CEN 502

PROJECT 1 - REPORT

GROUP 25

ABHISHEK JOSHI
ANIKET WANI
SIVAKUMAR VENKATARAMAN

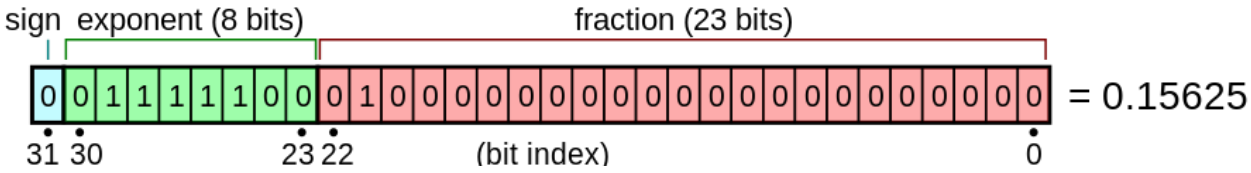
Tools:

- Language used: C++ for graph operations
- IDE used: Microsoft Visual Studio & MATLAB
- Software used for plotting Histogram : MATLAB

Data:

The Data is provided in floating point 32 bit in little endian format:

`sign * 2^exponent * mantissa`



[Img source: Wikipedia]

Each byte of data is read sequentially.
4 bytes are combined in proper order to account for the little endian format.
The bits are then processed to get the floating point number.

Data structure to represent the arrays: Adjacency matrix.

The adjacency matrix of a finite graph G on n vertices is the n × n matrix with values a(i,j)=1 if there is an edge between the elements i and j. a(i,j)=0 if there is no edge between the elements i and j.

Pearson Correlation Coefficients

We determine the relationship between two nodes X and Y. If the coefficient value is close to 1 then X and Y are strongly correlated. If the coefficient value is close to 0 then X and Y are not strongly correlated. A positive correlation indicates that if X increases then Y also increases, this is known as positive correlation. A negative correlation indicates that if X increases then Y decreases, this is known as negative correlation. The Pearson Correlation Coefficients were calculated using the formula.
[SRC: Project 1 description]

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{s_{xy}}{s_x s_y}$$

where,

$$S_{xx} = = \sum_{i=1}^{n} (X_i - \overline{X})^2$$

$$S_{yy} = = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

$$S_{xy} = = \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$$

Observe that $S_{xy} = S_{yx}$. Using this notation, we can define the *sample variance* of the Xs and Ys as:

$$s_x{}^2 = \frac{S_{xx}}{n-1}, \text{and}$$

$$s_y{}^2 = \frac{S_{yy}}{n-1}$$

Clustering Coefficient

We determined number of neighbors directly reachable with one hop from a node. We then checked for edges between all such nodes. Ev is number of edges between the nodes which are present in the cluster. Kv is the number of nodes in the cluster. Clustering Coefficient was calculated with the given formula.

$$\gamma_v = \frac{e(v)}{\binom{k(v)}{2}} = \frac{2 \cdot e(v)}{k(v) \cdot (k(v) - 1)}.$$

Characteristic path length:

We used djikstras algorithm to find the shortest path d_{ij} between every pair of node (i,j). Sum of all the d_{ij} was averaged over all ${}^{n}C_2$ nodes.

