Project Report

CEN-502 (Computer Systems II)

Project - 1

Submitted By:

Group no: 28

Ashish Aggarwal

Akash Goyal

Mengyuan Zhang

Introduction

In these sea ice concentration anomaly data, we have the values of the 304*448 cells in 27 years. Each cell (x, y) represents the percentage of deviation in ice concentration from the 27-year average for a given week. We import these big data into Java program to deal with.

In the 27 years analysis, we get the average value of each cell over 27*52 week, and in the separate phase time analysis, we do the same thing in each 9 years interval. After get the average value in a given time interval, we can deal with those big data without a time dimension consideration.

With a given threshold pearson correlation coeffecient r, we can generate a connection (u, v) if pearson correlation coeffecient of vertex u and vertex v is greater than threshold r. We store those connection and represent the graph by adjacency list. By comparison of setting a different value of r, say r =0.9 and r = 0.95, it is easy to see that with a greater threshold r, the connection of the graph will decrease a bit.

After the generation of graph, we can easily calculate the degree distribution and draw a historgram to describe it. The supernodes, cluster coefficient and characteristic path length of the graph and what's in random situation could also be obtained by the formula given in the project context.

By now, after achieving all the jobs required in the tasks. We are able to use the data mining approach to analyse the Sea ice concentration in northern hemisphere.

Answers for First Three Question

1. Construct a correlation-based graph $G_r = (V_r, E_r)$ for the complete sea ice anomaly dataset for each correlation threshold r ϵ { 0.9, 0.95 }

   a. When r = 0.9:

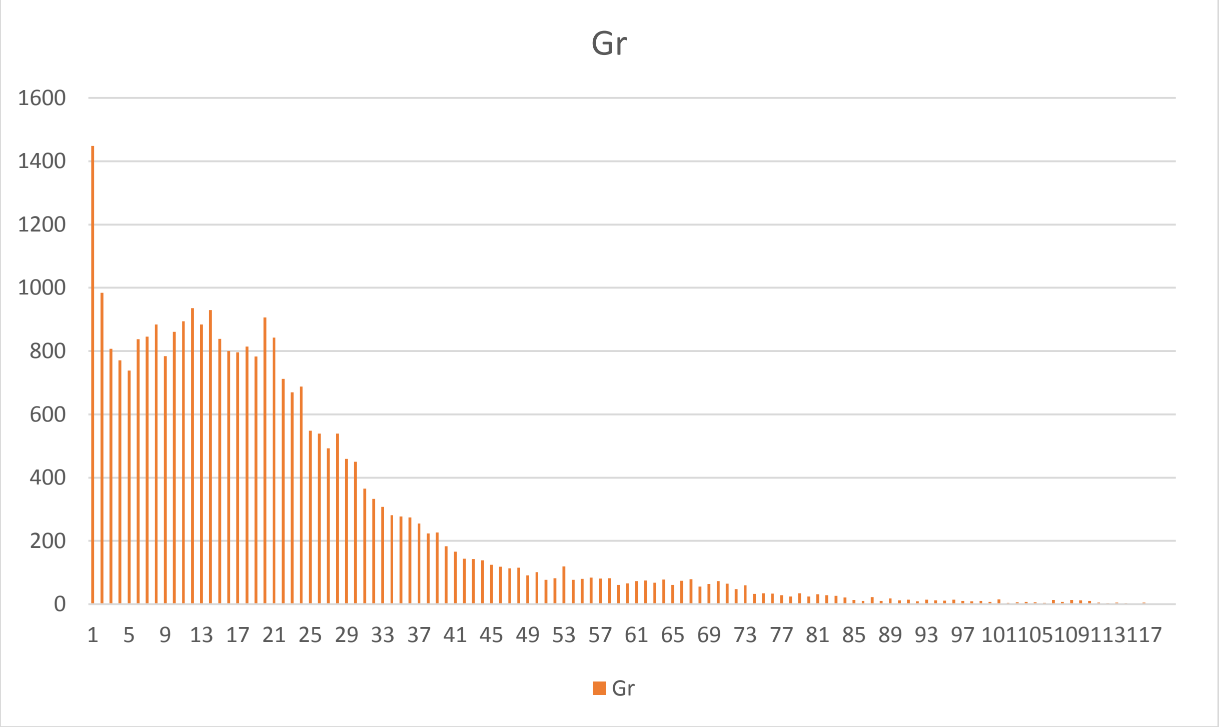

Figure 1. histogram of the degree distribution for r =0.9 (complete graph)

The number of the supernodes: 2854 (threshold 45)

i. Cluster coefficienct: $\gamma(G_r) = 0.2932332$

   Characteristics path length: L((G_r) = 15.7823016

   Random Clustering Coefficient: $\gamma(G_{random})$ = 1.3609763e-4

   Random Characteristic path length: L((G_{random}) = 5.0515376

ii. $\gamma \gg \gamma_{random}$, and L$\gg$ L_{random}

b. When r = 0.95

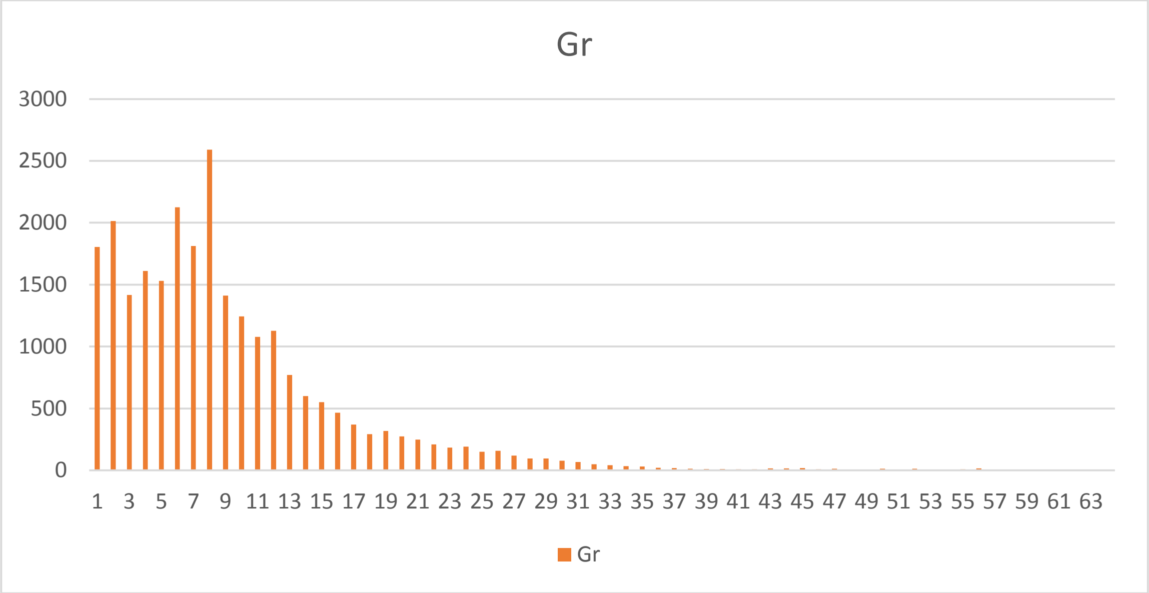

Figure 2. histogram of the degree distribution for r =0.95 (complete graph)

