

---

**ARIZONA STATE UNIVERSITY**  
**CEN 502 — Computer Systems II**  
(Fall 2015)

---

**Project #1 on Algorithm Design**  
**Analysis of Sea Ice Concentration Anomaly Data**  
**Using Correlation Based Graph**

Submitted by  
Ramsundar Kalpagam Ganesan (1207857868)  
Suresh Gururajan (1208567798)  
Aravind Rajendran (1208684590)

## Abstract

Arctic sea-ice levels are a proxy indicator of climate change and a lot of research is being conducted on this topic. Our aim is to analyze the longest continuous satellite record of sea-ice records from the data obtained from the meteorological satellites. The analysis presented in this report is conducted on the anomaly dataset comprising sea ice concentration (SIC), which spans 27 years. A correlation based graph is constructed using this dataset and various graph parameters are observed from the graph. We base our conclusions on these parameters. These findings could be useful in predicting the sea ice concentration and the consequent impact in climatic pattern of the entire world for the future and hence could help us create awareness and alleviate the reasons for it.

## 1. Introduction

The motivation behind this project is to analyze a large data set using efficient algorithms and representing it in a way that is intuitive and accurate. In our project, the data set represents sea ice concentration collected by NASA's satellite revolving around the Arctic Circle. The dataset is in the form of an anomaly data points, which is obtained by subtracting values from the long term average. An anomaly data set is used so as to remove seasonal trends which makes data obliging to statistical analyses. Such data sets are commonly used in National Centers for Environmental Protection (NCEP) for analyzing wind circulation. The data gives the percent deviation from the long term average. Every year as ice keeps melting, NASA has predicted that the Arctic would be ice free by 2050-2100. This has a direct impact on global climate. Hence, a lot of research is conducted on the Arctic sea ice. In our project, we analyze sea ice concentration (SIC) data set and find if it is a small world network.

A small world network is characterized by a high degree of local clustering, and small and random number of long range connections for information transfer. An advantage of "small world" network is that they facilitate efficient transfer of information and represent how well a unit of analysis is connected to another. This is because of its small number of long range connections characteristic [2]. Some of the applications [3] of small world networks include earth sciences (seismic network in California [3]), estimation of usability of data in huge databases [3] among others.

In our project, given an anomaly data set, we need to find if it is a small world network. To accomplish this, we need to calculate the following parameters:

1. Mean Clustering Coefficient of the Graph
2. Characteristic path length of the Graph ( $L$ )
3. Comparison with a Random Graph of the same size

For small world graph, the mean-clustering-coefficient ( $G$ )  $\gg$  mean-clustering-coefficient ( $G_{\text{random}}$ ) and  $L(G) \approx L(G_{\text{random}})$ . This leads us to the question of how to create the graph. The graph is a correlation based graph  $G = (V, E)$  where  $V$  is the set of vertices,  $E$  is the set of edges where edge  $e = \{u, v\}$ , where  $u, v$  belong to  $V$ . We add vertices and connect them by an edge only if the correlation between the vertices is above a "correlation threshold". We add vertices to the graph only if two points (vertices) have a strong correlation. We keep adding vertices to this graph until there's no more vertices (above the correlation threshold) left to add. Once we create the graph, we compute its mean clustering coefficient. Then, we calculate the characteristic path length ( $L$ ) of the graph  $G$ . Finally, we compare it to a random graph of the same size.

## 2. SIC Anomaly Dataset

We are given an SIC dataset comprising 27 years (from 1979 to 2005) of weekly anomaly data. We have based our findings on this dataset. Each week comprises a 304 x 448 floating point array. Each element (cell) denotes a region of 25 km<sup>2</sup> region. This dataset is specific to Arctic region and hence is localized to the northern hemisphere. There are 136,192 cells in total for each week in each year, for 27 years. We have ignored land data and the circular region above the North Pole where the satellite does not gather data due to its orbit.