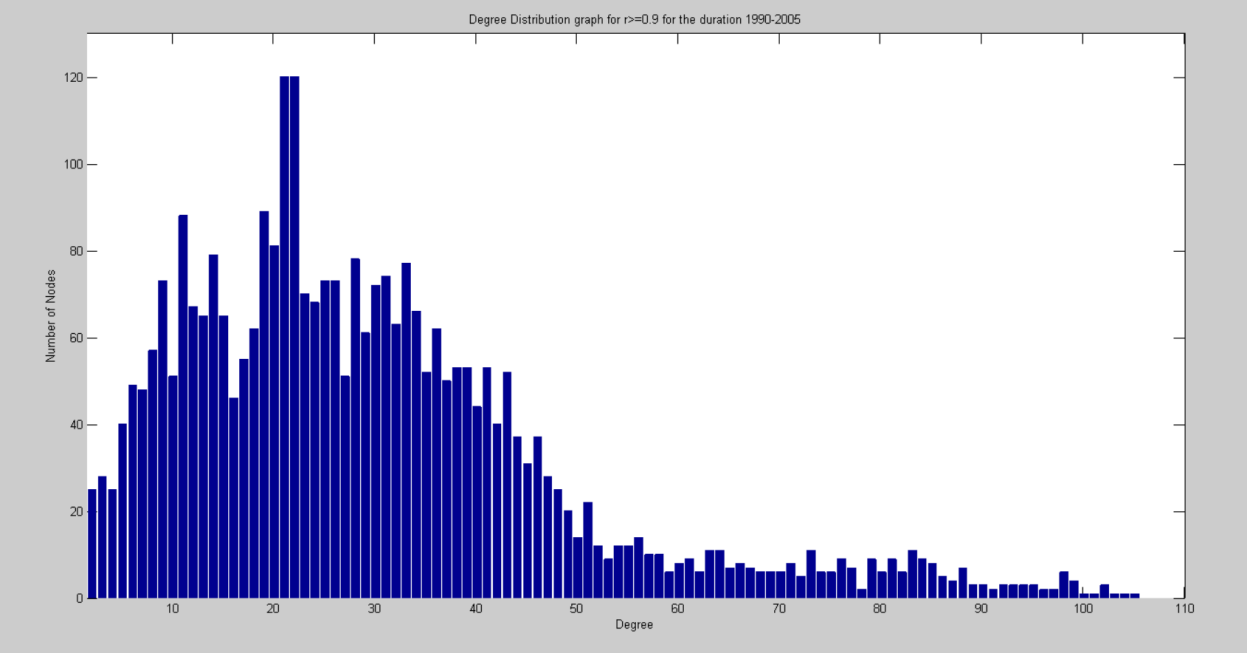
Small Data Set

- While we were able to read the big data set, we weren't able to store the graph as an adjacency matrix because the array size exceeds the size of the RAM on PC. Our code/logic is otherwise completely scalable to process the big data set.
- Our results show our interpretation/inference on the small data set.

Results:

- **Correlation = 0.9; Time Duration = 16 years :**

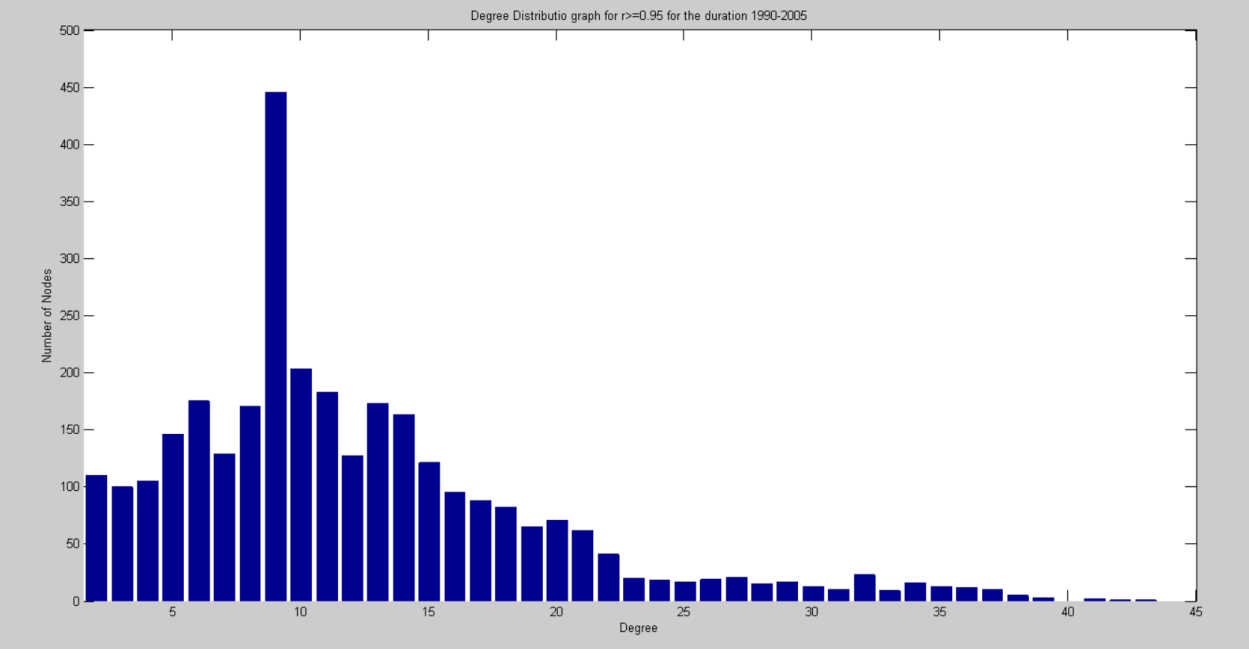
- The clustering coefficient and characteristic path length for our graph Gr is higher than the clustering coefficient and characteristic path length of a similar (same number of nodes and mean vertex degree) random graph. We can conclude that the data can be considered as a small world graph/network




```
Average degree of the nodes in the graph : 28.7749
Clustering coefficient of the graph: 0.618956
 Clustering coefficient of a random graph: 0.00907153
Characteristic path length : 12
Characteristic path length of random graph: 2.39979
```

