INFX 576: Problem Set 3 - Random Graphs and Network Tests*

Harkar Talwar

Due: Friday, April 20, 2018

Collaborators: Prateek Tripathi, Aakash Agrawal

Instructions:

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

1. Download the `problemset3.Rmd` file from Canvas. You will also need the `problemset3_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_ps3.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.

```
# Load standard libraries
library(statnet)

# Load data
load("problemset3_data.Rdata")
ls() # Print objects in workspace to see what is available
```

```
## [1] "kaptail.ins" "mids_1993"
```

Problem 1: Random Graphs

(a) Generating Random Graphs

Generate 100-node random directed graphs with expected densities of 0.0025, 0.005, 0.01, 0.015, 0.02, and 0.025, with at least 500 graphs per sample. Remember the `rgraph` function can draw more than one graph at a time. Plot the average Krackhardt connectedness, dyadic reciprocity, and edgewise reciprocity as a function of expected density. Use these to describe the baseline effect of increasing density on network structure.

*Problems originally written by C.T. Butts (2009)

```r
# Vector to store density values
densities = c(0.0025, 0.005, 0.01, 0.015, 0.02, 0.025)
# Vectors to store average GLI values for different density graph sets
connect = array(dim = c(6))
dyad.recip = array(dim = c(6))
edge.recip = array(dim = c(6))
for (i in 1:length(densities)) {
  # Generate a sample of random graphs for the given density
  graph.set = rgraph(n = 100, m = 500, tprob = densities[i])
  # Populate the average GLI values in the respective vectors
  connect[i] = mean(connectedness(graph.set))
  dyad.recip[i] = mean(grecip(graph.set, measure = "dyadic"))
  edge.recip[i] = mean(grecip(graph.set, measure = "edgewise"))
}

# Plot the average GLI values against density
par(mfrow=c(2,2))
plot(densities, connect, type = 'b', lty = 2, lwd = 2, col = 6,
     xlab = "Expected Density\n", ylab = "Krackhardt Connectedness",
     main = "Density vs Connectedness", sub = "\nFigure 1a.")
plot(densities, dyad.recip, type = 'b', lty = 2, lwd = 2, col = 1,
     xlab = "Expected Density\n", ylab = "Dyadic Reciprocity",
     main = "Density vs Dyadic Reciprocity", sub = "\nFigure 1b.")
plot(densities, edge.recip, type = 'b', lty = 2, lwd = 2, col = 2,
     xlab = "Expected Density\n", ylab = "Edgewise Reciprocity",
     main = "Density vs Edgewise Reciprocity", sub = "\nFigure 1c.")
```

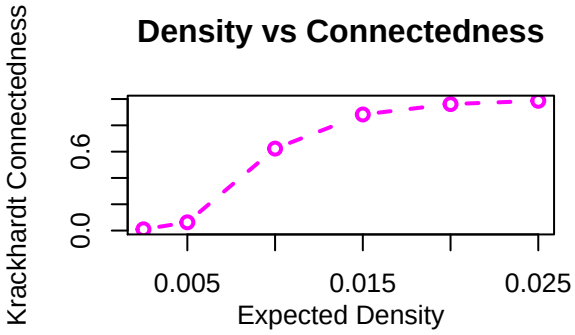


Figure 1a.

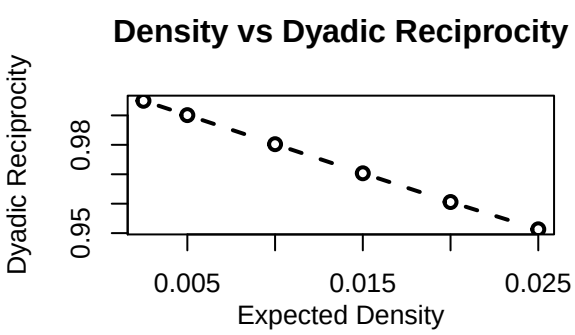


Figure 1b.

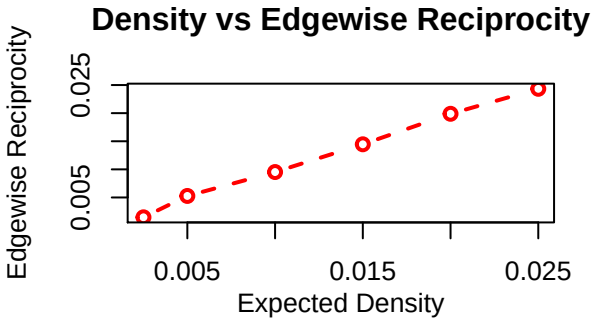


Figure 1c.

