University of California, Los Angeles
Statistical Analysis of Social Networks
(Stats 218, Fall 2018)

Course Project Report

Ruchen Zhen
UID: 205036408

Advisor: Professor Mark S. Handcock
Teaching Assistant: Bart Blackburn

2018/12/14

1. Introduction

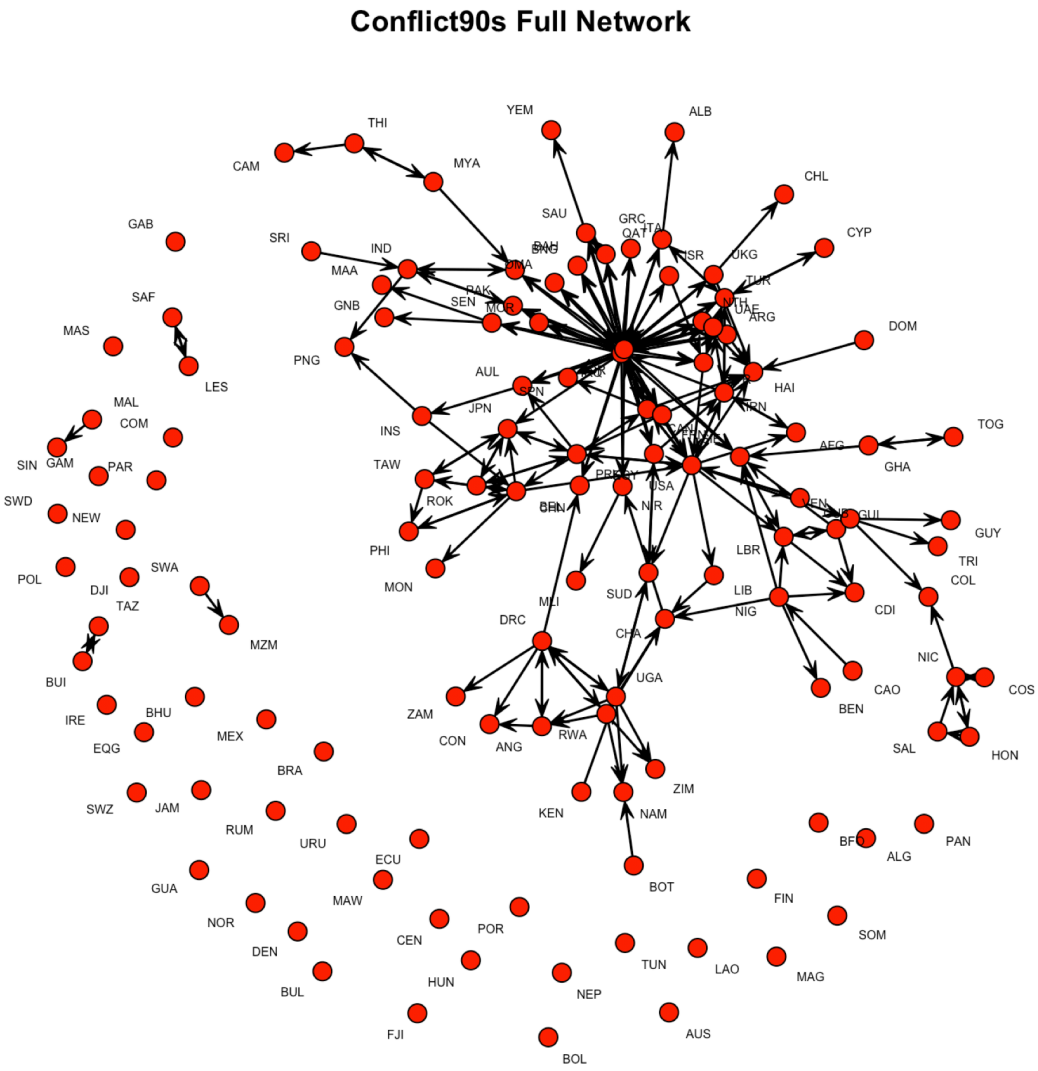
Social network analysis is an important study area. It can represent many social relations. For example, friendship, communication, international relationship, alliance, conflict, migration and so on. The statistical methods are often used to analysis the social network. There are two major aspects in statistical social network analysis. The descriptive methods focus on summarizing the macroscopic characteristic of network, such as degree distribution, centrality, connectivity, density and so on. These gives an overview of the existing network. The generative methods however, focus on finding out the intrinsic of the network. They try to fit the original network with some statistical models, then use the model to simulate other networks, and see if their statistics coincidence with the original network.

Among the social network analysis area, I am extreme interesting in international relationship networks. Study the country relationship can help us better understand the history, geology, international relation and international situation. The conflict90s network is about the trade conflicts during 1990s, including some nodal covariates (GDP, population, polity) as well as some dyadic covariates (number of conflicts, geographic distance, number of shared intergovernmental organizations). In this project, we will analyze this trade conflicts network using both descriptive methods and generative methods.

The rest of the report is organized as follows. Section 2 to 4 focus on descriptive analysis. Section 2 discuss the visualization and basic numerical summaries. Section 3 discuss the significant sub-graph. Section 4 talks about centralization. Section 5-8 focus on generative methods. Section 5 to 7 use the tapered Exponential-Family Random Graph Model (ERGM). Section 5 discuss the nodal covariates. Section 6 talks about the effect of triad census. Section discuss some other ERGM terms. Section 8 utilizes the latent position models for clustering.

2. Network Visualization and Numerical Summary

As a first part of the analysis, we visualize the network as follows.



As can be seen in the plot, most nodes are contained in the giant component. There are quite a few isolates, and a few pairs. This result is reasonable. Now the world is getting closer and closer to each other, most nations should be involved in the worldwide trade conflict. Some pairs may represent local conflicts be adjacent small countries. There should also be some neural countries. Some countries may claim that they are neural countries, others may be not that powerful, and are not willing to be involved in conflicts between alliances.

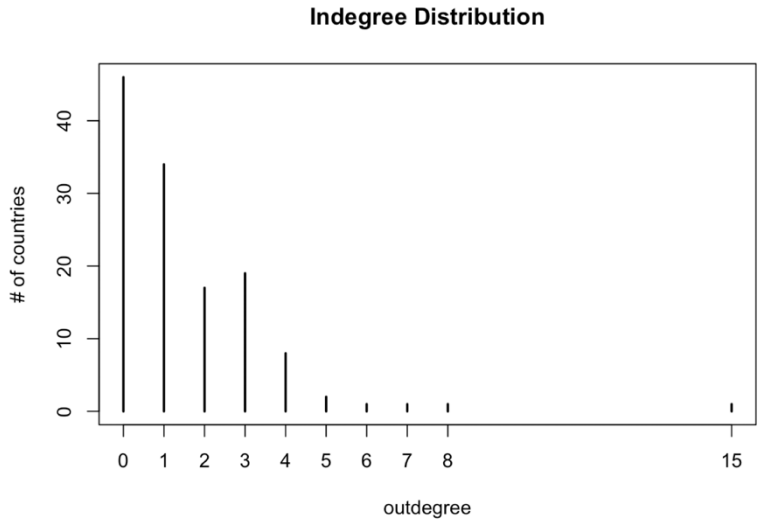
Given that we are analyzing the conflicts, the peaceful nations should be not helpful in analyzing the characteristics of conflicts. Thus, move the isolates should be a good idea. In most of the following analysis, we will perform the same analysis on both the whole network and the one with isolates removed, and argues that the results are similar.