- **Correlation = 0.95; Time Duration = 16 years :**

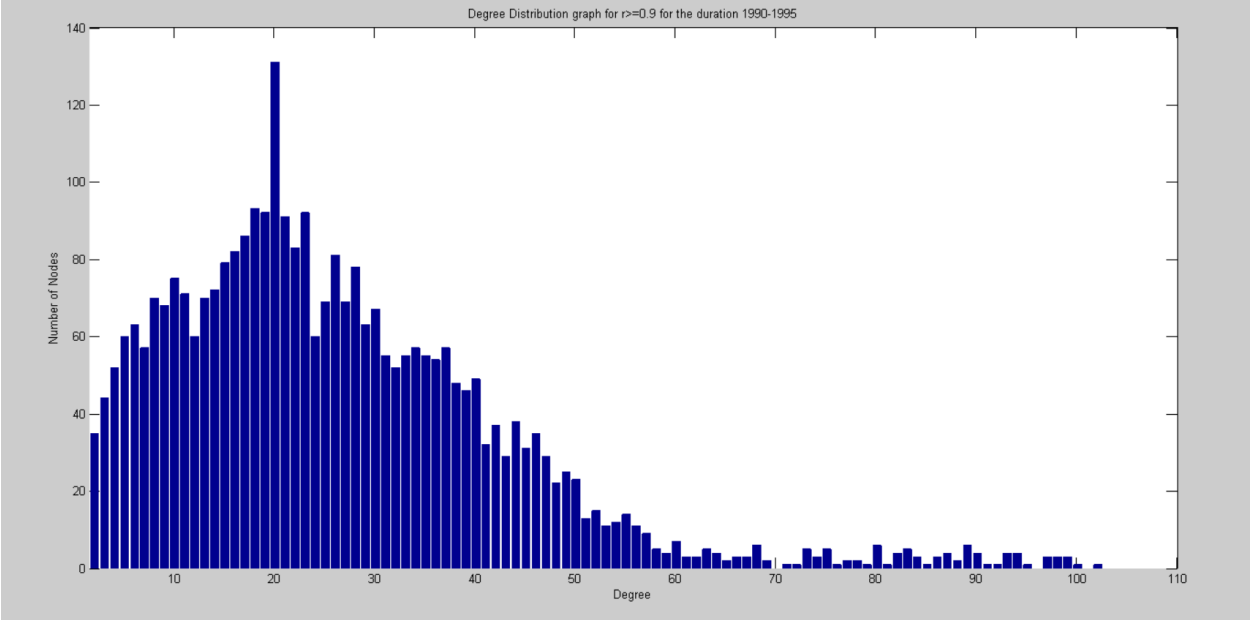
- The average/mean vertex degree is lower than the average degree we got for the graph using correlation threshold 0.9. This indicates that there are less number of edges in the graph.
- The characteristic path length of this graph is higher than the characteristic path length of the graph using correlation threshold 0.9. This implies that the average number of hops along the shortest paths for all possible pairs of network nodes is higher.




```
Average degree of the nodes in the graph : 11.1171
Clustering coefficient of the graph: 0.60235
 Clustering coefficient of a random graph: 0.00358733
Characteristic path length : 16
Characteristic path length of random graph: 3.33771
```