Observations:

- Krackhardt connectedness measures the fraction of all vertex pairs that are (at least) weakly connected, i.e., neglecting orientation, to what extent a path exists between each pair of vertices. The plot in figure 1a shows that as the density increases, the average Krackhardt connectedness increases as well. This is what one would expect intuitively as well, since increasing the density increases the proportion of dyadic connections present. The increase in connectedness value slows down beyond a certain density, which can be seen as the curve flattens out and peaks at about a connectedness of 0.98.
- Dyadic reciprocity is the proportion of dyadic connections that are mutual or null in nature. In figure 1b, we observe that on increasing the density, the dyadic reciprocity starts to fall. This can be attributed to the fact that an increase in density leads to a reduction of the number of null dyads (isolates) present in the network structure, which in turn reduces the dyadic reciprocity.

- Edgewise reciprocity measures the proportion of non null dyads that are mutual in nature. Increasing the density reduces the number of null dyads, which means that either mutual or asymmetric dyads could increase. However, each mutual dyad that is introduced has a slightly more pronounced effect on edgewise reciprocity than an asymmetric dyad ($recip = \frac{2M}{2M+A}$). Therefore in figure 1c, we see that edgewise reciprocity increases with the increase in density.

(b) Comparing GLIs

In this problem we will use the well-known social network dataset, collected by Bruce Kapferer in Zambia from June 1965 to August 1965, involves interactions among workers in a tailor shop as observed by Kapferer himself.[1] Here, an interaction is defined by Kapferer as "continuous uninterrupted social activity involving the participation of at least two persons"; only transactions that were relatively frequent are recorded.

Generate 500 random directed graphs whose dyad census is the same as that of `kaptail.ins`. Plot histograms for total degree centralization, betweenness centralization, transitivity, and Krackhardt connectedness from this random sample. On your plot mark the observed values of these statistics (from the `kaptail.ins` data) using a vertial line You mgith find the `abline` function helpful here. Try modifying the `lwd` argument to the plot function to make the vertical line stand out. How do the replicated graphs compare to the observed data.

```
# Compute the dyad census of the Kapferer network data
dyad.cen = dyad.census(kaptail.ins)
# Create random graphs having the same dyad census as the tailor shop network
g.uman.set = rguman(500, network.size(kaptail.ins), mut = dyad.cen[1,1],
                    asym = dyad.cen[1,2], null = dyad.cen[1,3], method = "exact")
# Compute GLIs for the random graphs (baseline)
# Total Degree Centralization
g.degree.cent = centralization(g.uman.set, FUN = 'degree', cmode = 'freeman')
# Betweenness Centralization
g.betw.cent = centralization(g.uman.set, FUN = 'betweenness')
# Transitivity
g.trans = gtrans(g.uman.set)
# Krackhardt Connectedness
g.connect = connectedness(g.uman.set)

# Compute GLIs for the our observed network data (tailor shop)
kap.degree.cent = centralization(kaptail.ins, FUN = 'degree', cmode = 'freeman')
kap.betw.cent = centralization(kaptail.ins, FUN = 'betweenness')
kap.trans = gtrans(kaptail.ins)
kap.connect = connectedness(kaptail.ins)

# Plot the GLI distributions for the random graphs and compare with observed
```

[1] Kapferer B. (1972). Strategy and transaction in an African factory. Manchester: Manchester University Press.

```r
# values
par(mfrow=c(2,2))  # 2 x 2 grid for 4 sub-plots
# Total Degree Centralization
hist(g.degree.cent, xlab = 'Total Degree Centralization\n',
     main = 'Histogram: Degree Centralization', sub = 'Figure 2a')
abline(v = kap.degree.cent, col = 3, lwd = 2)
# Betweenness Centralization
hist(g.betw.cent, xlab = 'Betweeness Centralization\n',
     main = 'Histogram: Betweeness Centralization', sub = 'Figure 2b')
abline(v = kap.betw.cent, col = 3, lwd = 2)
# Transitivity
hist(g.trans, xlim = c(0, kap.trans), xlab = 'Transitivity\n',
     main = 'Histogram: Transitivity', sub = 'Figure 2c')
abline(v = kap.trans, col = 3, lwd = 2)
# Krackhardt Connectedness
hist(g.connect, xlim = c(0.7, 1), xlab = 'Krackhardt Connectedness\n',
     main = 'Histogram: Krackhardt Connectedness', sub = 'Figure 2d')
abline(v = kap.connect, col = 3, lwd = 2)
```