The number of the supernodes: 1029 (threshold 25)

i. Cluster coefficiency: $\gamma(G_r) = 0.24303128$

Character path length L((G_r) = 32.69104522

Random Clustering Coefficient: $\gamma(G_{random})$ = 4.53658e-5

Random Characteristic path length: L((G_{random}) = 10.10307528

ii. $\gamma \gg \gamma_{random}$, and L$\gg$ L_{random}

2. Split the dataset into three equal parts of nine years each.

a. When r = 0.9

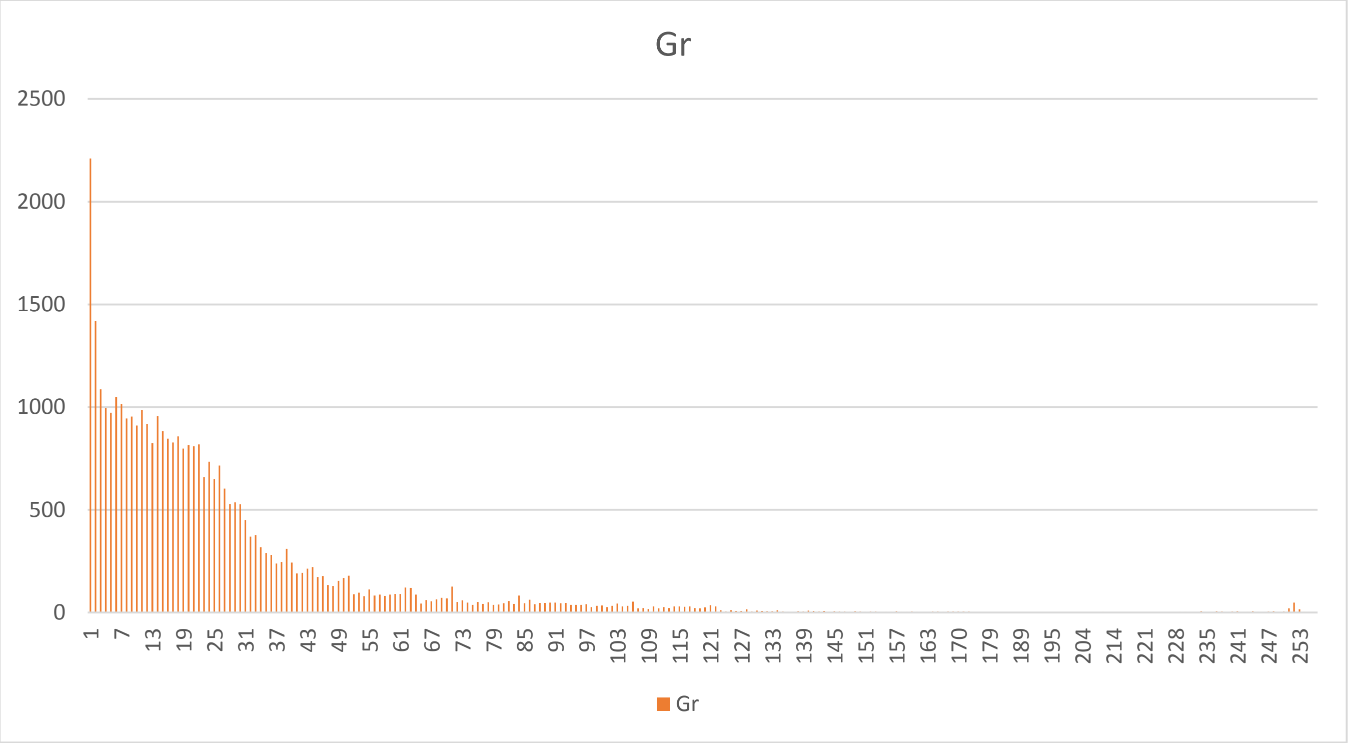

Figure 3. histogram of the degree distribution for r =0.9 (first nine year graph)

Identify the supernodes: 4328 (threshold 50)

i. Cluster coefficiency: $\gamma(G_r) = 0.36$

Character path length L((G_r) = 32.6289

Random Clustering Coefficient: $\gamma(G_{random})$ = 2.76e-5

Random Characteristic path length: L((G_{random}) = 18.42

ii. $\gamma \gg \gamma_{random}$, and L$\gg$ L_{random}

)

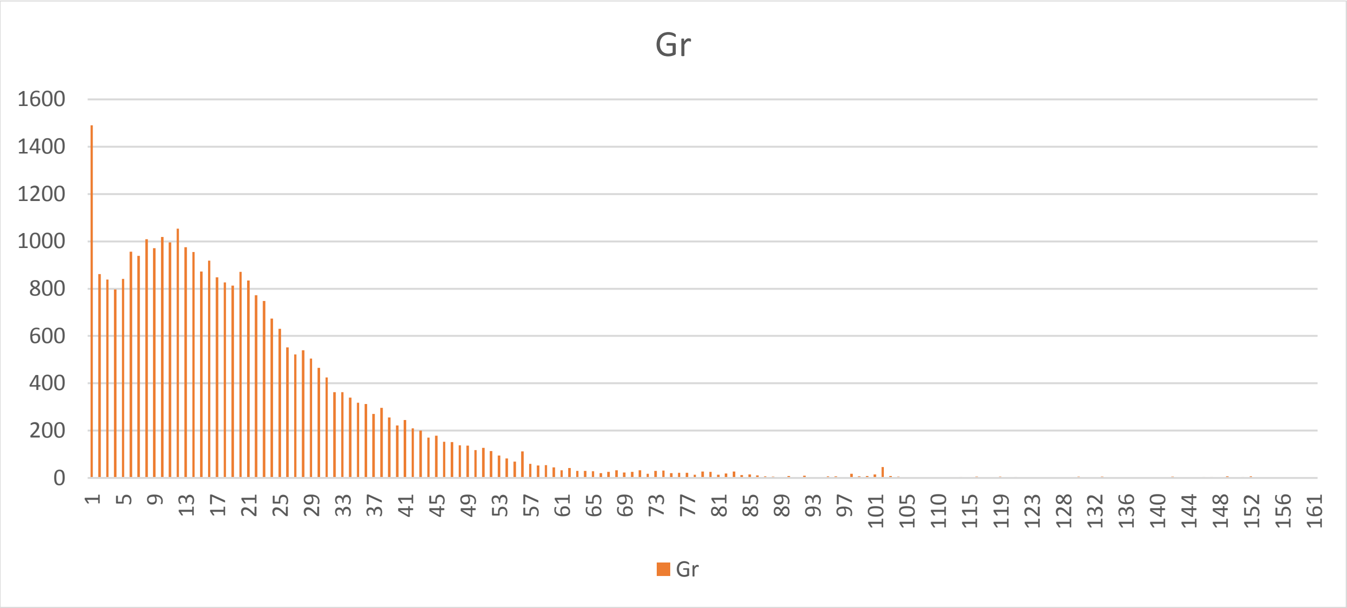

Figure 4. histogram of the degree distribution for r =0.9 (mid-nine year graph)