- **Correlation = 0.9; Time Duration = 1990-1995 :**

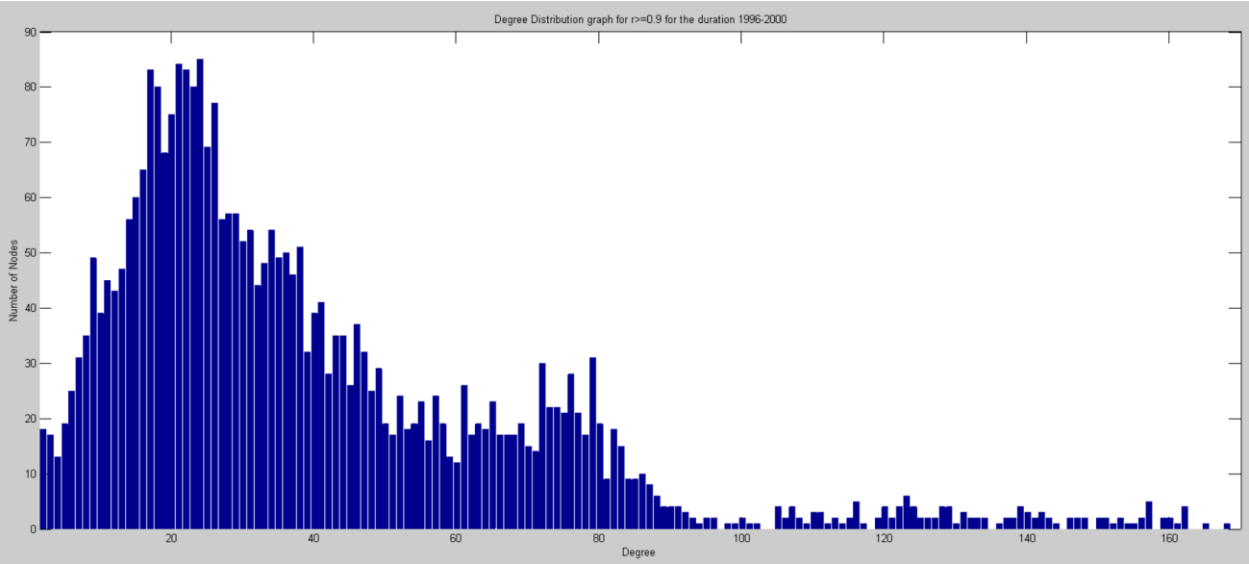
- The average/mean vertex degree is lower for the first 6 years than for the entire time duration of 15 years. Clustering coefficient is almost similar to the entire duration's clustering coefficient. Characteristic path length has increased by a small value. This reinforces the change in mean vertex degree.




```
Average degree of the nodes in the graph : 25.1819
Clustering coefficient of the graph: 0.62601
 Clustering coefficient of a random graph: 0.00797905
Characteristic path length : 13
Characteristic path length of random graph: 2.49744
```

- **Correlation = 0.9; Time Duration = 1996-2000** :