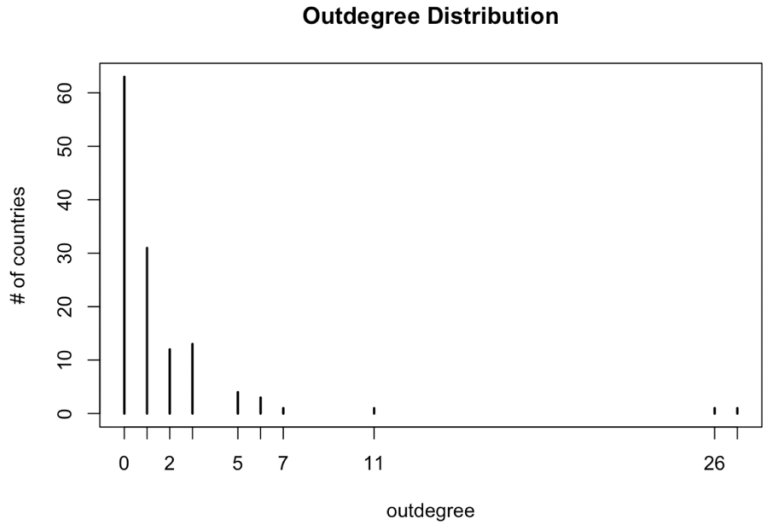
We perform some numerical summary measures of the network, as follows:

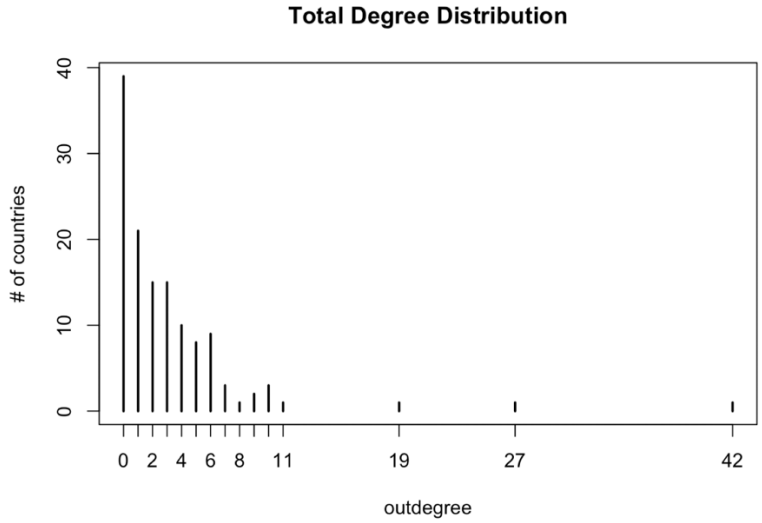
Number of Vertices	130
Number of Edges	203
Number of Isolates	39
Edge Density	0.012
Number of Transitive Triads	47
Number of Cyclic Triads	553

There are totally 39 isolates, indicating that 30% of the countries in the dataset are peaceful. The edge density is 0.012, which shows that the major tone of the world is still love and peace. There are 47 transitive triads as well as 553 cyclic triads.

Then we plot the degree distribution of indegree, outdegree and total degree, as follows:







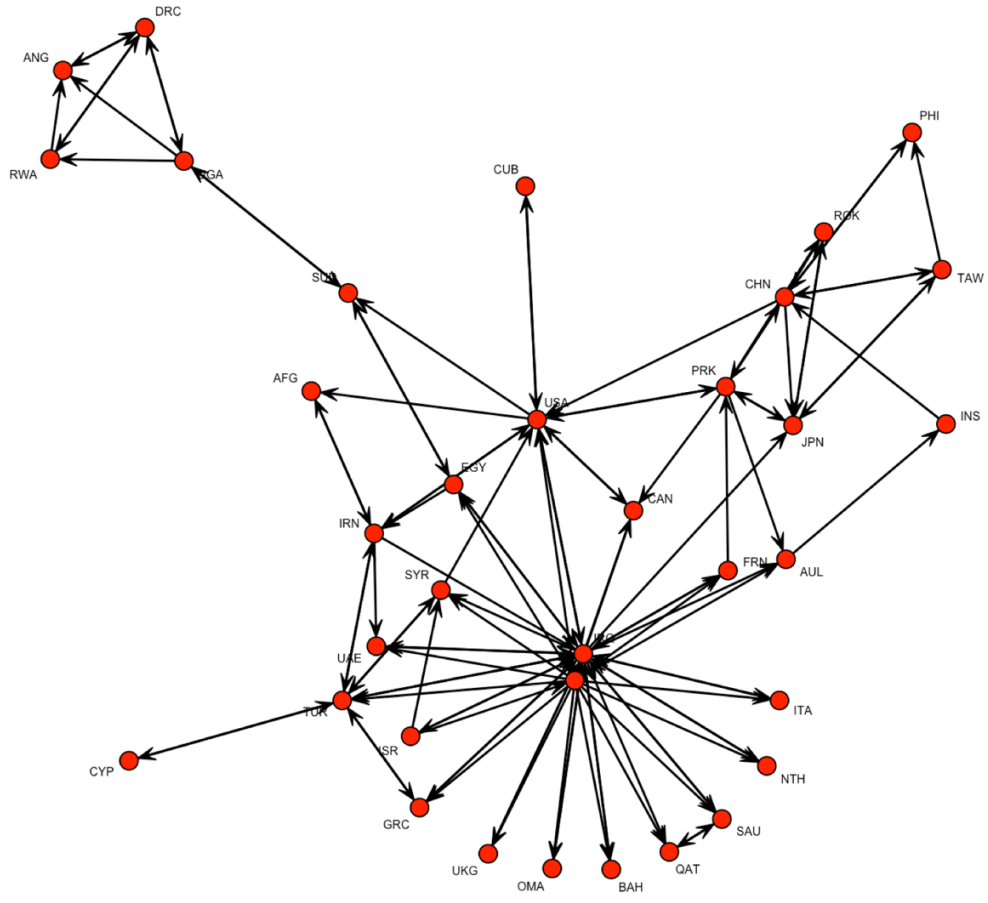
As can be seen in the degree distribution, there do exist some extremely aggressive countries, but they are not the majority. The highest outdegree is 27, second is 26, and third is only 11. As for the indegree, the highest is 15, while the second largest is only 8.

One thing extremely interesting is that, the country with the highest outdegree, 27, is Iraq. And the country with the highest indegree, 15, is also Iraq. Then what is the one with largest total degree? Definitely, Iraq. As the major player of the Guff War, the largest war in 1990s, Iraq fought again the alliance of some of its neighbors, Kuwait, Egypt, Syria, Qatar, the United Arab Emirates, Saudi Arabia, as well as some powerful western countries, the United States, the United Kingdom and France. No wander Iraq was also in the center of trade conflicts during 1990s.

