0       1      0         0       0       0     1    1
## FLORA          1     0       1      0         0       0       0     1    1
##           VERNE MYRNA KATHERINE SYLVIA NORA HELEN DOROTHY OLIVIA FLORA
## EVELYN        2     2         2      2    2     1       2      1     1
## LAURA         2     1         1      2    2     2       1      0     0
## THERESA       3     2         2      3    3     2       2      1     1
## BRENDA        2     1         1      2    2     2       1      0     0
## CHARLOTTE     1     0         0      1    1     1       0      0     0
## FRANCES       1     1         1      1    1     1       1      0     0
## ELEANOR       2     1         1      2    2     2       1      0     0
## PEARL         2     2         2      2    2     1       2      1     1
## RUTH          3     2         2      3    2     2       2      1     1
## VERNE         4     3         3      4    3     3       3      1     1
## MYRNA         3     4         4      4    3     3       4      1     1
## KATHERINE     3     4         6      6    5     5       4      1     1
## SYLVIA        4     4         6      7    6     6       4      1     1
## NORA          3     3         5      6    8     6       3      2     2
## HELEN         3     3         5      6    6     7       3      1     1
## DOROTHY       3     4         4      4    3     3       4      1     1
## OLIVIA        1     1         1      1    2     1       1      2     2
## FLORA         1     1         1      1    2     1       1      2     2
```

Figure 3a. Row Projection Matrix

```r
plot.sociomatrix(sw.p2p, cex.lab = 0.25, xlab = 'Figure 3b. Row Projections - Sociomatrix Form')
```

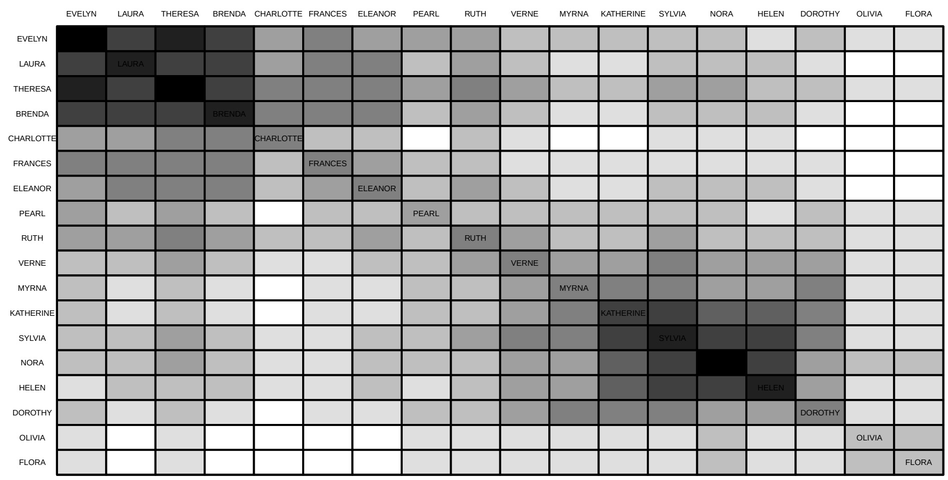

Figure 3b. Row Projections − Sociomatrix Form

- What does the row projection tell us about how people are connected in this social group?
  Performing a row-projection helps convert the two-mode data into one-mode, thereby helping us analyze if there is a tie between two given women, as shown in Figure 3a. A tie here indicates that two women are active in one or more of the same events. So, for instance, the element (Evelyn, Laura), whose value is 6, indicates that there are 6 events in common between Evelyn and Laura. We also observe from the sociomatrix in Figure 3b that Evelyn and Theresa seem to have the highest number of events in common, which is 7, as per Figure 3a. Most pairs of women seem to be connected. Moreover, Olivia and Flora seem to have the least common events with other women.

- Does the group seem to have subdivision?
  Figure 3b indicates that there exists some sort of sub-division among the Southern Women, with roughly two different sub groups, one towards the top left of the sociomatrix, and one towards the bottom right. This claim seems to be corroborated by the adjancency matrix in part Figure 3a, with one group of women being active mainly in events ranging from E1 to E9 and other group being active primarily in events E6 to E14, with some overlap between the two sub-groups. Figure 3b shows how Evelyn, Laura, Theresa, Brenda form one of the sub-divisions and Katherine, Sylvia, Nora, Helen form the other. This is based on an assumption that a group of women sharing 5 or more events could be indicative of some sub-division.

- Do some members seem more "central" than others? If so, who?
  From Figure 3a, it can be seen that some members are more central than others. Evelyn, Theresa, Brenda and Sylvia, for instance, are women with whom 4 or more other women share more than 4 events. This means that these women may be acting as key influencers in the entire network.

(c) Entailment Structures

Now, we are going to explore the *entailment structures* of women and events. We can construct a row-wise entailment matrix using the following code. The new matrix will be a person by person matrix such that `sw.r.entail[i,j]==1` if person j attends all of person i's events.

Use this function to create the entailment matrices (row-wise and column-wise) and produce a visualization of the entailment network for each case.