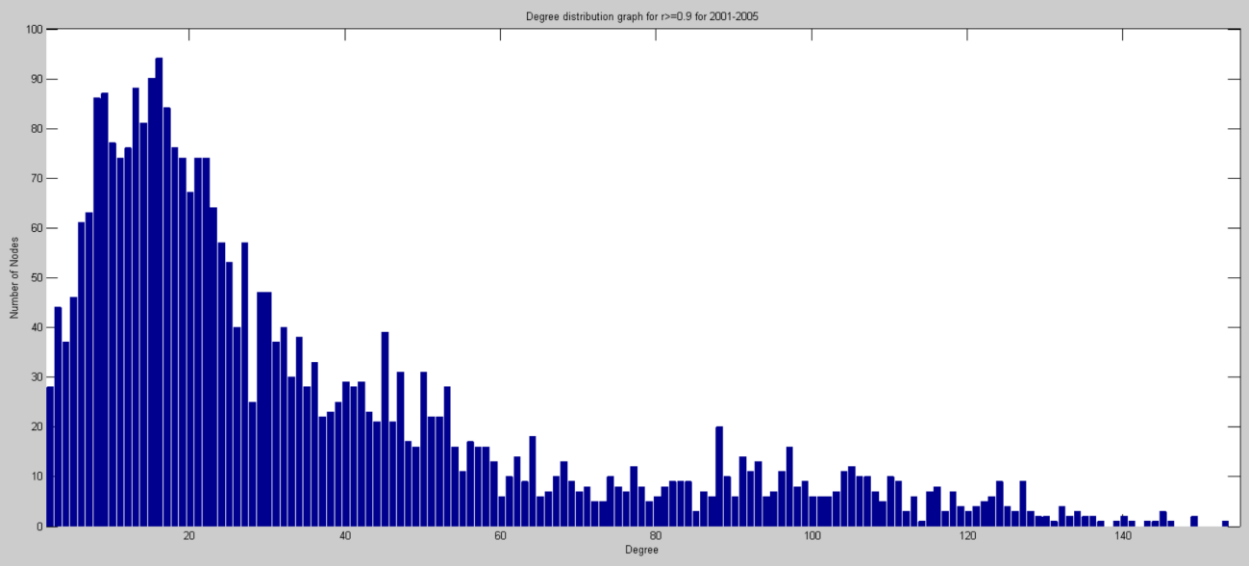
- The mean vertex degree is higher for this duration (year 1996 to year 2000) when compared to the graph for the entire duration. Clustering coefficient is also higher by a small value indicating higher number of connections between the nodes. The characteristic path length is also lower thus indicating a higher ease of reachability.




```
Average degree of the nodes in the graph : 38.5431
Clustering coefficient of the graph: 0.639723
Clustering coefficient of a random graph: 0.0121128
Characteristic path length : 10
Characteristic path length of random graph: 2.20859
```

- **Correlation = 0.9; Time Duration = 2001-2005 :**

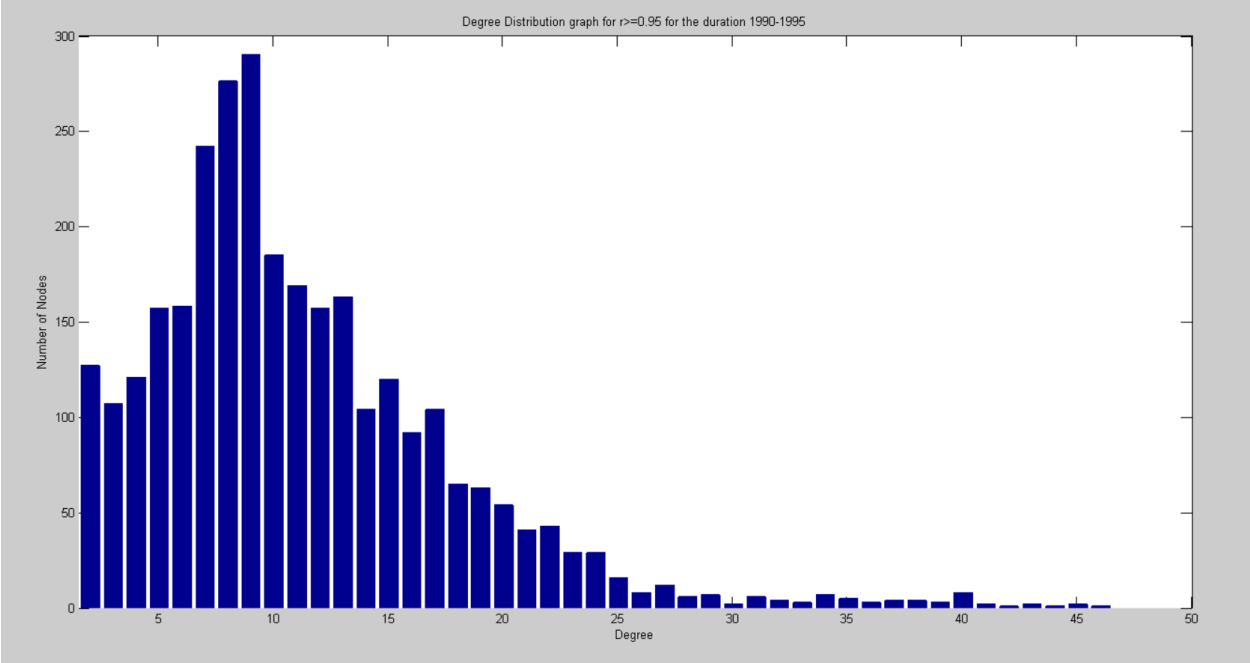
- Mean vertex degree of this graph is also higher than the graph obtained for the entire duration. It is however smaller than the mean vertex degree of the second 5 years. The characteristic path length of this graph is comparable yet smaller than the one obtained for the entire duration. Clustering coefficient is also bit higher for this duration.




```
Average degree of the nodes in the graph : 35.1462
Clustering coefficient of the graph: 0.631202
Clustering coefficient of a random graph: 0.0110732
Characteristic path length : 11
Characteristic path length of random graph: 2.26512
```