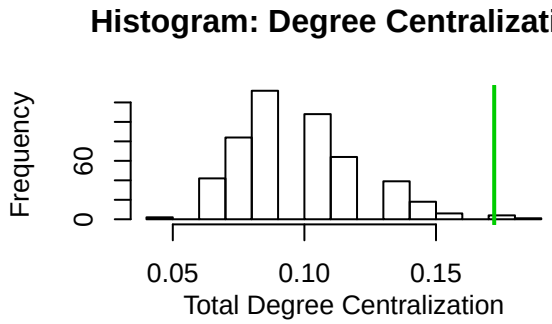


Figure 2a

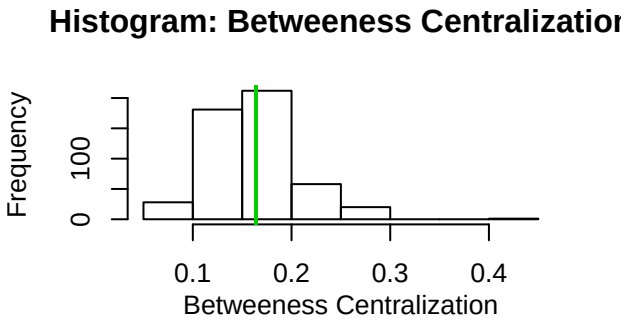


Figure 2b

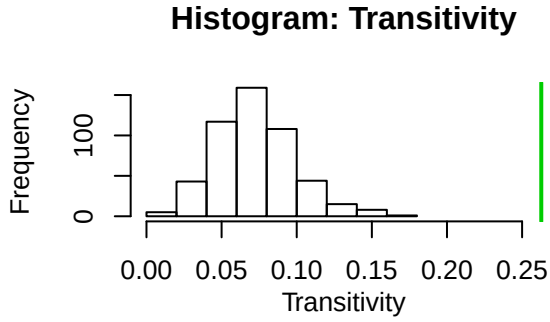


Figure 2c

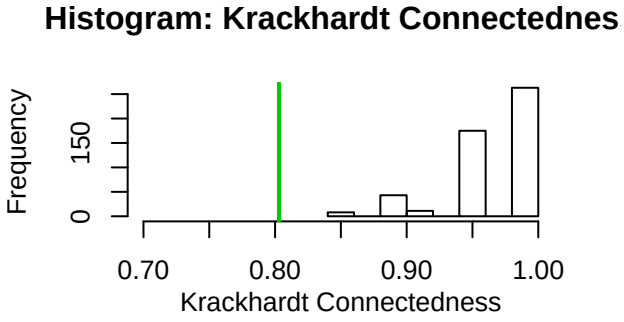


Figure 2d

Observations:

- Figure 2a shows the distribution of Degree centralization values for random graphs having the same dyad census as our tailor shop network. We can see that the observed value of degree centralization is significantly higher than almost all of our random graphs. This alludes to the fact that there is a relative dominance of some workers in the network.

- Figure 2b depicts the betweenness centralization distribution for the random graphs as well as the observed value for the network being analyzed. The observed value in the tailor shop network is close to the expected value from the sample of random graphs. So, the structure does not seem to be biased in terms of a few workers exerting excessive control over the flow/transmission of information.
- Figure 2c shows a marked departure in the transitivity features of the tailor shop network, when compared to the expected values in our random graph samples. The proportion of interactions between workers, where i->j, j->k implies i->k is significantly higher among these workers, than in the generated baseline graphs.
- Figure 2d depicts that relation between the observed Krackhardt Connectedness versus the baseline values. Again we see that our tailor shop network is biased in this regard. The proportion of workers that are at least (weakly) connected (i.e. regardless of orientation) is smaller than what we observe in our random graph samples. This also suggests that the network could be more centalized than hierarchical, for networks of its dyad census configuration.