Identify the supernodes: 1727 (threshold 50)

i. Cluster coefficienct: $\gamma(G_r) = 0.268126024$

Characteristics path length: L((G_r) = 14.4682167

Random Clustering Coefficient: $\gamma(G_{random})$ = 2.8384635E-5

Random Characteristic path length: L((G_{random}) = 3.5365658

ii. $\gamma \gg \gamma_{random}$, and L≫ L_{random}

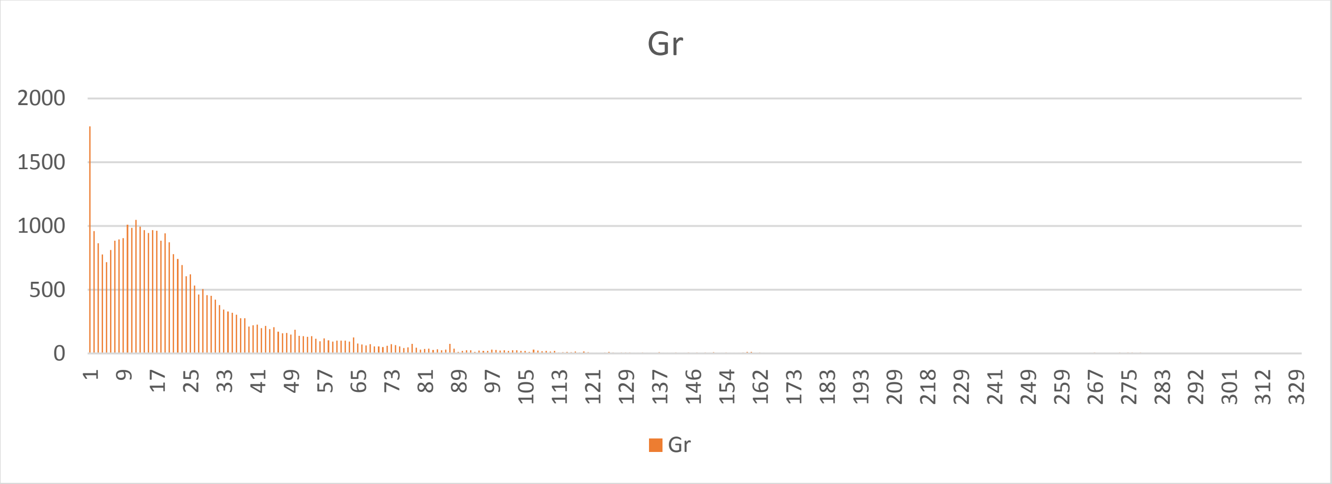

Figure 5. histogram of the degree distribution for r =0.9 (last nine year graph)

Identify the supernodes: 4033 (threshold 50)

i. Cluster coefficiency: $\gamma(G_r) = 0.42$

Character path length L((G_r) = 33.5349

Random Clustering Coefficient: $\gamma(G_{random})$ = 1.96e-5

Random Characteristic path length: L((G_{random}) = 20.34

ii. $\gamma \gg \gamma_{random}$, and L$\gg$ L_{random}

b. When r = 0.95

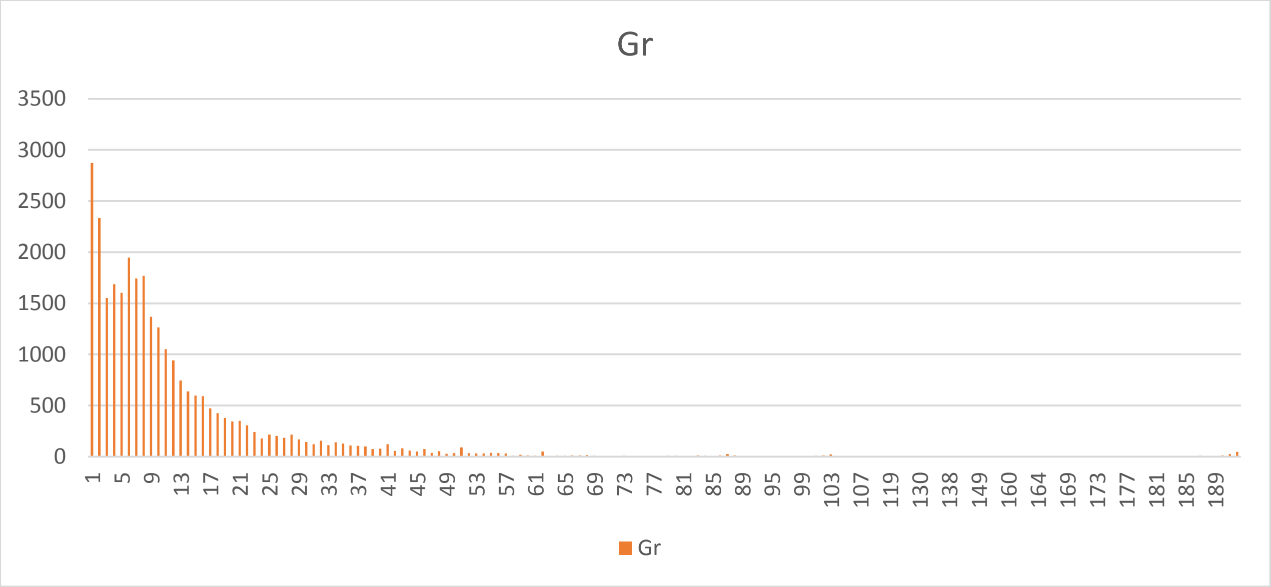

Figure 6. histogram of the degree distribution for r =0.95 (first nine year graph)