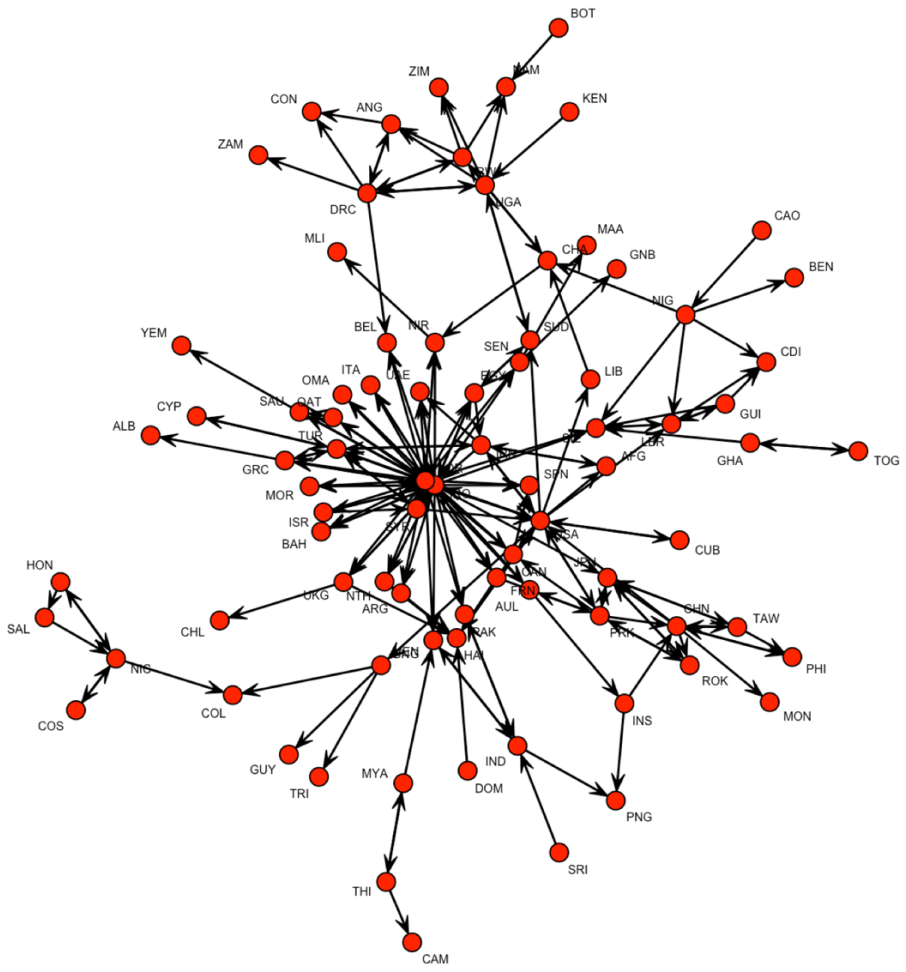
3. Significant Sub-Group in the Network

As can be seen in the plot of the whole network, the network is mainly consisted of a large giant component, a few pairs, and some isolates. We plot the giant components, both strongly connected and weakly connected, as follows, to see the size as well as structure of them.

Giant Component of Conflicts90s Network (Strongly Connected)



Giant Component of Conflicts90s Network (Weakly Connected)



In order to quantitively analyze how important of this giant component, we summary the number of vertices and edges of the full network and three subnetworks, as the following table.

Network Type	Number of Vertices	Percentage of Vertices	Number of Edges	Percentage of Edges
Full Network	130	100%	203	100%
Remove Isolates	91	70%	203	100%
Giant Component (Weakly Connected)	83	63.8%	197	97%
Giant Component (Strongly Connected)	35	26.9%	111	54.7%

There are 39 isolates in the original network, which is exactly 30% of all the vertices. There are totally 4 pairs, two of them are bidirectional and the others are single directional. As for the weakly connected giant component, which allows single directional connections, contains 83 vertices, 63.8 of the total vertices, and 197 edges, which is impressively to be 97% of the total edges. Even the strongly connected giant component, which has only 26.9% of vertices, contains 54.7% of the total edges. Therefore, we can draw a strong conclusion that the giant component represents the majority of the network.

These results seem quite reasonable. As the world getting closer and closer, most nations should be involved in the worldwide trade conflict. The peaceful countries might be some widely agreed neural (such as Finland in the plot), or some geographical isolated countries (such as Australia in the plot), or some small countries which is not powerful and not willing to be involved in large alliance (such as the Lao People's Democratic Republic in the plot). This network strongly supports this idea.

4. Centralization of the Network

We consider all the four types of centrality: degree centrality (based on degree), closeness centrality (based on average distances), betweenness centrality (based on geodesics), and eigenvector centrality (the centrality of each vertex is proportional to the sum of the centralities of its neighbors).

Intuitively, closeness centrality and betweenness centrality are not appropriate in. this problem. What we are considering is a trade conflict network, and conflict is between two countries. The occurring of a third country, or belonging in a certain alliance may leads to conflict or peace. However, the distance in this network is meaningless because, nations do not need some “steps”, or saying go through some countries, to conflict with another country. Similar with the betweenness centrality, the geodesic distance is meaningless in this trade conflict network.

The degree centrality should be significant, given that high outdegree shows that a country is aggressive (like the USA), while the high indegree node may represent a vulnerable country (like Afghanistan). The eigenvector centrality should also be important, since that if a nation has conflict with a conflict-centered nation, it is very likely to be involved in a large-scale trade conflict between opposed alliances.

We first take a look as the centrality scores of the full network, as shown in the following table. As expected, the closeness centrality and betweenness centrality is extremely low. The closeness centrality is even equal to zero. The eigenvector centrality is as large as 0.517. The degree centrality is not that significant,

Degree Centrality	0.153
Closeness Centrality	0
Betweenness Centrality	0.088
Eigenvector Centrality	0.517

Then we see the centrality scores of the network without isolates. The scores are similar with that of the full network. The eigenvector centrality is also the most significant, and the closeness centrality is equal to zero too.

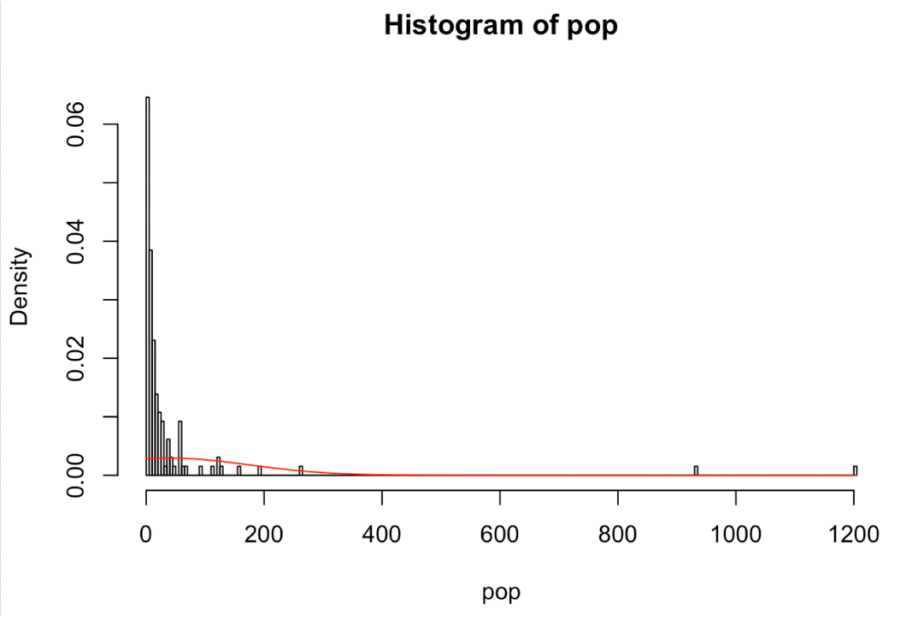
Degree Centrality	0.213
Closeness Centrality	0
Betweenness Centrality	0.179
Eigenvector Centrality	0.505