Problem 2: Testing Structural Hypotheses

Consider the following set of propositions, which may or may not be true of given dataset. For this problem, use the rdata from the Correlates ofWar project. For each, do the following:

1. Identify a statistic (e.g. GLI) whose value should deviate from a random baseline if the proposition is true.
2. Identify the approporate baseline distribution to which the statistic should be compared.
3. Determine whether the proposition implies that the statistic should be greater or lower than its baseline distribution would indicate.
4. Conduct a conditional uniform graph test based on your conclusions in 1-3. In reporting your results, include appropriate summary output from the `cug.test` function as well as the resulting distributional plots. Based on the results, indicate whether the data appears to support or undermine the proposition in question. Be sure to justify your conclusion.

Note: For the below parts, our chosen significance level is $\alpha = 0.05$

(a) In militarized interstate disputes, hostile acts are disproportionately likely to be responded to in kind.

Steps

1. **Identification of test statistic**: Reciprocity is deemed to be an appropriate graph level index for this proposition, since we are interested in knowing how likely a country is to respond in kind to war waged against it by another.
2. **Identification of baseline distribution**: The uniform conditional model on the number of edges would be appropriate in this case, since a model based on fixed dyad census would end up fixing the value of our test statistic as well.
3. **Determination of direction of deviation**: Here, our interest would be in an upper tail test, to see if the reciprocity in our network is higher than what would be expected from similar density structures.
4. **Conducting a Conditional Uniform Graph test**:

```
# Perform the CUG test, based on our chosen baseline distribution and test statistic
cug.recip = cug.test(mids_1993, FUN = "grecip", mode = "digraph", cmode = "edges",
                     reps = 2000, FUN.args = c(measure = "edgewise"))
print(cug.recip)
```

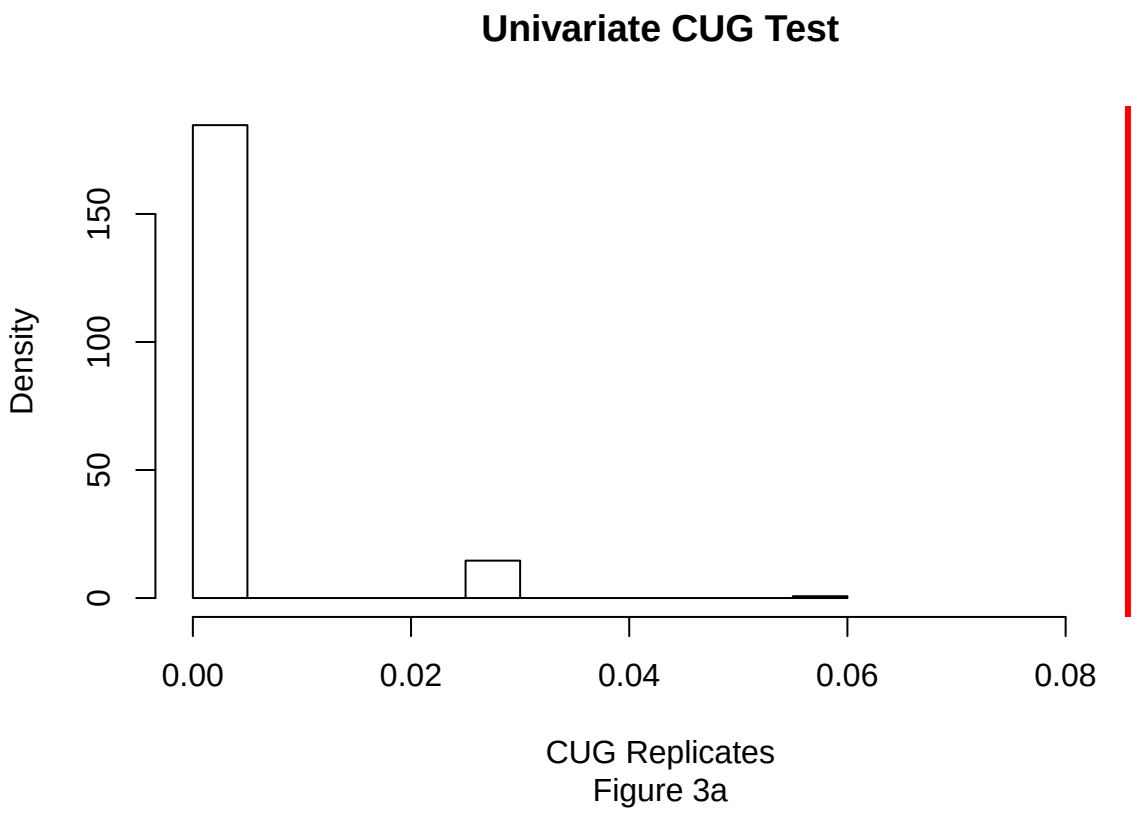
```
##
## Univariate Conditional Uniform Graph Test
##
## Conditioning Method: edges
```

```
## Graph Type: digraph
## Diagonal Used: FALSE
## Replications: 2000
##
## Observed Value: 0.08571429
## Pr(X>=Obs): 0
## Pr(X<=Obs): 1
```

```
summary(cug.recip)
```

```
##          Length Class  Mode
## obs.stat    1   -none- numeric
## rep.stat 2000   -none- numeric
## mode        1   -none- character
## diag        1   -none- logical
## cmode       1   -none- character
## plteobs     1   -none- numeric
## pgteobs     1   -none- numeric
## reps        1   -none- numeric
```

```
plot(cug.recip, sub = 'Figure 3a')
```