Identify the supernodes: 3137 (threshold 27)

i. Cluster coefficienct: $\gamma(G_r) = 0.12925$

Characteristics path length: L((G_r) = 5.1367345

Random Clustering Coefficient: $\gamma(G_{random})$ = 1.843634E-5

Random Characteristic path length: L((G_{random}) = 9.753752

ii. $\gamma \gg \gamma_{random}$, and L$\gg$ L_{random}

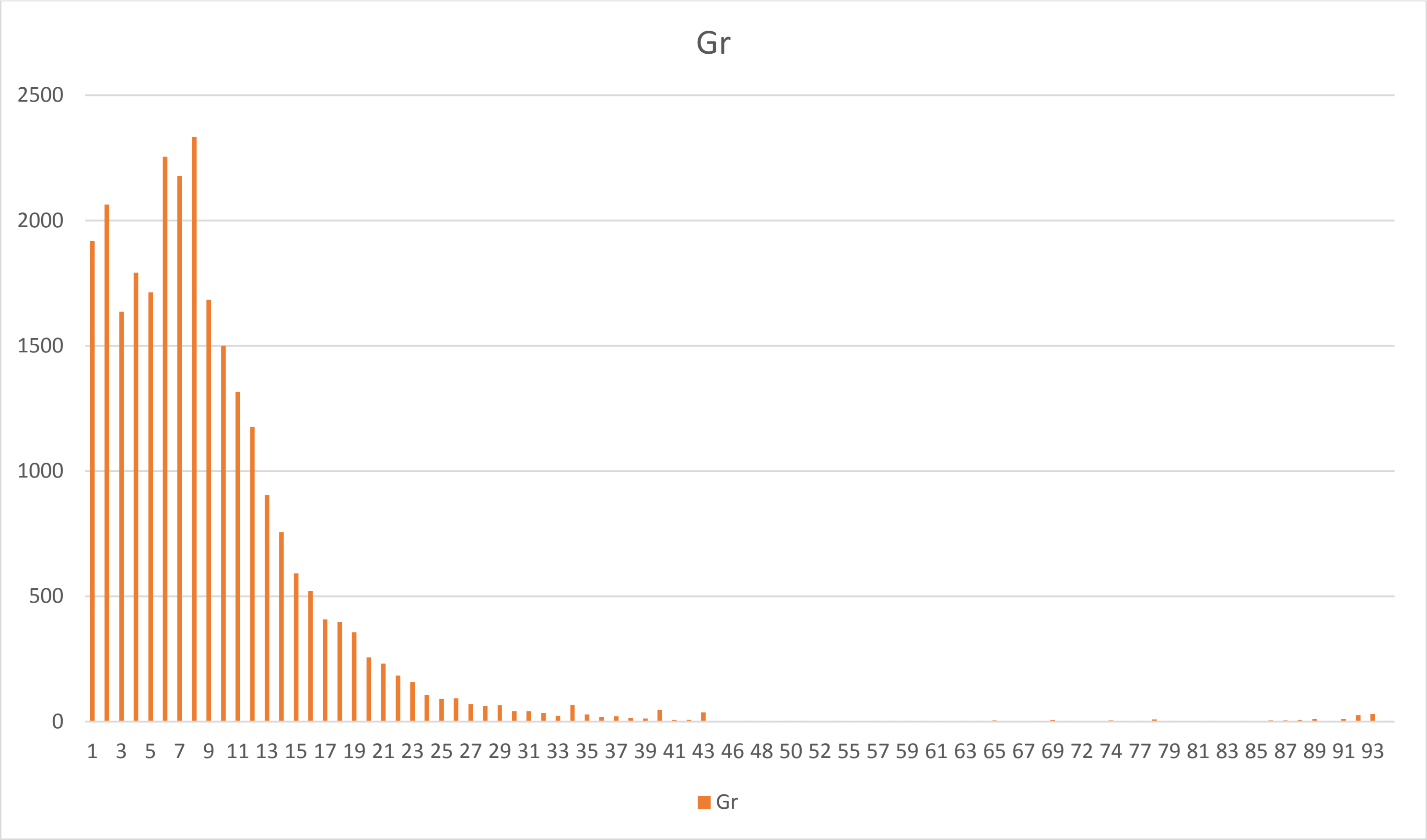

Figure 7. histogram of the degree distribution for r =0.95 (mid- nine year graph)

Identify the supernodes: 718 (threshold 27)

iii. Cluster coefficienct: $\gamma(G_r) = 0.026922$

Characteristics path length: L((G_r) = 15.185367

Random Clustering Coefficient: $\gamma(G_{random})$ = 0.09436E-5

Random Characteristic path length: L((G_{random}) = 0.057575

iv. $\gamma \gg \gamma_{random}$, and L≫ L_{random}

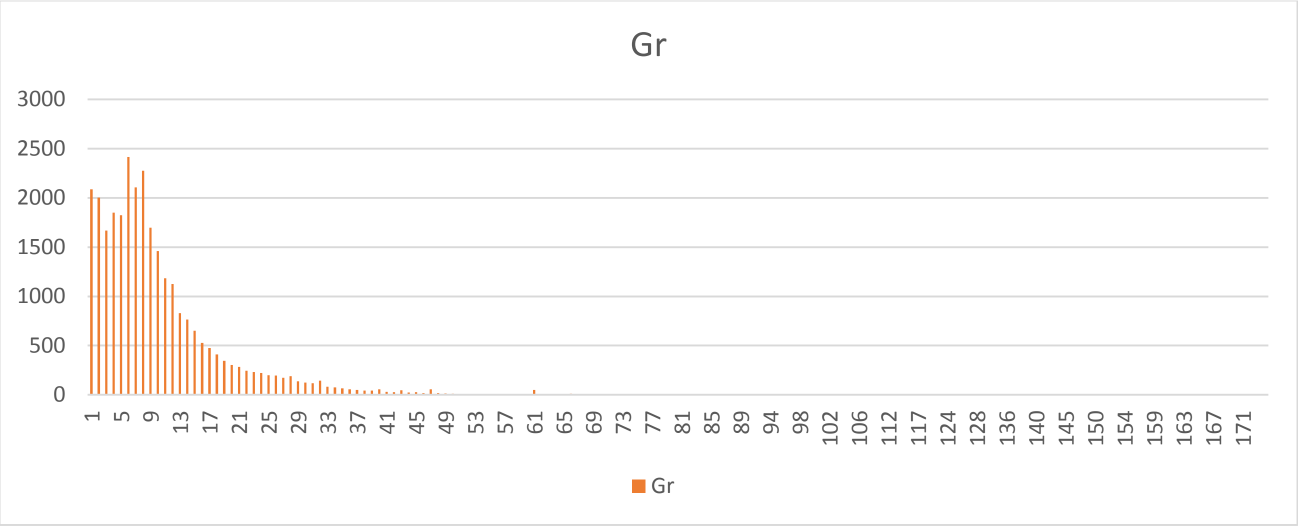

Figure 8. histogram of the degree distribution for r =0.95 (last nine year graph)