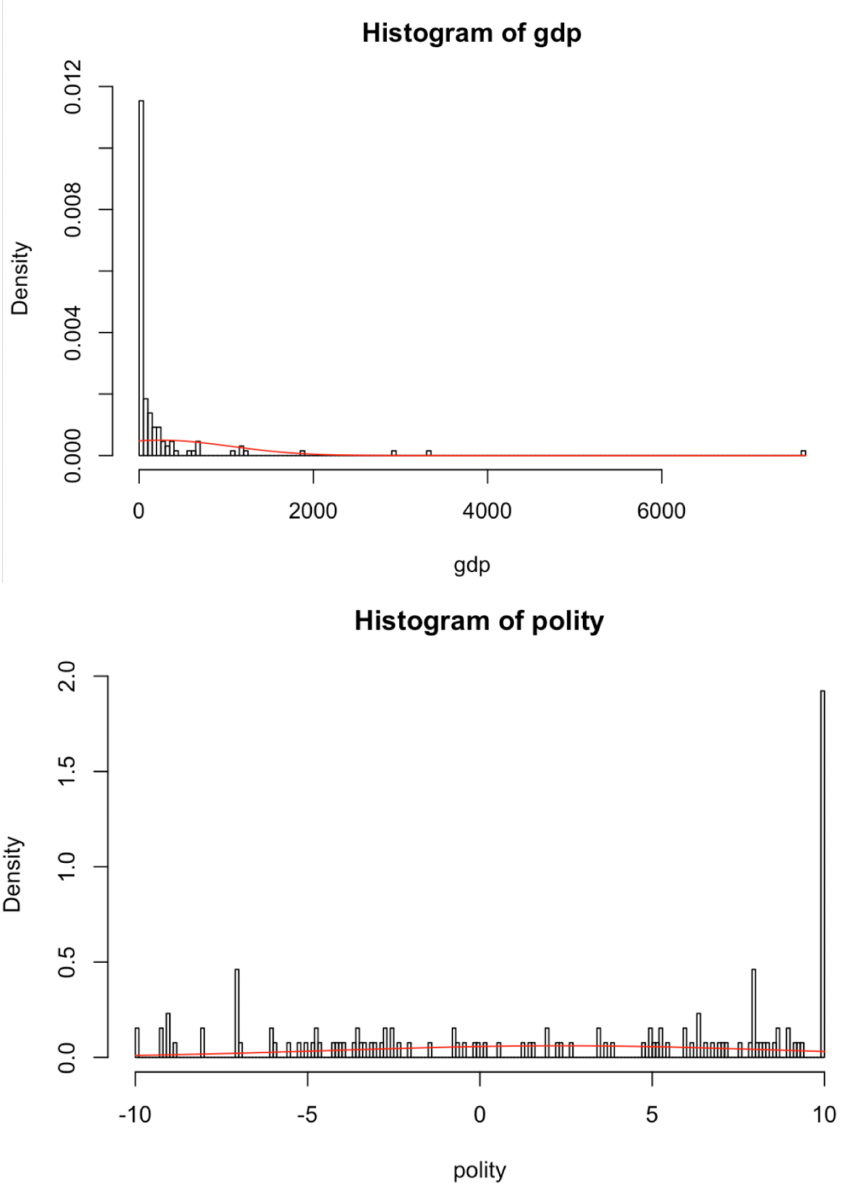
5. Nodal Covariates Analysis with ERGM

There are three nodal covariates: GDP, population and polity score. We would like to analyze their effects by fitting the ERGM model. We want to: 1. Analyze the effect of the nodal covariates and, 2. Check whether there is uniform homophily and differential homophily. We use the network with isolates removed in this section. Intuitively, there should be differential homophily between polity scores, given that different ideology is very likely to lead conflicts. Besides, there might be uniform homophily in GDP, since that large economic entities are more likely to have trade conflicts because of huge amount of interest.

However, all these three covariates are numerical, while checking homophily must set the covariates to be categorical. Thus, we analyze the distribution of the three covariates and find out an appropriate approach to divide the data. We first check out the mean, median, and standard deviation, then fit a normal distribution of them. The measures are as follows:

Covariate	Mean	Median	Standard Deviation
gdp	245.63	28.38	801.03
pop	38.86	9.59	135.48
polity	2.36	3.73	6.59





As can be seen in the previous results, GDP and populations varies a lot. The mean of GDP, 245.62, is much greater than the median 28.38. Same thing happens with population, which has a mean of 38.86 and median 9.59. These results show that there is a huge wealth gap between rich country countries. Besides, there are huge nations with large populations, however most countries are with small populations. Thus, we set the countries which GDP and population to be larger than the sum of mean and standard deviation, which is the top 16%, to be “rich” and “large” in population, otherwise to be “poor” and “small” in population.

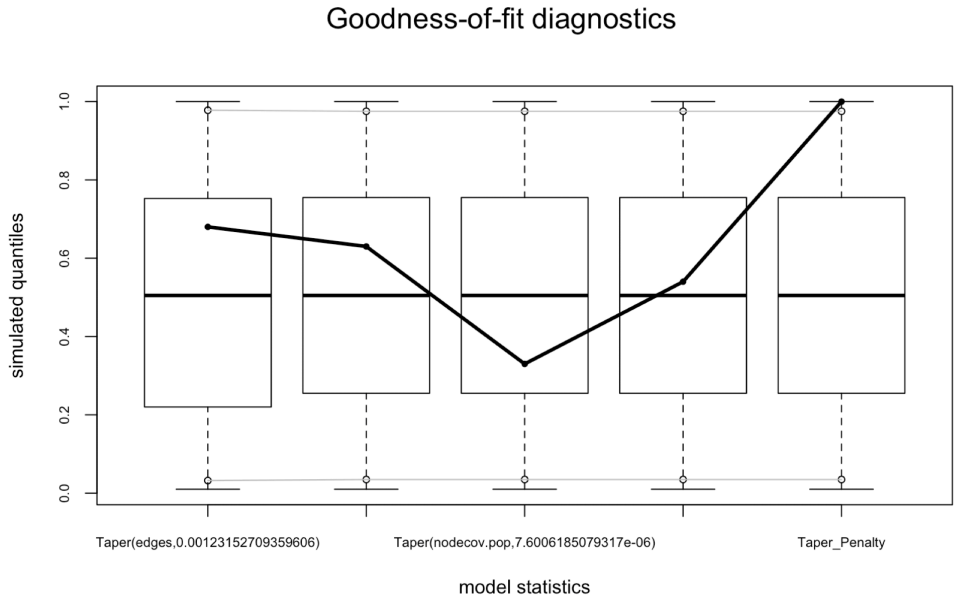
The polity score seems much more uniformly distributed, compared with GDP and popularity. Thus, we set the two categories to be “positive” and “negative” in polity score.