Analysis:

H_o: In militarized interstate disputes, hostile acts are proportionately likely to be responded to in kind.

As the results above show, given that the null hypothesis holds true, the probability of observing a value as extreme or more extreme than our observed reciprocity of 0.086 is 0 ($< \alpha$). This is also corroborated by the

plot in Figure 3a, where the density of values equal to or higher than our observed statistic is 0. Therefore we are able to reject our null hypothesis. This means that hostile acts are more likely to be responded to in kind in interstate disputes, than one would expect in random uniform networks with similar density configuration. Thus, the data supports the proposition.

(b) When engaging in disputes, nations behave in accordance with the notion that "the enemy of my enemy is not my enemy".

Steps

1. **Identification of test statistic**: Transitivity is deemed to be an appropriate graph level index for this proposition, since we are interested in knowing that for any countries i, j, k if i and j are enemies, j and k are enemies, how likely is it that i and k are also enemies.
2. **Identification of baseline distribution**: The uniform conditional model based on the dyad census would be appropriate in this case, since it would more closely mirror the dyadic relations depicting disputes between various countries represented by the network. Further, our test statistic (transitivity) would also not be affected by fixing the dyad census of our replicates.
3. **Determination of direction of deviation**: Here, our interest would be in a lower tail test, to see if the transitivity in our network is lower than what would be expected from similar dyadic configuration structures. This is based on the intuition that if the enemy of my enemy is not my enemy, then the transitivity would be reduced.
4. **Conducting a Conditional Uniform Graph test**:

```
# Perform the CUG test, based on our chosen baseline distribution and test statistic
cug.trans = cug.test(mids_1993, FUN = "gtrans", mode = "digraph",
                     cmode = "dyad.census", reps = 2000)
print(cug.trans)
```

```
##
## Univariate Conditional Uniform Graph Test
##
## Conditioning Method: dyad.census
## Graph Type: digraph
## Diagonal Used: FALSE
## Replications: 2000
##
## Observed Value: 0.02409639
## Pr(X>=Obs): 0.0495
## Pr(X<=Obs): 0.9505
```

```
summary(cug.trans)
```

```
##          Length Class  Mode
## obs.stat    1   -none- numeric
## rep.stat 2000   -none- numeric
## mode        1   -none- character
## diag        1   -none- logical
## cmode       1   -none- character
## plteobs     1   -none- numeric
## pgteobs     1   -none- numeric
## reps        1   -none- numeric
```

```
plot(cug.trans, sub = 'Figure 3b')
```