Identify the supernodes: 1900 (threshold 27)

v. Cluster coefficienct: $\gamma(G_r) = 0.05693$

Characteristics path length: L((G_r) = 1.1867

Random Clustering Coefficient: $\gamma(G_{random})$ = 0.008436E-5

Random Characteristic path length: L((G_{random}) = 0.557573

vi. $\gamma \gg \gamma_{random}$, and L$\gg$ L_{random}

3. Consider a time lag of s∈ {1,2,3,4}.

a. When r = 0.9, time lag =1

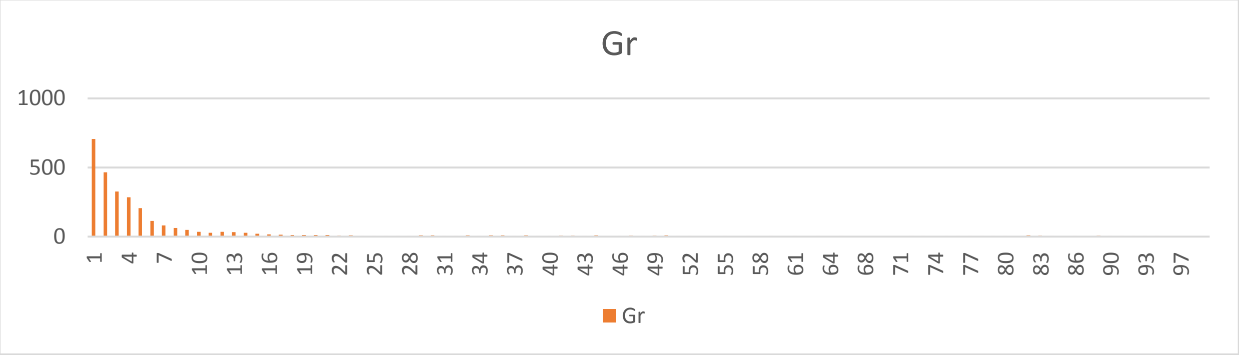

Figure 9. . histogram of the degree distribution for r =0.9 (time lag = 1 week)

Number of supernodes: 419 (threshold 13)

i. Cluster coefficienct: $\gamma(G_r) = 0.02692$

Characteristics path length: L((G_r) = 15.185367

Random Clustering Coefficient: $\gamma(G_{random})$ = 1.488436E-5

Random Characteristic path length: L((G_{random}) = 4.557575

ii. $\gamma \gg \gamma_{random}$, and L$\gg$ L_{random}

b. When r = 0.95, time lag =1

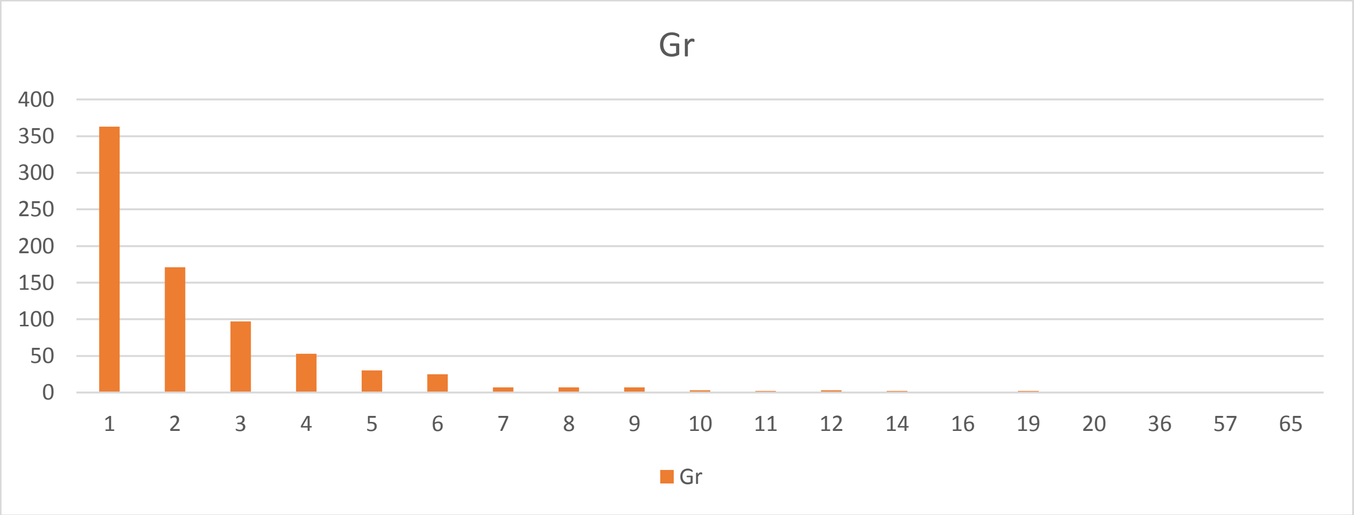

Figure 10. histogram of the degree distribution for r =0.95 (time lag = 1 week)

Number of supernodes: 146 (threshold 3)