```
# Code to determine the row-wise entailment structure
# Create a new empty matrix
sw.r.entail <- matrix(0, nc=nrow(sw.incidence), nr=nrow(sw.incidence))
# Populate the matrix using a nested 'for' loop
for (i in 1:nrow(sw.incidence)){ # Pick an women i
  for (j in 1:nrow(sw.incidence)){ # And and women j
    sw.r.entail[i,j] <- all(sw.incidence[j,] >= sw.incidence[i,]) # Compare them
  }
}
rownames(sw.r.entail) <- rownames(sw.incidence) # Renames the nodes
colnames(sw.r.entail) <- rownames(sw.incidence)

# Plot the row-wise entailment structure
gplot(sw.r.entail, label=rownames(sw.r.entail), label.cex=.5,
      boxed.labels=FALSE, vertex.cex=1.25, xlab="Figure 4. Row Entailment Structures")
```

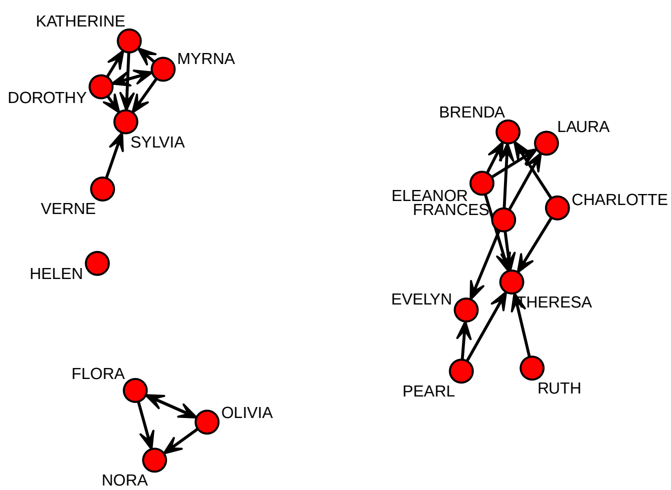

Figure 4. Row Entailment Structures

```
# Code to determine the column-wise entailment structure
# Create a new empty matrix
sw.c.entail <- matrix(0, nc=ncol(sw.incidence), nr=ncol(sw.incidence))
# Populate the matrix using a nested 'for' loop
for (i in 1:ncol(sw.incidence)){ # Pick an event i
  for (j in 1:ncol(sw.incidence)){ # And an event j
```

```
    sw.c.entail[i,j] <- all(sw.incidence[,j] >= sw.incidence[,i]) # Compare them
  }
}
rownames(sw.c.entail) <- colnames(sw.incidence) # Renames the nodes
colnames(sw.c.entail) <- colnames(sw.incidence)

# Plot the columne-wise entailment structure
gplot(sw.c.entail, label=rownames(sw.c.entail), label.cex=.5,
      boxed.labels=FALSE, vertex.cex=1.25, xlab="Figure 5. Column Entailment Structures")
```

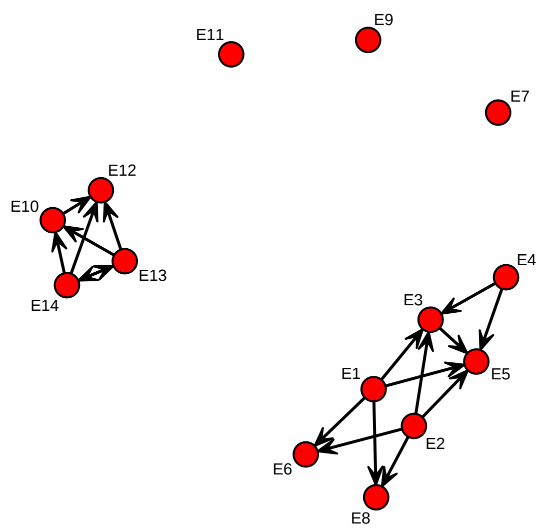

Figure 5. Column Entailment Structures

Use the matrices and visualizations to answer:

- What does a path tell us?
  In the row wise entailment plot represented by Figure 4, a path from vertex i to vertex j signifies that the events for which the woman at vertex i is active are a proper subset of the events for which the woman at vertex j is active, i.e., $E_i \subset E_j$*, where* E_i *is the set of events for which i is active.In other words, the path from Olivia to Nora indicates that Nora attends all events that Olivia does. On the other hand, in the column entailment structure of Figure 5, a path from event i to event j, indicates that the women active in event i are a proper subset of the women active in j, i.e.,* $W_i \subset W_j$*, where* W_i *is the set of women active in event i. The path from E4 to E3 therefore means that all women who are active in the event E3 are also active in the event E4.*

- What do mutual (i.e. bidirectional) dyads mean?
  In Figure 4 (Row Entailment Structures), a mutual dyad between two vertices represents two women, the set of whose events are equal, i.e., $E_i = E_j$*. This implies that, Olivia and Flora, for instance, are active in exactly the same events.*
  Similarly, mutual dyads in Figure 5 (Column Entailment Structures), represent events for which the set

of active women are the same, i.e., $W_i = W_j$. *So, for instance, the women who attend E13 and E14 are the same.*

- What is special about isolates?
  In Figure 4 (Row Entailment Structures), an isolate represents a woman, the set of whose events are neither a subset nor a superset of the events for any other woman. This means that there is no woman that attends all of the events that Helen, for instance, does. Likewise, there is no such woman, all of whose events are attended by Helen. In Figure 3 (Column Entailment Structures), an isolate represents an event, for which the set of active women is neither a superset nor a subset of the set of active women for any other event. So, for instance, there is no event for which all women attending E9 (isolate) are also active. Similarly, there is no such event for which all women active are also active in E9.