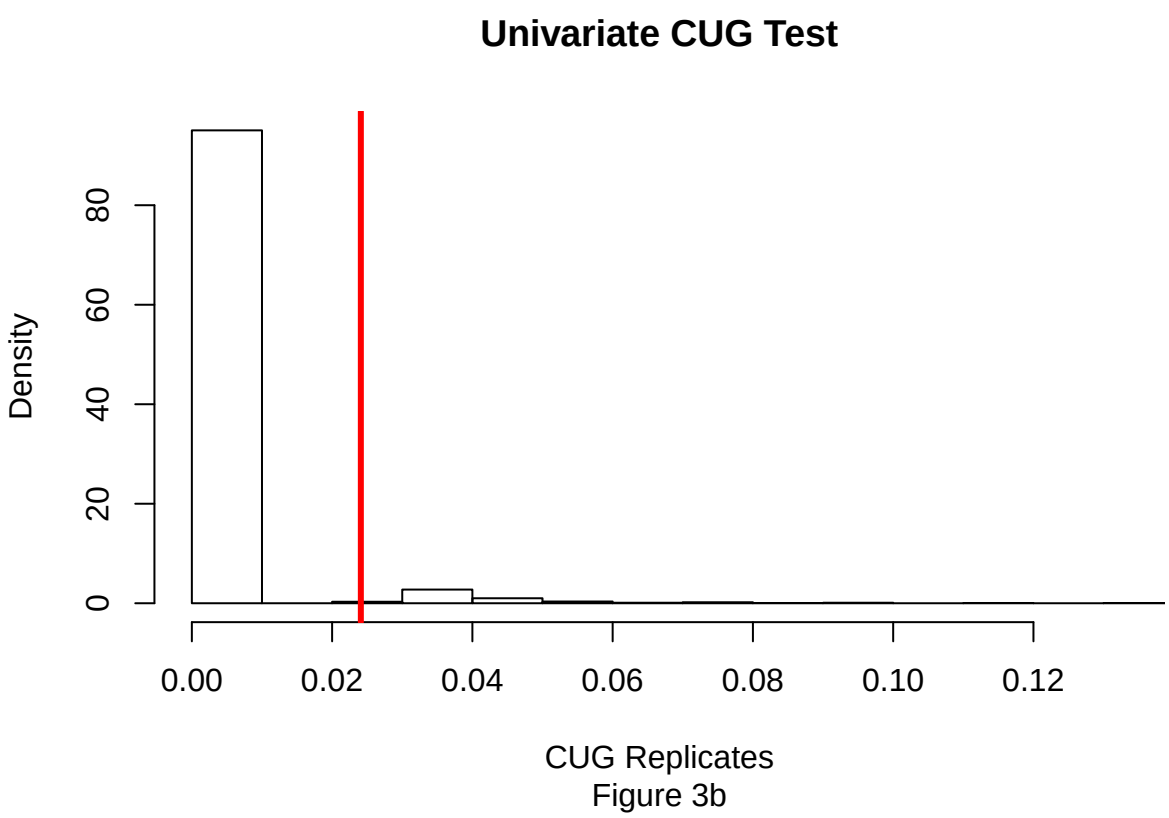


Figure 3b

Analysis:

H_o: When engaging in disputes, nations do not behave in accordance with the notion that "the enemy of my enemy is not my enemy".

As the results above show, given that the null hypothesis holds true, the probability of observing a value as low or lower than our observed transitivity of 0.024 is 0.943 ($> \alpha$). That is, more than 90% of the values in our replicates are less than the observed value. So, given nations i, j, k, where i->j and j->k are linked by enemity, the likelihood of i and k being enemies in our network is more than the baseline distributions. This is also supported by the plot in Figure 3b, where the density for values less than or equal to our observed statistic is pretty high. Therefore it is quite likely to observe such transitivity values in networks with similar dyadic configuration. This means that we are not able to reject our null hypothesis that nations do not behave in accordance with the notion that "the enemy of my enemy is not my enemy". In other words, the data undermines the proposition.

(c) Given the number of disputes at any given time, as small number of nations will receive a disproportionate share of aggressive acts.

Steps

1. **Identification of test statistic**: Indegree Centralization is deemed to be an appropriate graph level index for this proposition, since we are interested in knowing if a few countries are at the receiving end of a large number of aggresive acts of military action. This effect can be analyzed using Indegree Centralization.

2. **Identification of baseline distribution**: The uniform conditional model based on the number of edges would be appropriate in this case, since the proposition describes fixing the 'number of disputes', which translates to the edges in our network.
3. **Determination of direction of deviation**: Here, our interest would be in an upper tail test, to see if the indegree centralization in our network is higher than what would be expected from similar density structures. This would help us test the possibility of a few countries being at the receiving end of a disproportionate number of acts of military aggression.
4. **Conducting a Conditional Uniform Graph test**:

```
# Perform the CUG test, based on our chosen baseline distribution and test statistic
cug.indeg.cent = cug.test(mids_1993, FUN = "centralization", mode = "digraph",
                          cmode = "edge", reps = 2000,
                          FUN.args = c(FUN = "degree", cmode = "indegree"))
print(cug.indeg.cent)
```

```
##
## Univariate Conditional Uniform Graph Test
##
## Conditioning Method: edges
## Graph Type: digraph
## Diagonal Used: FALSE
## Replications: 2000
##
## Observed Value: 0.05773557
## Pr(X>=Obs): 0
## Pr(X<=Obs): 1
```

```
summary(cug.indeg.cent)
```

```
##          Length Class  Mode
## obs.stat    1   -none- numeric
## rep.stat 2000   -none- numeric
## mode        1   -none- character
## diag        1   -none- logical
## cmode       1   -none- character
## plteobs     1   -none- numeric
## pgteobs     1   -none- numeric
## reps        1   -none- numeric
```

```
plot(cug.indeg.cent, sub = 'Figure 3c')
```