## 3. Implementation Details

We have used MATLAB® as the primary environment to read and process the dataset and visualize the results. The reason we chose MATLAB® is partly attributed to its implementation of algorithms, plotting functionalities and syntax and partly to the availability of libraries that are optimized to handle huge datasets, which help in reducing the overall run time of the code.

Some important formulae used in our implementation are given below.

### 3.1 Pearson Correlation Coefficient (r)

*Pearson correlation coefficient*( $r$ ) is used to measure the correlation between 2 regions' data. This gives an idea of how the data is connected and the pattern variation in one region with respect to another. It is defined as,

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

Where,

$$\begin{aligned} S_{xx} &= \sum_{i=1}^n (X_i - \bar{X})^2 \\ S_{yy} &= \sum_{i=1}^n (Y_i - \bar{Y})^2 \\ S_{xy} &= \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) \end{aligned}$$

### 3.2 Clustering Coefficient ( $\gamma$ )

*Clustering coefficient* ( $\gamma$ ) denotes the ratio of the number of edges between 2 vertices (within the neighborhood) to the maximum possible edges that could exist between these two vertices. It is defined as,

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

#### 4. Design Decisions

1. We used MATLAB for the entire project because of its versatility and optimizations in performing computation on vectors and matrices.
2. When we read the data from the data files, instead of transforming it to a matrix representation (2D), we used vector representation (1D) because vector computations are faster.
3. *Graph representation*
  - We used adjacency matrix to represent the correlation graph
  - There are no vertices with vertex degree 0 (zero) i.e., we do not consider such a data point.
  - When we perform computations such as calculating the clustering coefficient and correlation graph, we use sparse matrix because it is memory efficient
4. Our project was executed on the following *resource specifications*:
  - 16 GB RAM
  - Intel i7 processor clocked @ 3.5 GHz
  - Number of cores = 8
5. The values of parameters estimated, reported and discussed in this project are with respect to these resources.

#### 5. Code Optimizations

The inbuilt functions present in MATLAB and libraries available for MATLAB were exploited for optimizing our code. Various techniques and practices followed during the coding phase of the project are discussed in detail in this section.

##### 5.1 Correlation Coefficient Computation

- Since there are  $n = 136,192$  cells in an array (i.e. 1 week of data), and we needed to compute the correlation coefficient for  $(n * (n-1)) / 2$  pairs of cells, it takes a lot of computation and memory to move forward.
- We decided to remove unused data from memory and exclude land data or missing values from computations. We focused on storing variables that were most relevant to the current point of execution and most likely to be reused in further stages of the project, so as to save memory.
- For calculating the correlation coefficient, we used the *corr()* function provided by MATLAB [1]. By default, the *corr()* function computes the Pearson correlation coefficient. We used this function because it allows us to compute correlation between two blocks of data points, which is faster than computing between two data points.
- The block size can be specified in the function call and it is a tradeoff between memory and running time. In our project, we set the block size to 10000 (equal to 10000 data points).
- We found that in order to build the correlation graph, the calculated value of  $r$  (correlation coefficient) meet the threshold. There are no other places where the value of  $r$  is reused. Since storing so many values of  $r$  takes a lot of memory, we decided against storing it.

## 5.2 Correlation graph

The data structure for representing the correlation graph is a sparse matrix where each value is stored as a Boolean (logical 0 or 1) which greatly reduces the space occupied by the matrix.

## 5.3 Characteristic Path Length Computation

For computing the characteristic path length, we used the MatlabBGL package [4] written by David Gleich [5]. We used a function, *shortest\_paths()*, that uses Dijkstra's algorithm. The function calculates the shortest path from one vertex to all the other vertices in a single function call. The runtime of this function is  $O(|V| \log |V|)$  where  $|V|$  = number of vertices [6] (p.42). We found this function to be very efficient in terms of runtime.

## 5.4 Other Optimizations & Coding Practices

Other optimizations include storing as less temporary variables in our code as possible so as to save memory. In addition, once we decide that a variable is not going to be used further in a program, we clear the variables and free the occupied memory.