- **Correlation = 0.95; Time Duration = 1990-1995 :**

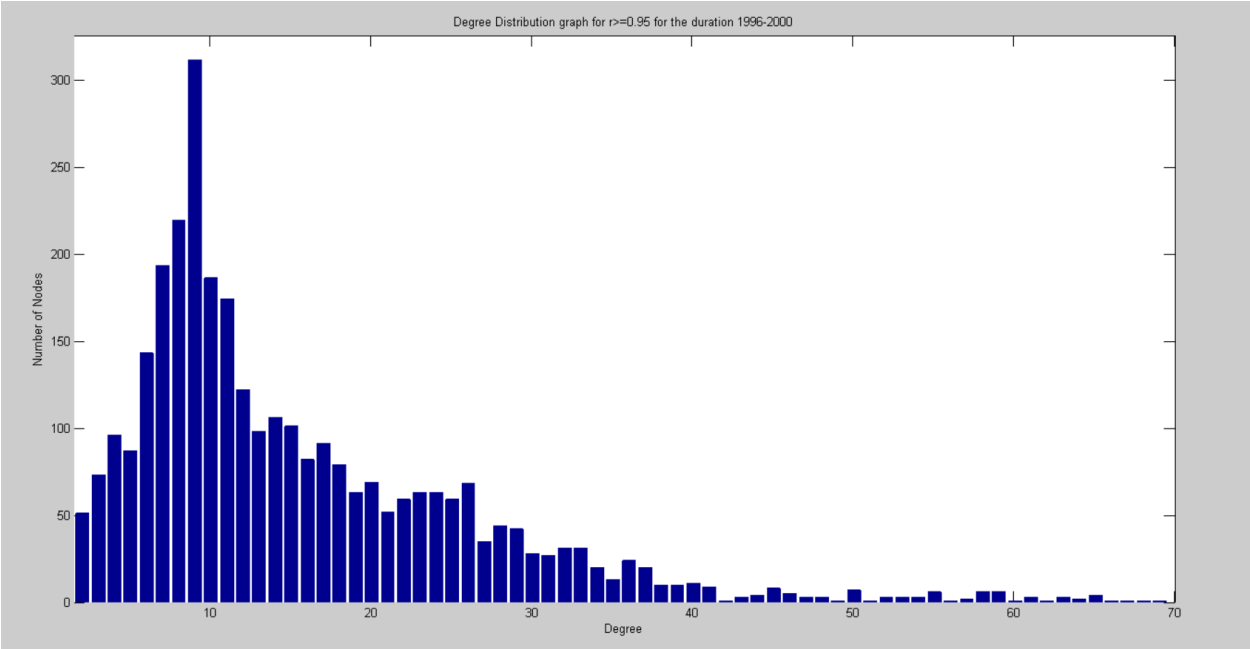
- Considering the original graph obtained for the entire duration with correlation threshold 0.95,the average/mean vertex degree is lower for the first 6 years. Clustering coefficient is almost similar to the entire duration's clustering coefficient. Characteristic path length has increased by a small value. This reinforces the change in mean vertex degree.




```
Average degree of the nodes in the graph : 10.1732
Clustering coefficient of the graph: 0.612277
 Clustering coefficient of a random graph: 0.00338767
Characteristic path length : 16
Characteristic path length of random graph: 3.45182
```