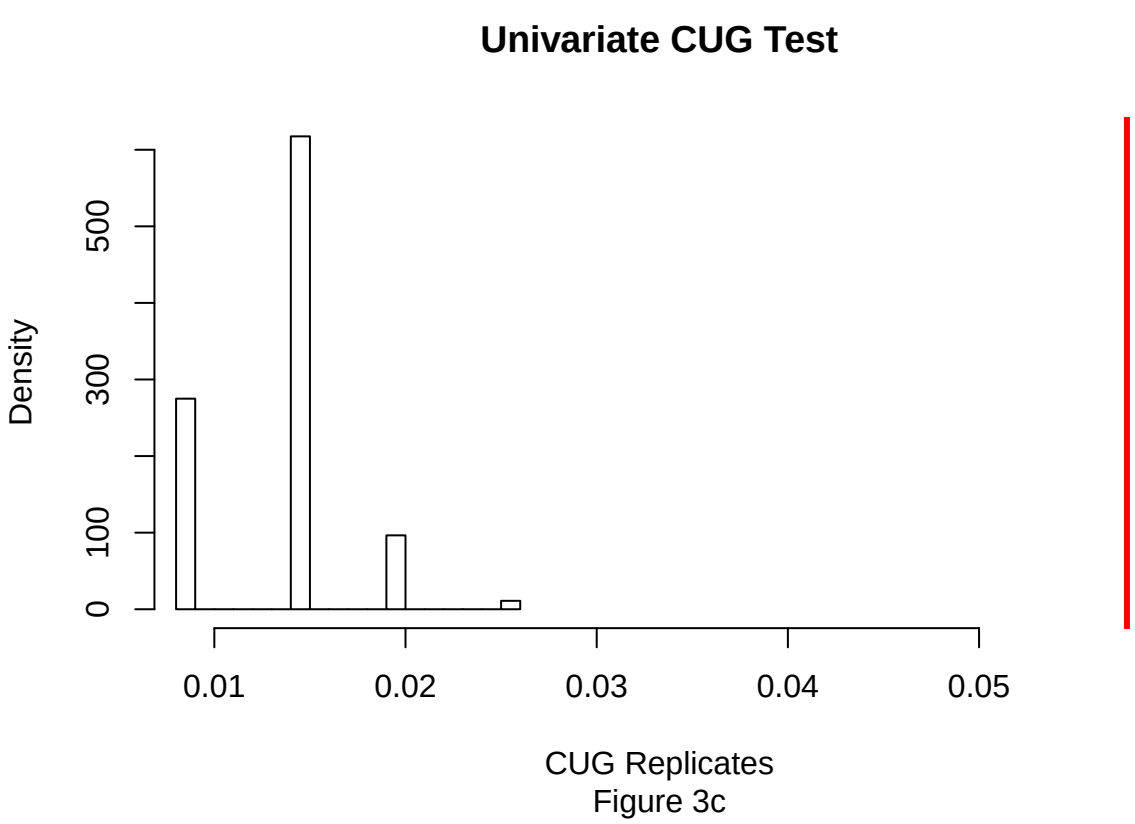


Figure 3c

Analysis:

H_o: Given the number of disputes at any given time, a small number of nations will not receive a disproportionate share of aggressive acts.

As the results above show, given that the null hypothesis holds true, the probability of observing a value as extreme or higher than our observed indegree centralization of 0.058 is 0 ($< \alpha$). This is supported by the plot in Figure 3c, where the density of values more than or equal to our observed statistic is 0. This means that we can reject our null hypothesis. In other words, the test provides evidence that a small number of nations are likely to be at the receiving end of a disproportionate number of acts of aggression. Thus, the data supports the proposition.