i. Cluster coefficienct: $\gamma(G_r) = 0.02692$

Characteristics path length: L((G_r) = 15.185367

Random Clustering Coefficient: $\gamma(G_{random})$ = 1.488436E-5

Random Characteristic path length: L((G_{random}) = 4.557575

ii. $\gamma \gg \gamma_{random}$, and L$\gg$ L_{random}

c. When r = 0.9, time lag = 2 week

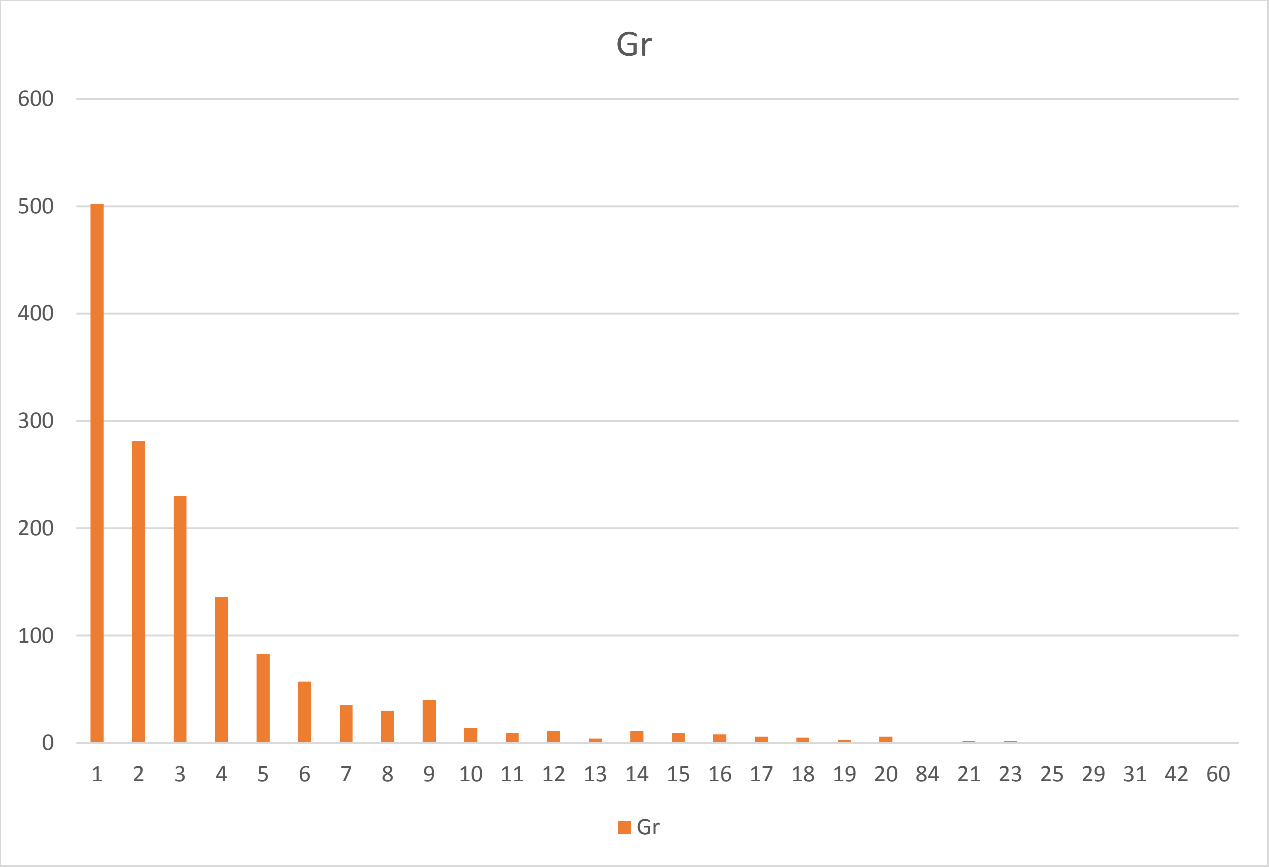

Figure 11. histogram of the degree distribution for r =0.9 (time lag = 2 week)

Number of supernodes: 477 (threshold 3)

i. Cluster coefficienct: $\gamma(G_r) = 0.02692$

Characteristics path length: L((G_r) = 15.185367

Random Clustering Coefficient: $\gamma(G_{random})$ = 1.488436E-5

Random Characteristic path length: L((G_{random}) = 4.557575

ii. $\gamma \gg \gamma_{random}$, and L$\gg$ L_{random}

d. When r = 0.95, time lag = 2 week.

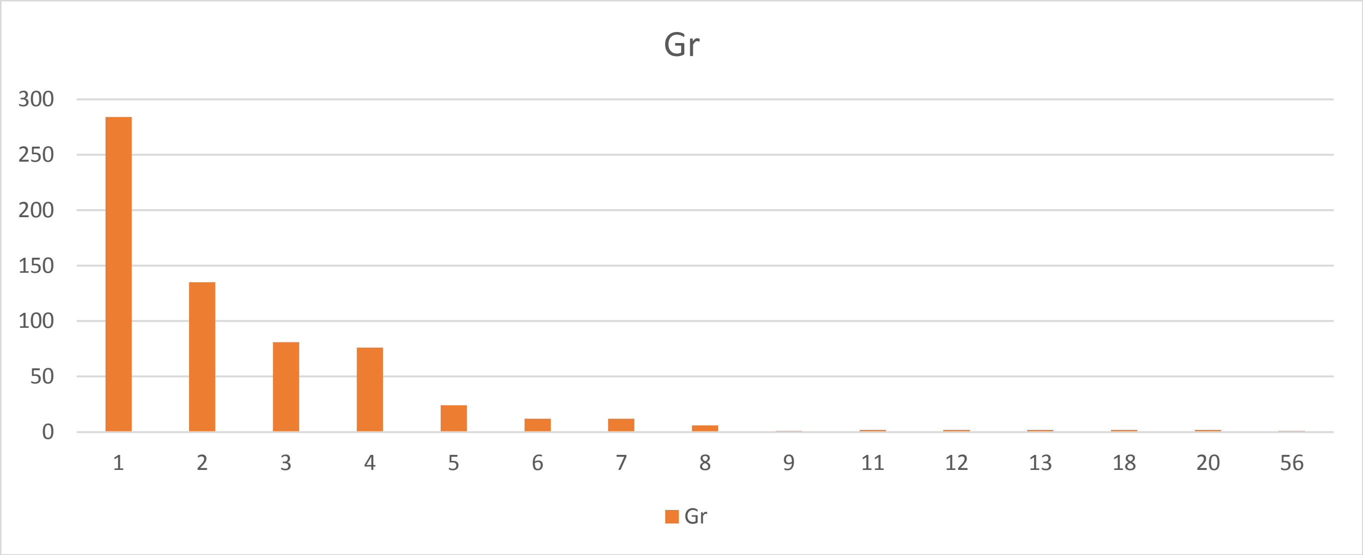

Figure 12. histogram of the degree distribution for r =0.95 (time lag = 2 week)

Number of supernodes: 142 (threshold 3)

i. Cluster coefficienct: $\gamma(G_r)$ = 0.02692

Characteristics path length: L((G_r) = 15.185367

Random Clustering Coefficient: $\gamma(G_{random})$ = 1.488436E-5

Random Characteristic path length: L((G_{random}) = 4.557575

ii. $\gamma \gg \gamma_{random}$, and L$\gg$ L_{random}

e. When r = 0.9, time lag = 3

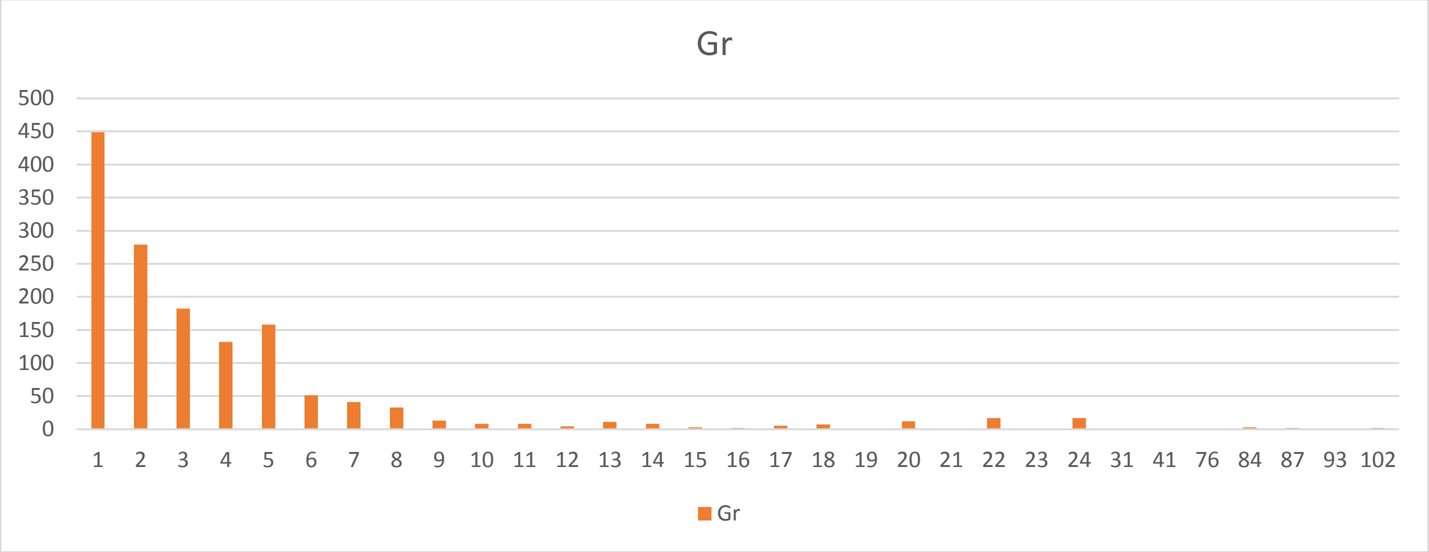

Figure 13. histogram of the degree distribution for r =0.9 (time lag = 3 week)