- **Correlation = 0.95; Time Duration = 1996-2000 :**

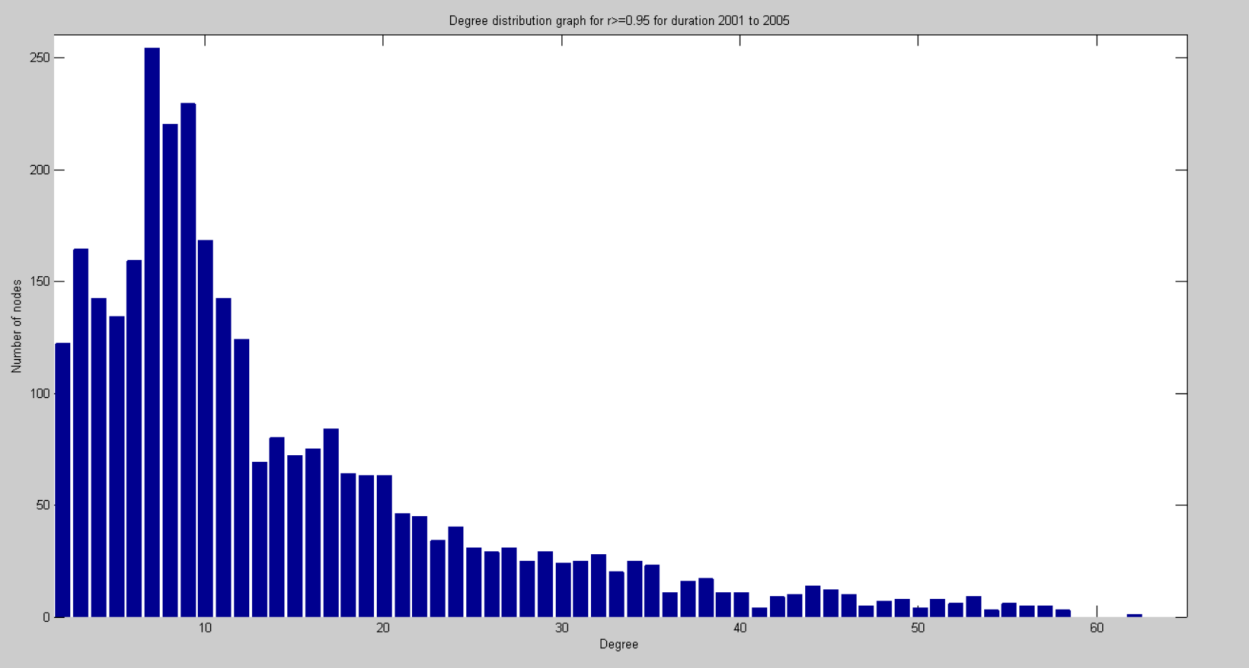
- The mean vertex degree of this graph is higher for this duration than for the entire duration with 0.95 correlation threshold. It is also higher than the mean vertex degree obtained for the 1st 6 years duration with 0.95 correlation threshold. Clustering coefficient is similar to the one obtained for the entire duration.




```
Average degree of the nodes in the graph : 14.4976
Clustering coefficient of the graph: 0.616154
Clustering coefficient of a random graph: 0.00460681
Characteristic path length : 16
Characteristic path length of random graph: 3.01206
```