Some interesting facts are: Some countries with full score in polity: United States, United Kingdom, Australia. And some countries with extremely low score: Saudi Arabia with -10, Iraq with -9. And China, -7, seems not too bad.

In the first model, we fit the tapered ERGM model to the three covariates. By observing the MCMC diagnostics, the model converges. As for the goodness of fit, the model fits great in the number of triad census, however not good enough in indegree and outdegree. We perform the summary of the model and the goodness of fit diagnostics plot as follows:

```
==========================
Summary of model fit
==========================

Formula:   conf_no_iso_net ~ Taper(~edges + nodecov("gdp") + nodecov("pop") +
    nodecov("polity"), coef = .taper.coef, m = .taper.center)
<environment: 0x122e1a460>

Iterations:  30 out of 30

Monte Carlo MLE Results:
                 Estimate Std. Error MCMC % z value Pr(>|z|)
edges          -3.8695338  0.1358580      2 -28.482   <1e-04 ***
nodecov.gdp     0.0006879  0.0014082      0   0.488    0.625
nodecov.pop    -0.0026296  0.0043981      1  -0.598    0.550
nodecov.polity -0.0423926  0.0973760      0  -0.435    0.663
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

     Null Deviance: 11354  on 8190  degrees of freedom
 Residual Deviance:  1898  on 8186  degrees of freedom

AIC: 1906    BIC: 1934    (Smaller is better.)
```

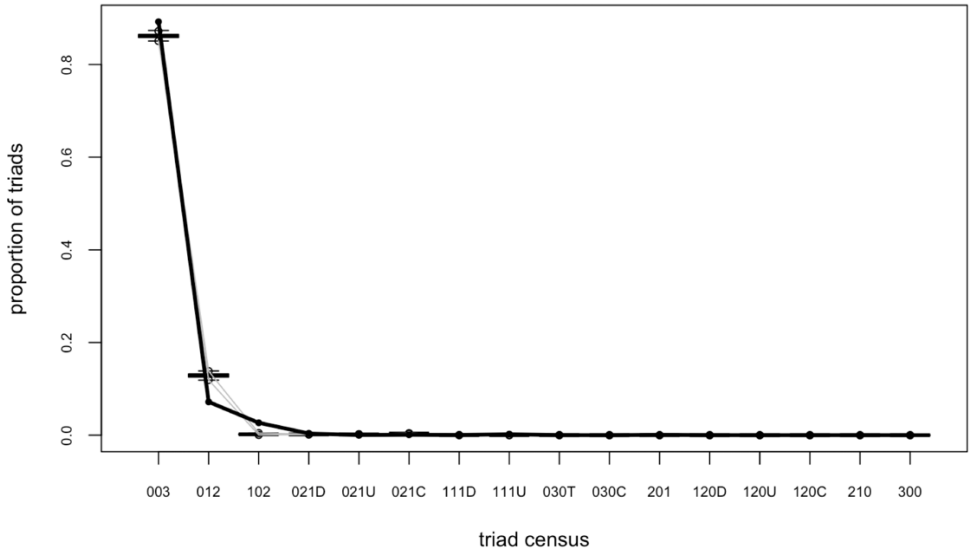
This summary shows that the three covariates, GDP, population and polity do not matter a lot in the tapered ERGM model, although the z-score and p-value say that the conclusion is not that convincing. The result is kind of surprising, or maybe tapered ERGM model is not that good in mining the characteristics of this network. The goodness of fit diagnostics plot shows that the model fits good.

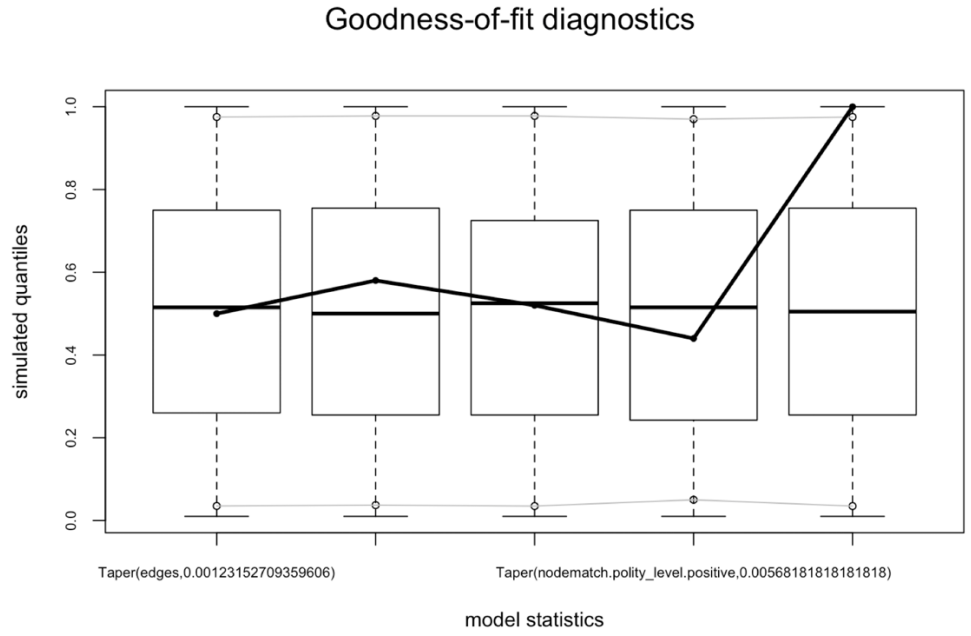


Then we check the homophily. Based on the discussion in the beginning of this section, intuitively there might be differential homophily in polity and uniform homophily in GDP. We fit a tapered ERGM to figure this out:

```
==========================
Summary of model fit
==========================

Formula:   conf_no_iso_net ~ Taper(~edges + nodematch("gdp_level") + nodematch("po
lity_level",
    diff = T), coef = .taper.coef, m = .taper.center)
<environment: 0x1203beea8>

Iterations:  7 out of 30

Monte Carlo MLE Results:
                               Estimate Std. Error MCMC % z value Pr(>|z|)
edges                           -3.0515     0.1547      0 -19.729   <1e-04 ***
nodematch.gdp_level             -0.9316     0.1815      0  -5.133   <1e-04 ***
nodematch.polity_level.negative  0.7955     0.1870      0   4.254   <1e-04 ***
nodematch.polity_level.positive -0.3018     0.2200      0  -1.372     0.17
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

     Null Deviance: 11354  on 8190  degrees of freedom
 Residual Deviance:  1858  on 8186  degrees of freedom

AIC: 1866    BIC: 1894    (Smaller is better.)
```

As can be seen in the summary above, overall there is a differential homophily of trade conflict potential in polity style. The nations with negative polity score are more likely to show conflict towards the positive polity score nations, with a large z-score and small p-value. The countries with positive polity score seem not like to conflict towards nations with negative polity score. This result might because some democratic nations are peaceful countries (isolates in the network), such as Australia (with score 10 in polity). Besides, this model fits pretty well, especially regarding triad censes. The goodness of fit diagnostics plot and trad census goodness of fit are as follows:





6. Triad Census Analysis of the Network with ERGM

Triad census is an important characteristic in directed network. In this section, we consider the network's tendency of forming a triad or not. We first summarize the triad census in the network, then fit a tapered ERGM model to figure out the tendency of have triads. We will also compare the analysis between the full network and the network with isolates removed. We will argue that the isolates can be removed.