Number of supernodes: 544 (threshold 3)

i. Cluster coefficienct: $\gamma(G_r) = 0.02692$

Characteristics path length: L((G_r) = 15.185367

Random Clustering Coefficient: $\gamma(G_{random})$ = 1.488436E-5

Random Characteristic path length: L((G_{random}) = 4.557575

ii. $\gamma \gg \gamma_{random}$, and L$\gg$ L_{random}

f. When r = 0.95, time lag = 3

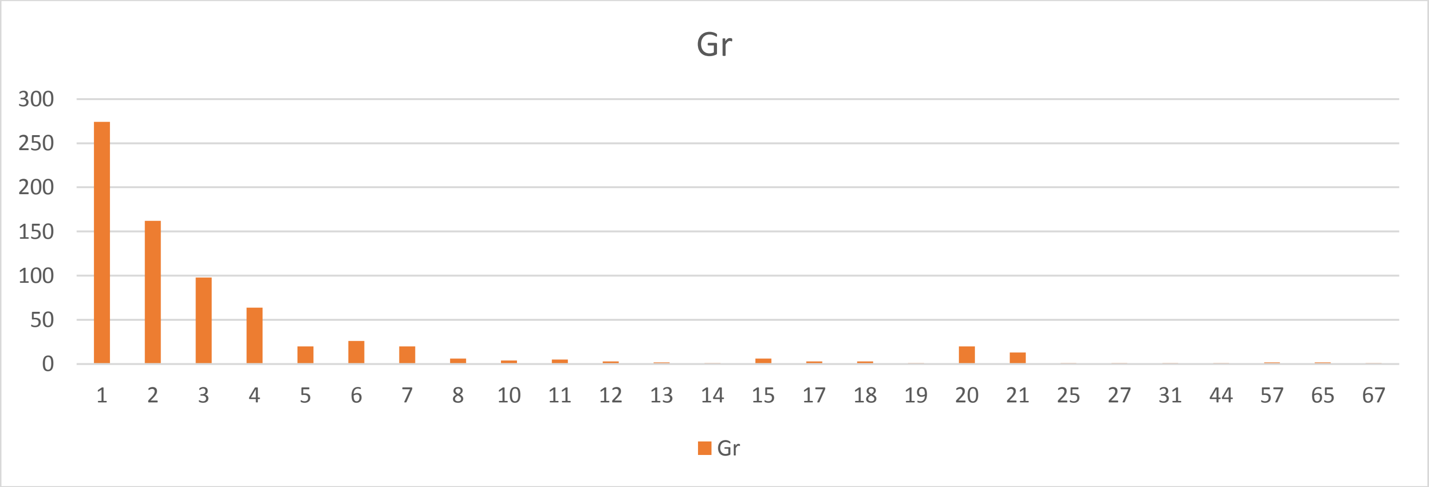

Figure 14. histogram of the degree distribution for r =0.95 (time lag = 3 week)

Number of supernodes: 206 (threshold 25)

i. Cluster coefficienct: $\gamma(G_r) = 0.03692$

Characteristics path length: L((G_r) = 15.185367

Random Clustering Coefficient: $\gamma(G_{random})$ = 1.488436E-5

Random Characteristic path length: L((G_{random}) = 4.557575

ii. $\gamma \gg \gamma_{random}$, and L$\gg$ L_{random}

g. When r = 0.9, time lag = 4 week

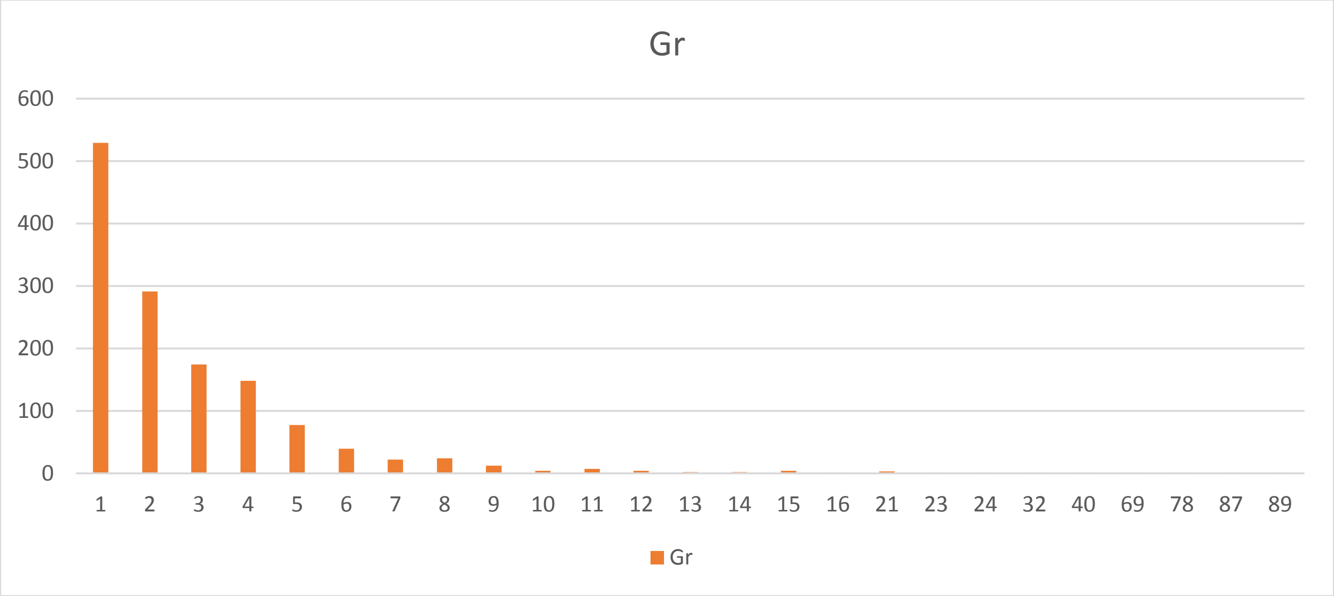

Figure 15. histogram of the degree distribution for r =0.9 (time lag = 4 week)