## 5.5 Summary of code optimizations

	Estimate for large dataset <sup>1</sup>		Recorded parameters <sup>2</sup> (After code optimization)	
	Memory	Running Time (Sec)	Memory	Running Time (Sec)
Pearson correlation coefficient	2.03 GB (if all computed values of $r$ are stored)	40,000 (~11 Hours)	600 MB (storing intermediate results temporarily)	130 (~2 Mins)
Correlation Graph	14.7 MB <sup>3</sup> (using sparse matrix) ~4 GB (using full matrix)		14.7 MB (using sparse matrix)	
Clustering coefficient and Characteristic path length	227 KB	110,000 (~30 Hours)	227 KB + 64 KB (temporary storage)	700 (~11.5 Mins)


1. The estimate is based on worst case scenario, considering no optimization in the code. The actual un-optimized performance may vary.

2. Values are obtained with reference to the large data set.

3. Building the correlation graph and storing it in a sparse matrix.

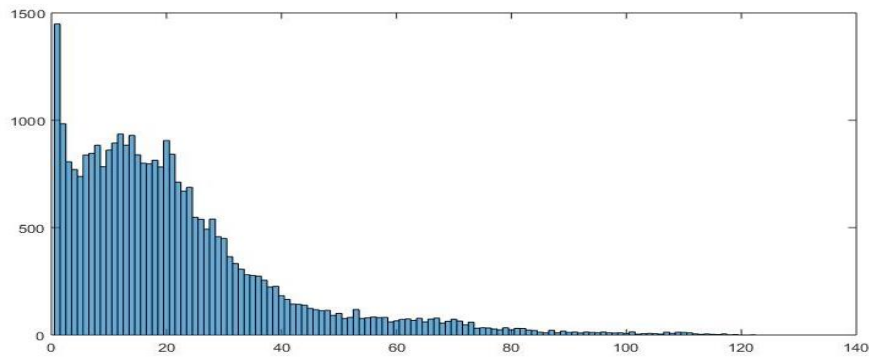
## 6. Results and Discussion

In this section, we present the results of Tasks 1, 2 and 3 and compare and contrast their differences in graph statistics. The results for each task are presented in the following order:

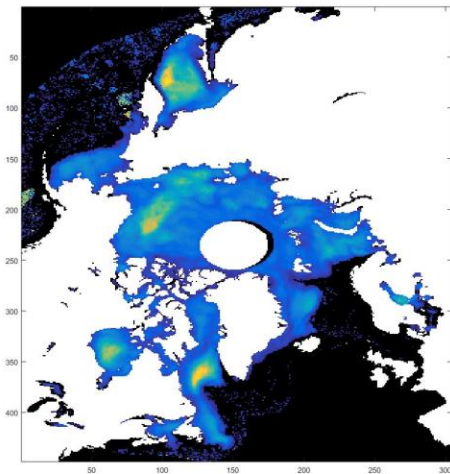
<p><b>Task #</b></p> <ul style="list-style-type: none"> <li>- <b>A histogram of the degree distribution of vertices in graph <math>G_r</math></b></li> <li>- <b>Graph representation</b> <ul style="list-style-type: none"> <li>- Graphs are represented in pseudo color.</li> <li>- The graph on <b>Left</b> with <b>all the nodes</b> is represented according to the degree distribution</li> <li>- The graph on <b>Right</b> with all the <b>Super Nodes</b> is represented according to the degree distribution</li> </ul> </li> </ul>	<p>ColorMap Legend</p>  <p>Min Degree Max Degree</p>
---	--