First, we take a look of all kinds of triads in the network, the above is the count for the full network, below is the count for the one with isolates removed:

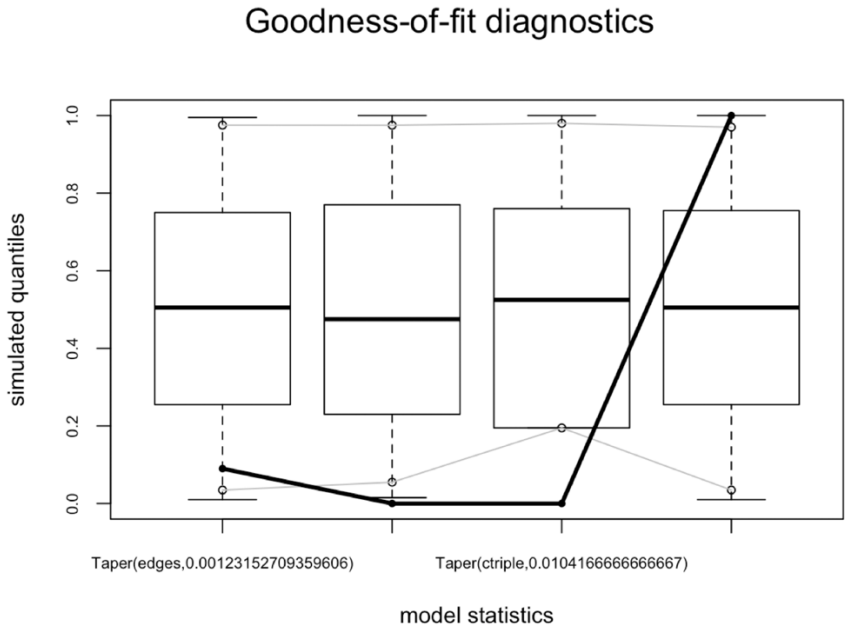
```
> triad.census(conf_net)
        003   012  102 021D 021U 021C 111D 111U 030T 030C 201 120D 120U 120C 210 300
[1,] 338443 13292 4929  439   57  120   51  260   23    0 102   18    4    8  12   2
> triad.census(conf_no_iso_net)
        003  012  102 021D 021U 021C 111D 111U 030T 030C 201 120D 120U 120C 210 300
[1,] 108408 8729 3252  439   57  120   51  260   23    0 102   18    4    8  12   2
```

As can be seen, removing the isolates causes dramatic reduce in 003, 012 and 102, and have no effect to other types of triads. That is, all cyclic triads and transitive triads are maintained after removed the isolated. As we have mentioned in the second section, there are totally 47 transitive triads and 553 cyclic triads in the network.

The network is obviously not balanced, given that there are lots of unbalanced triads in the network, for example, 021D, 021U and so on.

Then we fit tapered ERGM models which includes the number of cyclic triads and transitive triads, and compare the performance. First is the model for the whole network:

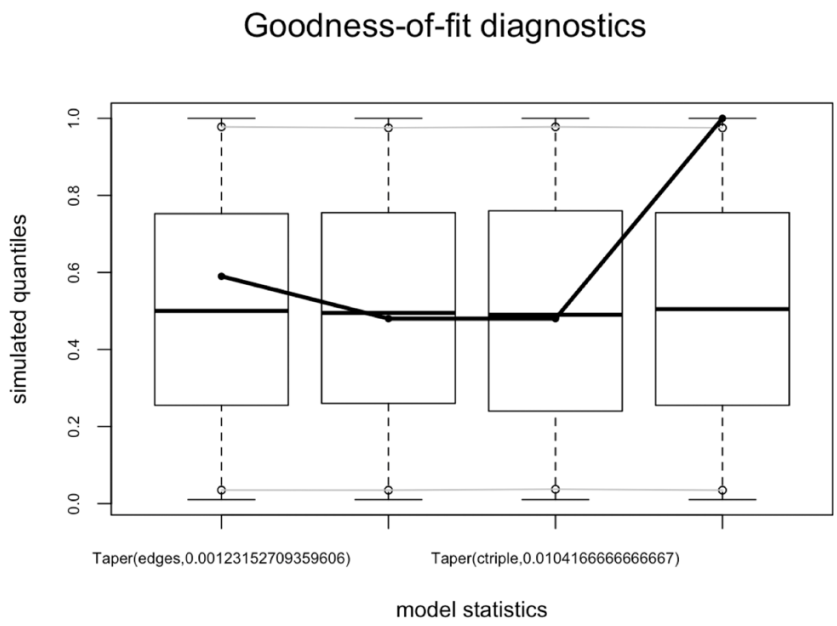
```
==========================
Summary of model fit
==========================

Formula:   conf_net ~ Taper(~edges + ttriad + ctriad, coef = .taper.coef,
    m = .taper.center)
<environment: 0x1204cc678>

Iterations:  8 out of 30

Monte Carlo MLE Results:
        Estimate Std. Error MCMC % z value Pr(>|z|)
edges   -4.66315    0.08841      0 -52.744   <1e-04 ***
ttriple  0.83519    0.05839      1  14.303   <1e-04 ***
ctriple -0.51504    0.20777      0  -2.479   0.0132 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

     Null Deviance: 23248  on 16770  degrees of freedom
 Residual Deviance:  2038  on 16767  degrees of freedom

AIC: 2044    BIC: 2067    (Smaller is better.)
```



As can be seen, there is a strong tendency of forming transitive triads, which is very significant given the large z-score and small p-value. Meanwhile, there is a strong tendency of not forming cyclic triads, which is also quite significant. The goodness of fit is pretty bad.

As for the model for the isolates removed network:

```
==========================
Summary of model fit
==========================

Formula:   conf_no_iso_net ~ Taper(~edges + ttriad + ctriad, coef = .taper.coef,
    m = .taper.center)
<environment: 0x14271b600>

Iterations:  27 out of 30

Monte Carlo MLE Results:
        Estimate Std. Error MCMC % z value Pr(>|z|)
edges   -4.02537    0.08841      0 -45.532   <1e-04 ***
ttriple  0.69714    0.06252      0  11.150   <1e-04 ***
ctriple -0.28476    0.21272      0  -1.339    0.181
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

     Null Deviance: 11354  on 8190  degrees of freedom
 Residual Deviance:  1792  on 8187  degrees of freedom

AIC: 1798    BIC: 1819    (Smaller is better.)
```



The model also suggests a significantly strong tendency of forming transitive triads, as well as not forming cyclic triads, with small p-value supported. Besides, the model fits pretty well by observing the goodness-of-fit diagnostics plot. We can therefore draw two conclusions: First, the trade conflict network has a strong tendency of forming transitive triads, as well as not forming cyclic triads. Second, removing isolates does not affect the analysis of the conflicts. Ignoring these conflict-free nodes helps better analyze the conflicts.

7. Explore ERGM with Other Parameters

In this we consider the four terms in tapered ERGM model: geometrically weighted dyadwise shared partner distribution, geometrically weighted edgewise shared partner distribution, geometrically weighted in-degree distribution and geometrically weighted out-degree distribution. This idea comes because the indegree, outdegree and partners in this trade conflict network is crucial. The outdegree shows the aggressiveness and willing to participate in conflicts. The indegree show that whether a country is against the world. The dyadwise as well as edgewise shared partners shows who is you ally and who is your enemy. In real worlds, allies usually fight together, whether they are really willing to do so. For example, a few days ago, the United States forbad Huawei from being in the US market. Later, some other countries, including the United Kingdom, forbad Huawei as well, showing the support to their ally.