- **Correlation = 0.95; Time Duration = 2001-2005 :**

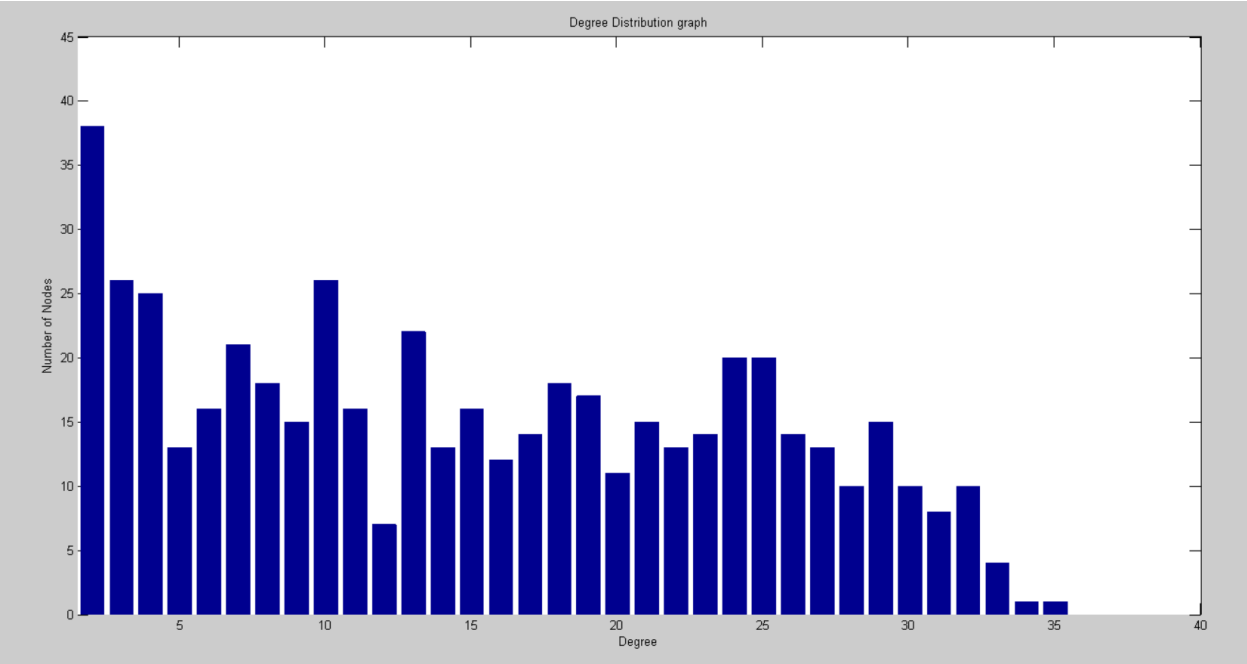
- Mean vertex degree is lesser than the mean vertex degree for the second 5 years duration. However, it is higher than the mean vertex degree for the entire duration. Clustering coefficient is almost similar to the entire duration. Characteristic path length is also lesser than the one obtained for the entire duration indicating a higher reachability in this span of time.




```
Average degree of the nodes in the graph : 13.18
Clustering coefficient of the graph: 0.617028
Clustering coefficient of a random graph: 0.004282
Characteristic path length : 15
Characteristic path length of random graph: 3.11476
```

- **Lag by one week and for correlation = 0.9:**

- The mean vertex degree is much lower than the graph that was obtained with no time lag.
- The number of nodes and number of edges for this graph are also less. However, the clustering coefficient is a bit higher for this graph.




```
Average degree of the nodes in the graph : 14.1719
Clustering coefficient of the graph: 0.663624
 Clustering coefficient of a random graph: 0.0276794
Characteristic path length : 3
Characteristic path length of random graph: 2.35297
```

- For correlation threshold = 0.9, there were no nodes and hence no graph for time lag of 2, 3 & 4 weeks. For correlation threshold = 0.95, there were no nodes and hence no graph for time lag of 1, 2, 3 &4 weeks.

Handling Missing Data/Errors:

- Our code handles missing data(value 157) and land data(value 168) and excludes them appropriately during calculations.
- We also avoid the divide by zero condition that might occur while calculating the correlation step, when Sxx=0 or Syy=0. For such entries, the correlation is simply zero and there is no edge between them.

Optimization:

- We realized that we were using more space to store the values of Sxy between every element and correlation between every element. Hence, we decided to not store Sxy and correlation as we are just using it for comparison in one step.

Time and Space Complexity:

We observe that edge calculation has the highest number of nested loops. The outer two loops together run n times where n is the number of elements. The next two loops run n times to calculate the correlation between each pair of nodes. Finally the inner loop will runs for the total number of weeks. Hence, worst case time complexity can be given by $O(n^3)$.

Edge data is stored in n*n adjacency matrix. And so the space complexity is given by (n*n).

Inference- Small World Networks:

The clustering coefficient and the characteristic path length for a random graph G_{random} is calculated and compared with our graph G_r. We observe that the clustering coefficient and characteristic path length for our graph G_r is higher than the clustering coefficient and characteristic path length of a similar (same number of nodes and mean vertex degree) random graph G_{random}. We can conclude that the data can be considered as a small world graph/network.