### Task 1

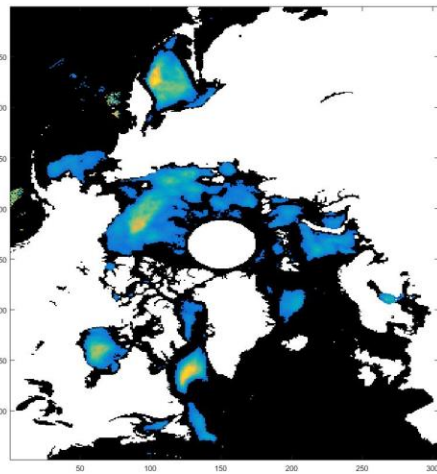
For Threshold ( $r$ ) = 0.90



Histogram showing the degree distribution of  $G_r$



All nodes obtained from the graph  $G_r$



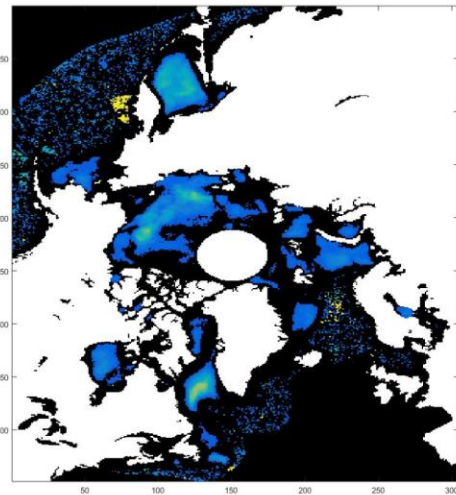
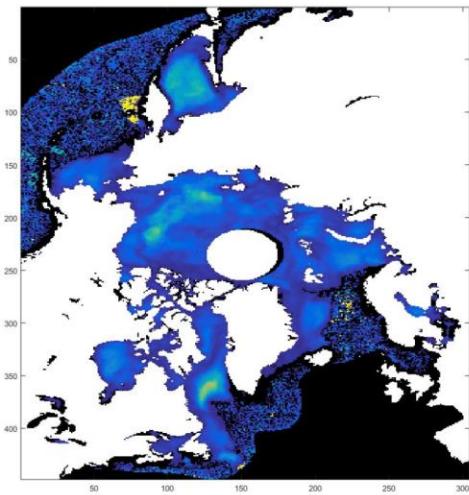
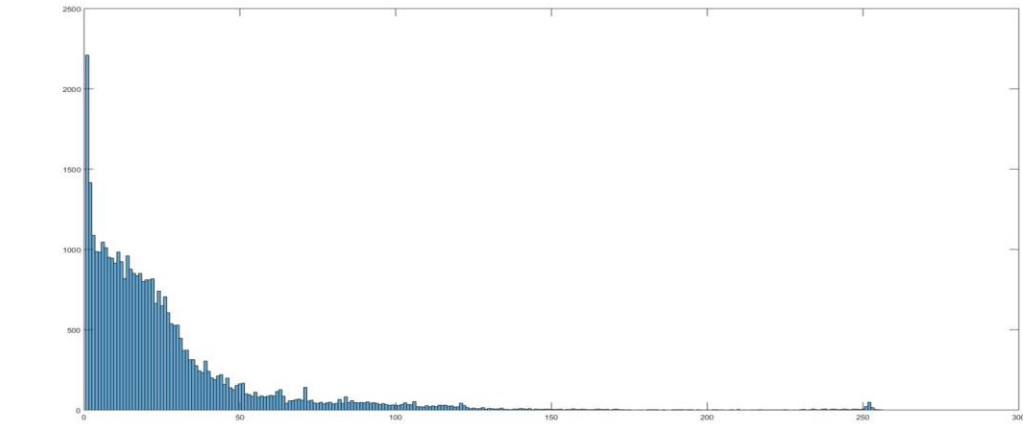
Super Nodes represented by density.