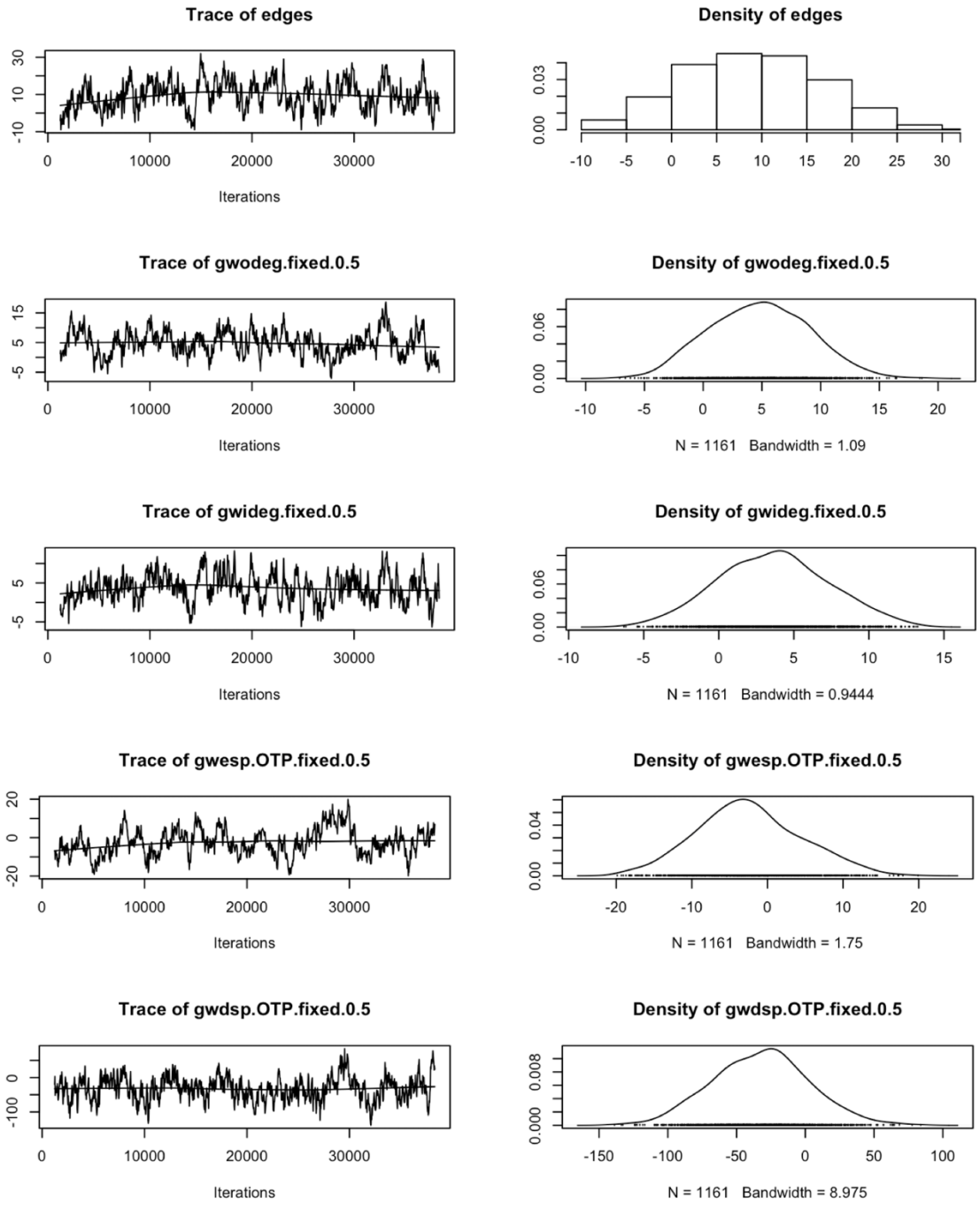
The summary of the model is as follows:

```
Summary of model fit
=========================

Formula:   conf_no_iso_net ~ Taper(~edges + gwodegree(0.5, fixed = TRUE) +
    gwidegree(0.5, fixed = TRUE) + dgwesp(0.5, fixed = TRUE) +
    dgwdsp(0.5, fixed = TRUE), coef = .taper.coef, m = .taper.center)
<environment: 0x1250e0270>

Iterations:  30 out of 30

Monte Carlo MLE Results:
                   Estimate Std. Error MCMC % z value Pr(>|z|)
edges              -4.61028    0.40504     20 -11.382  < 1e-04 ***
gwodeg.fixed.0.5   -0.91193    0.44940     19  -2.029  0.04244 *
gwideg.fixed.0.5    1.44804    0.51389      8   2.818  0.00483 **
gwesp.OTP.fixed.0.5  1.27052    0.28994     22   4.382  < 1e-04 ***
gwdsp.OTP.fixed.0.5  0.03706    0.05110     21   0.725  0.46825
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

     Null Deviance: 11354  on 8190  degrees of freedom
 Residual Deviance:  1818  on 8185  degrees of freedom

AIC: 1828    BIC: 1863    (Smaller is better.)
```

As can be seen in the result, the geometrically weighted in-degree distribution and the geometrically weighted edgewise shared partner distribution have strong effect to form trade conflict within the network, which is quite significantly given the tiny p-value.

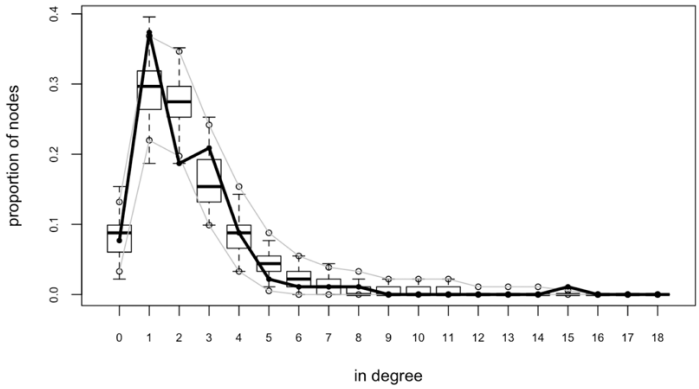
Besides, geometrically weighted out-degree distribution has a strong negative effect to form trade conflict, which is also quite significant with a p-value of 0.04.