Number of supernodes: 357 (threshold 3)

i. Cluster coefficienct: $\gamma(G_r) = 0.021923$

Characteristics path length: L((G_r) = 1.153674

Random Clustering Coefficient: $\gamma(G_{random})$ = 0.008436E-5

Random Characteristic path length: L((G_{random}) = 0.5757523

ii. $\gamma \gg \gamma_{random}$, and L$\gg$ L_{random}

h. When r =0.95, time lag = 4 week

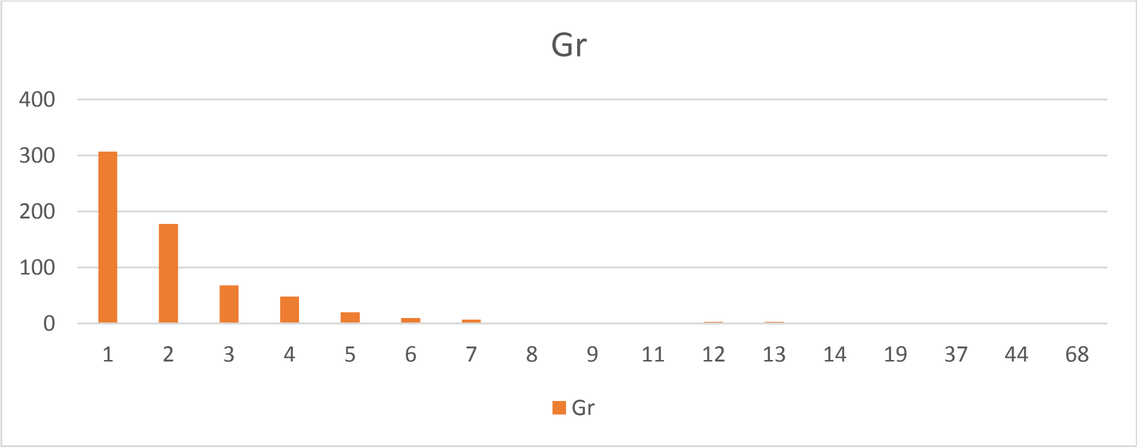

Figure 16. . histogram of the degree distribution for r =0.95 (time lag = 4 week)

Number of supernodes: 100 (threshold 3)

i. Cluster coefficienct: $\gamma(G_r) = 0.03691$

Characteristics path length: L((G_r) = 1.18453652

Random Clustering Coefficient: $\gamma(G_{random})$ = 0.018446E-5

Random Characteristic path length: L((G_{random}) = 0.5572434

ii. $\gamma \gg \gamma_{random}$, and L$\gg$ L_{random}

Representation of the graph

We preferred the adjacency list over the adjacency matrix for the representation of graph. Since the adjacency list is O(m+n) whereas adjacency matrix is O(n2), it was obvious choice.

We are maintaining the adjacency list in the form of Map<Integer, List<Integer>>. The size of Map is 66129, corresponding to total number of vertices (which acts as key of the Map). And for every vertex there is list of Vector values which forms an edge with the corresponding key vertex.

Optimizations made in the graph

We faced many challenges in implementing this project. The main bottleneck was managing the big data.

1. The very first challenge we faced was ‘Out of heap space error’ – The default heap space of JVM i.e. 512 MB. We increased it to 2048 MB.
2. While reading files, we used JAVA API named ‘ByteBuffer’. This library helps in conversion between the Little-Endian data form to the Big-Endian data form.
3. The next major challenge was choosing the optimal data structure for storing the data. Optimal in the sense, that can give us the fastest calculations.
    a. First data structure we used was MAP. Map <Integer, float [ ] [ ]>, we were creating mapped list where week number was our key and we were storing the data in the form of 2-Dimensional array. But the processing time for this was too large, of about 550 hrs.
    b. Since processing 1-Dimensional array is faster than 2-Dimensional array, we converted the 2-Dimensional data into 1-Dimensional data. We used List<float[ ]> , created a list of floating arrays. The size of one float array was 66129, corresponding to the number of data points after the deletion of vectors values {168 (these are the land values), 157 (these are unknown values)} and the number of lists were 1404 corresponding to the number of weeks. This was taking about 200 hrs for execution.
    c. Then we optimized the structure further, we interchanged the size of array and number of lists. Now the size of one float array is 1404, corresponding to the number of weeks and there are 66129 lists in

total, corresponding to the number of data points after the deletion of vectors values {168, 157}. The final execution time is about 1.5 hrs.

4. During read files multithreading feature of Java is exploited, where every thread is reading different files simultaneously to speed up the process of reading data.
5. While calculating the Pearson Correlation Coefficient we eliminated the values of Sxx = 0 to increase efficiency. Since in the formula Sxx is in denominator and value 0 for that would have given us undefined values for Pearson Correlation Coefficient, we were able to eliminate so.

The worst-case bound of the algorithm

1. Worst- case time:

In this algorithm, we cost O(n) time to read and store each file. To get the connection in all cells, we spent O(n^2) time, and do Breadth First Search in O（m+n）. In all, at worst case, it cost O(n^2).
Where, m = number of edges & n = number of vertices.

2. Worst-case bound on space:

We have used data structure List<float[ ]> to represent the complete data set. To store all those data, we spend totally O(K*n) to store the data, in which n = number of vertices and k = number of weeks.

Summary of the report

1. Does the sea ice concentration data represent a scare-free model with small world property？

Now We get the number of supernodes for complete graph in 27 years and separate phase graph in 9 years when r =0.9 and r=0.95 respectively. For example, In the complete graph when r=0.9, the average degree is 10, we define a supernode which has the degree greater than 45, and we get 2854 supernodes which only 2% of the total nodes. In fact, all of those statistics shows sea ice concentration graph has very few nodes with large number of degrees. So we can safely conclude the sea ice concentration network matches the scare free network which is not only stable but also efficient to pass on information.

To consider in other side, In the complete 27 years data for r = 0.95, we get the cluster coeffecient which is $\gamma = 0.243$, $\gamma_{random} = 4.536\mathrm{E} - 5$, and the characteristic path length L=32.69104522, $L_{random} = 10.103$. We can easily tell $\gamma \gg \gamma_{random}$, and L$\gg$ L_{random}. The similar result could come when r = 0.9 or when

we analyse the separate time phase data. So we find sea ice concentration matches the scale-free model with a small world property.

2. What change does the sea ice concentration have by analysing the separate time phase.

Now we compare the graph in each separate time phase when r = 0.95. In the first nine years , the average degree is 14 and it has 4328 supernodes; in the mid-nine years, the average degree is 10 and it has 1727 supernodes; in the third nine year, the average degree is 13 and it has 4033 supernodes. So we can see a trend that sea ice concentration experience 27 years in which the network is less stable from the first nine year to the second nine years, and then it is more stable in the third nine years. That is probably for the reason that sea ice concentration is deteriorated from first nine year to second nine year, and because some negative feedback mechanism, the sea ice concentration is restored by itself.

3. What can we find from analysis of time lag data.

It is obvious when we consider time lag, the total connections of the graph decrease dramatically as shown in the graph above. Especially the more the time lag, the less connection the graph would has. That maybe the situation of sea ice concentration is changed with the time greatly.