## Task 2

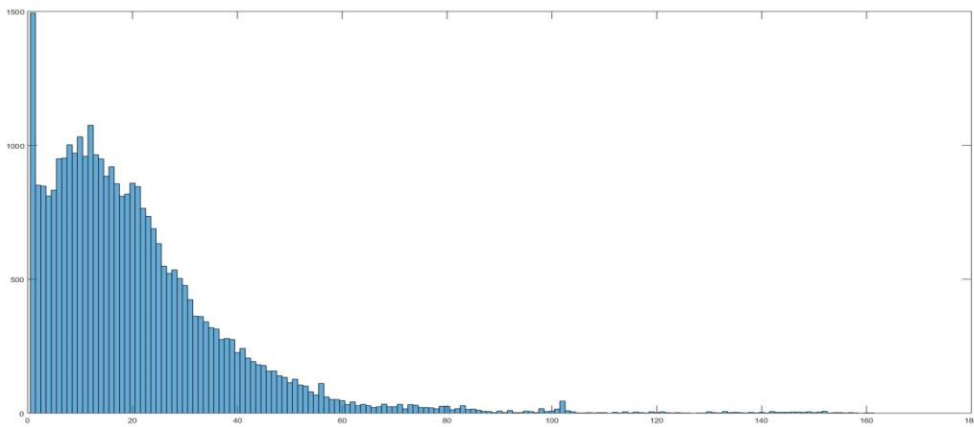
Construction of a correlation based graph for 9-year periods.