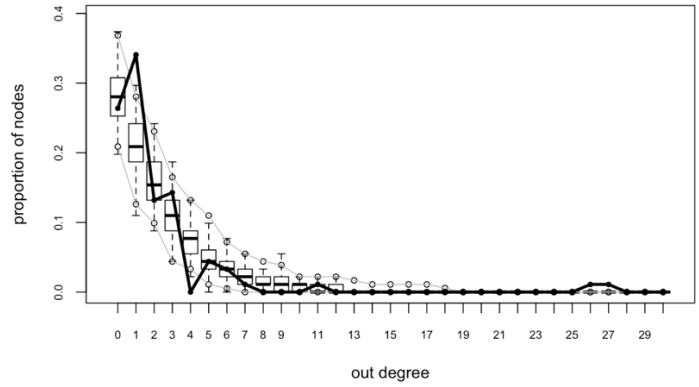
The MCMC diagnostics result shows that the fitting converge:

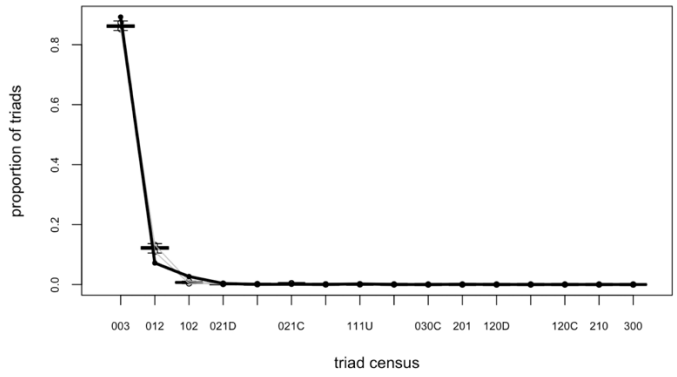


The goodness-of-fit plots shows that the model fits pretty well in triad census (even though

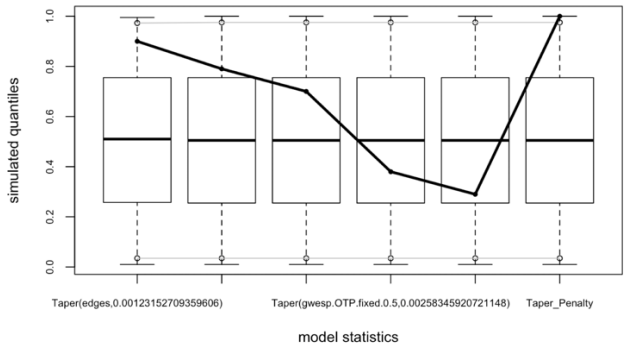
we did not add the ctriple and ttriple terms), fits so-so in indegree and outdegree.







Goodness-of-fit diagnostics



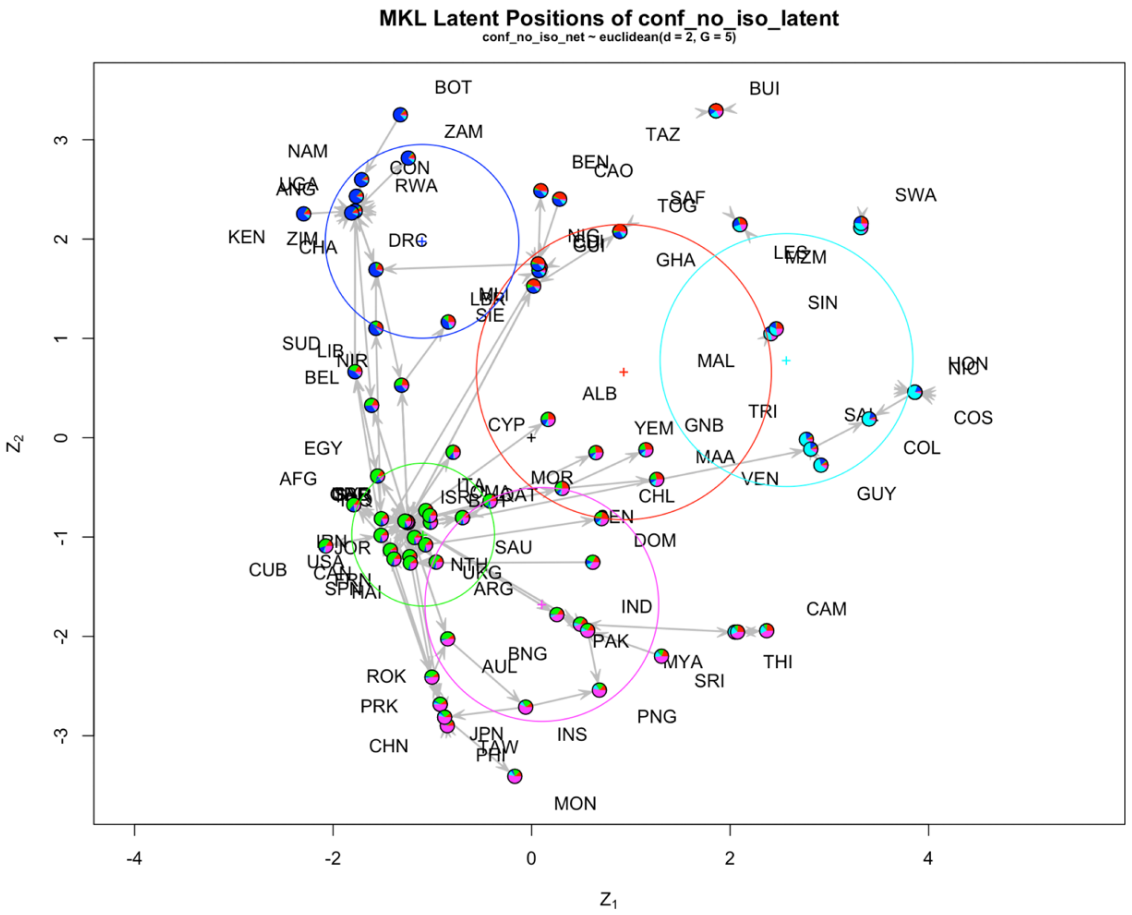
8. Latent Position Models for Clustering

In order to see whether the network has intrinsic of clustering, we fit some latent position models. We select the dimension to be 2, and try different numbers of groups. Then we check the Bayesian Information Criterion (BIC) to find evidence of clustering. The BIC scores of models with different number of groups are as follows:

Number of groups	Latent space/clustering BIC	Overall BIC
1	753.0556	2011.47
2	721.3381	2013.147
3	718.9186	2015.472
4	682.9981	1981.276
5	626.7312	1934.979
6	629.6177	1932.36

There is strong evidence for clustering, given that the one-group model has large BIC values. The BIC values go down until the 5-groups model, then starts rising at the 6-groups model. We may draw a conclusion that 5-groups is strongly supported. It is quite interesting that, there are 7 continents in the world: Asia, Africa, North America, South America, Antarctica, Europe, and Australia. Except Antarctica, and Australia, which is a peaceful country occupying the whole continent, there are actually 5 continents. Same number with the number of groups in the trade conflict network. But are the conflict groups coincidence with the continents?

We then take a look at the clustering plots to see how the groups are defined.



As can be seen in the network, the United States, as well as its allies, the United Kingdom, France, Canada, Israel are both in the green group. Their enemy in the Gulf War, which is the largest war in 1990s, Iraq, is also in the green group. Other countries which participate in the Gulf War like Qatar, is also in the green component. We may conclude that the trade conflict along with the Gulf War is represents by the green sub-network. In the pink network, some important Asian countries appears, like China, Japan and India. These results support the clustering by the latent model.

Acknowledgement

At the end of the final report, I would like to extend my great thank to Professor Handcock and my TA Bart. As an Electrical and Computer Engineering major student, this is the first statistics course I have taken in my graduate study at UCLA. Professor Handcock's lectures attracted me a lot. This course actually brought my knowledge in social network analysis from zero to one, which is quite an important first step. Bart also helped me a lot in the office hours, with great patience to a non-statistics student. I really appreciate their help during the whole quarter.