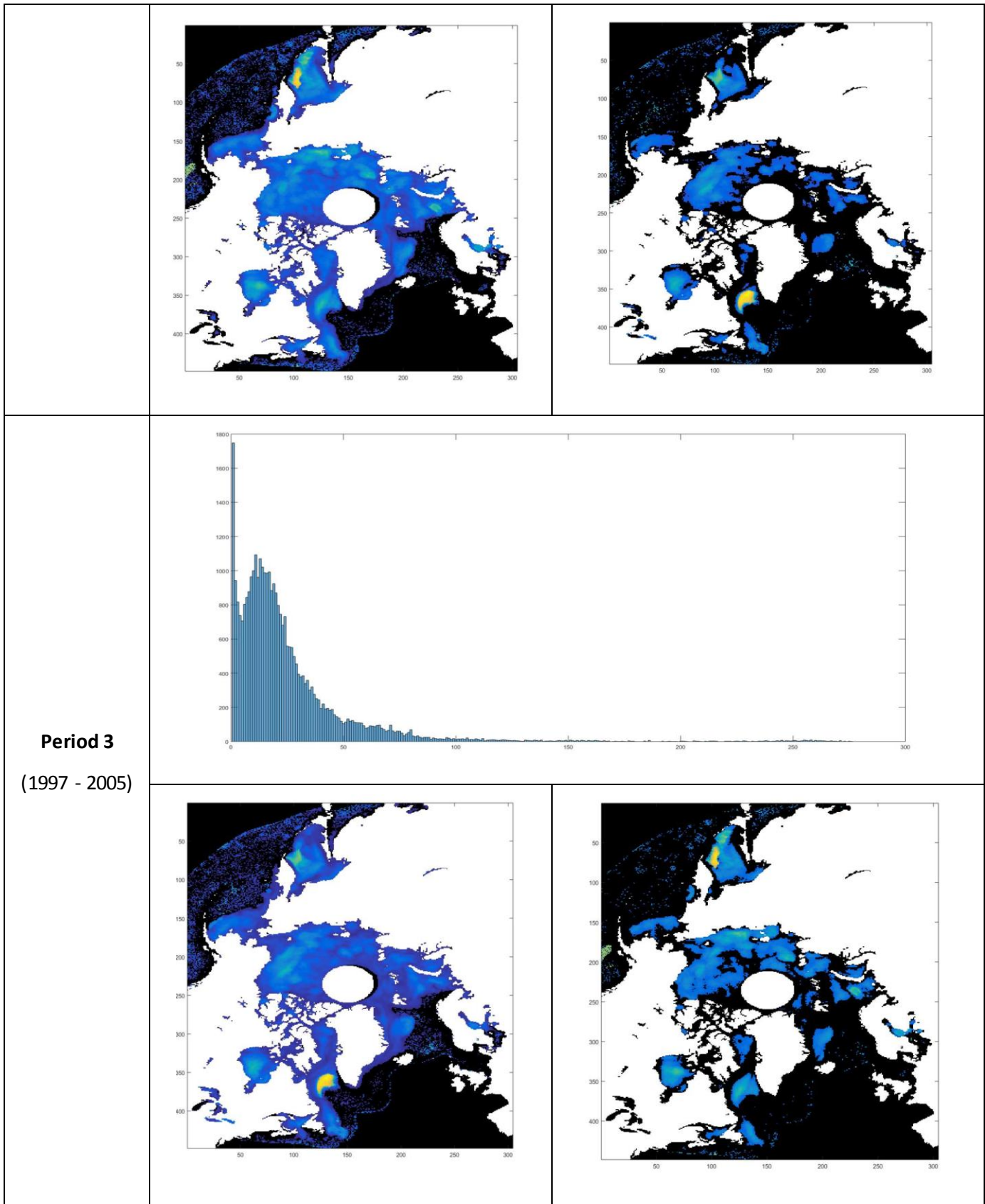
For Threshold  $(r) = 0.90$

**Period 1**  
(1979 - 1987)



**Period 2**  
(1988 - 1996)

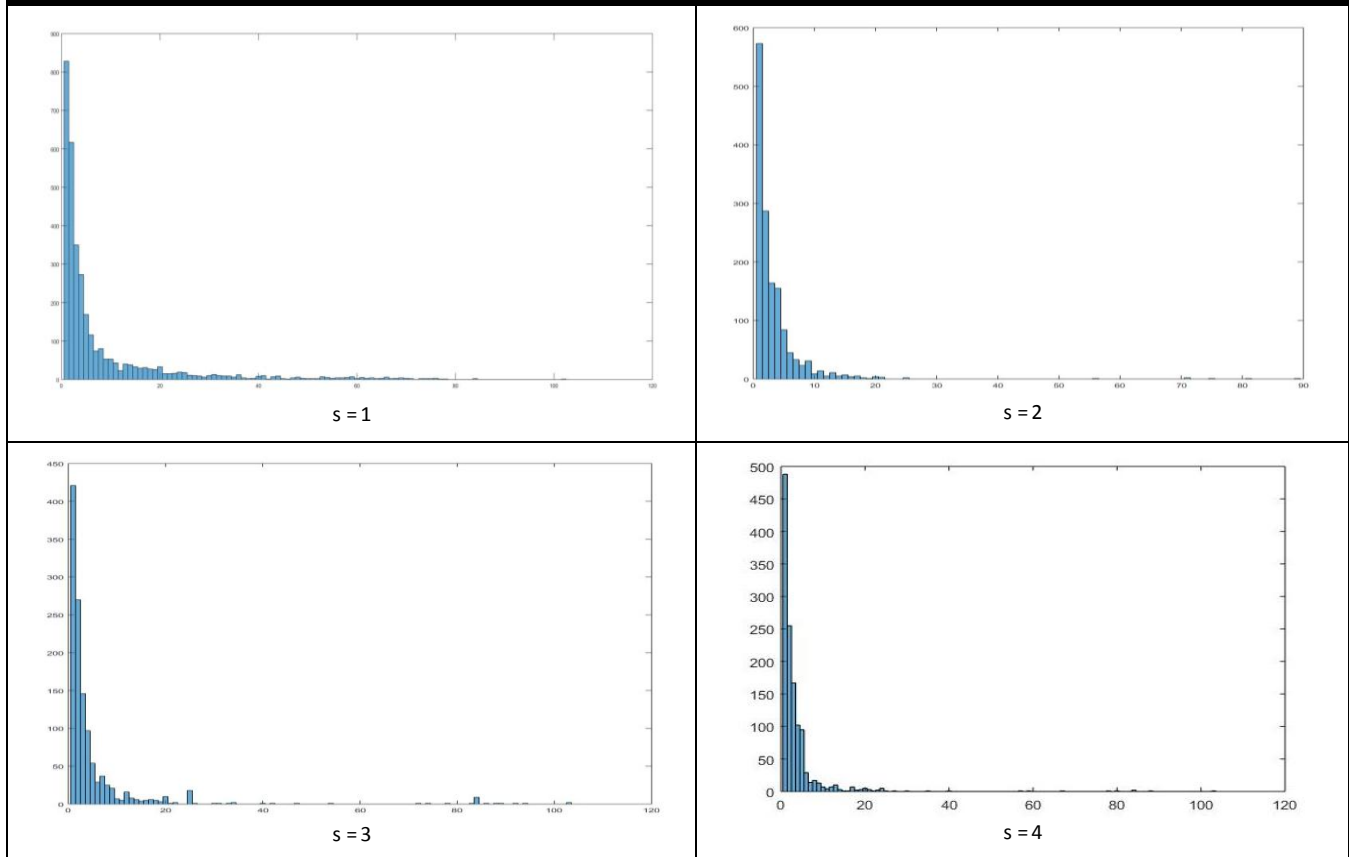




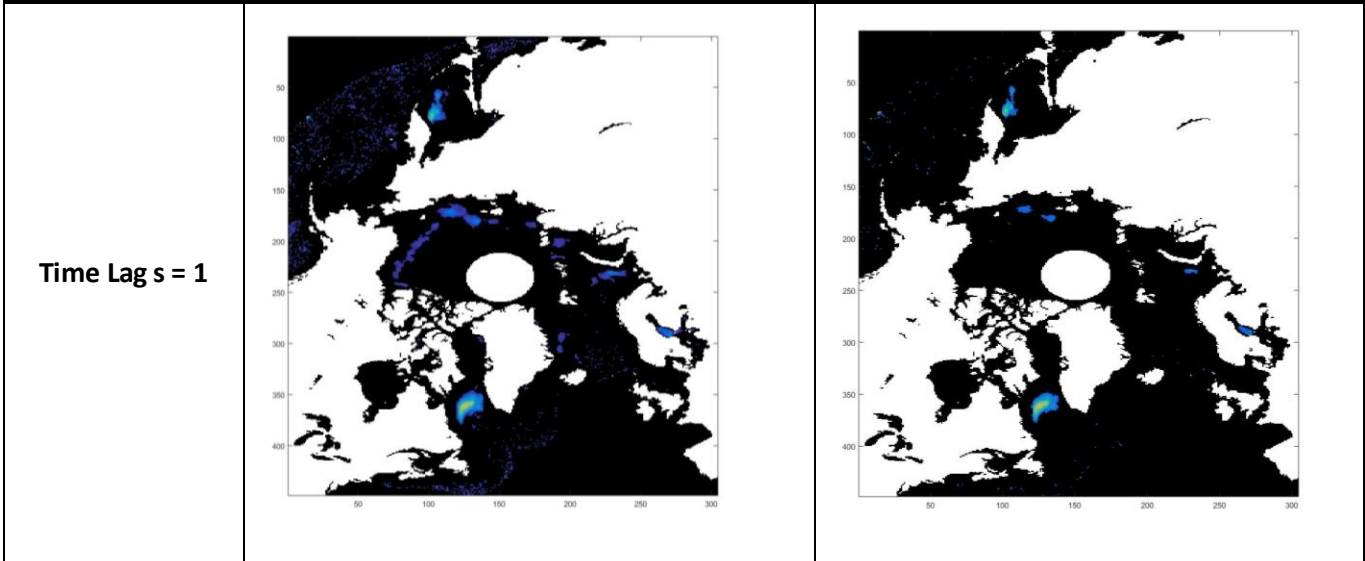


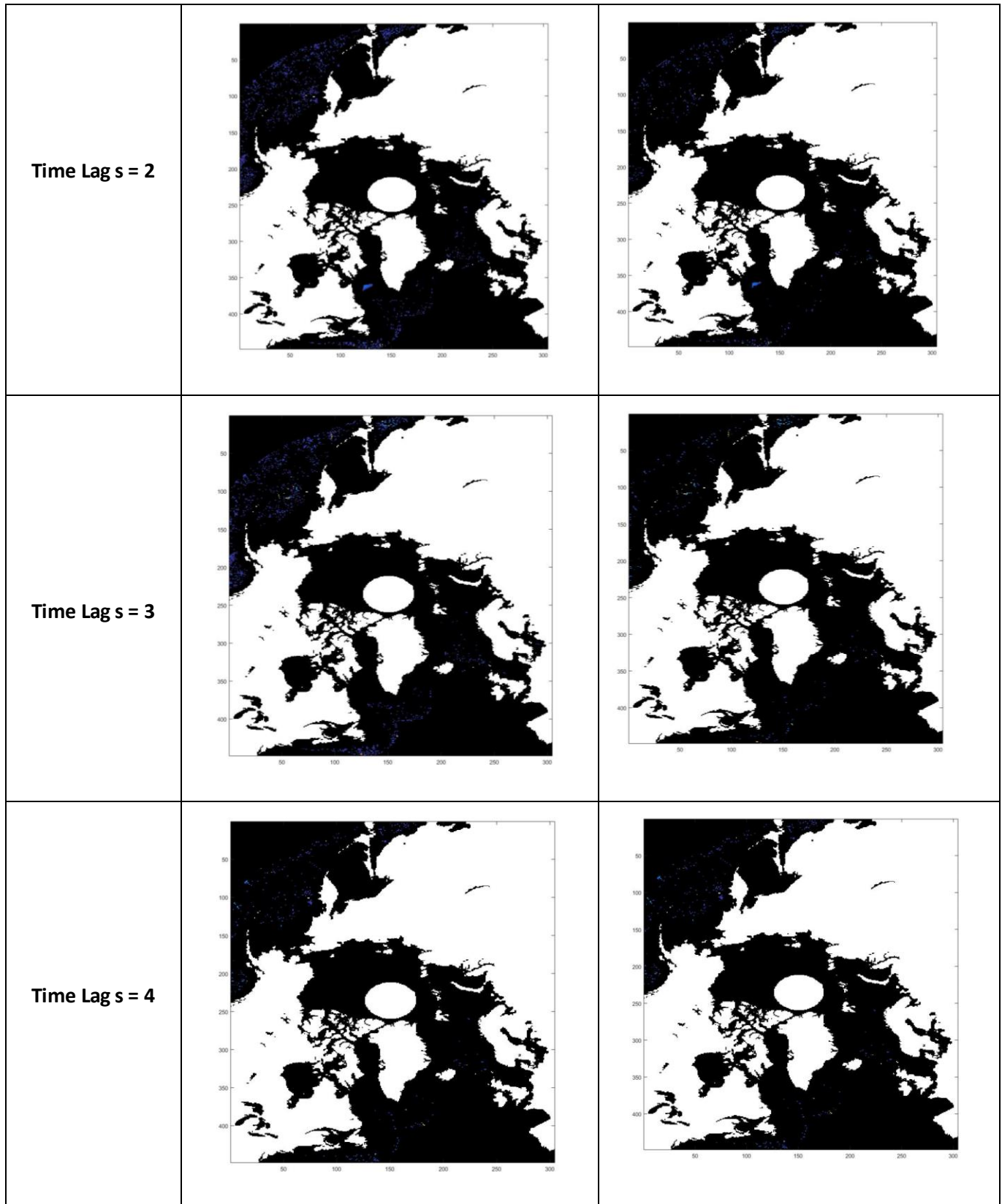
### Task 3

**Histogram for Threshold ( $r$ ) = 0.90**



**Threshold ( $r$ ) = 0.90**





## 7. Comparison

The various parameters that we obtained for the three tasks (Threshold  $r = 0.90$ ) are listed in the table below. (The parameters obtained for threshold of  $r = 0.95$  can be found in the appendix).

Parameters	Threshold of $R = 0.90$							
	Task 1	Task 2			Task 3			
		Period 1 (1979 - 1987)	Period 2 (1988 - 1996)	Period 3 (1997 - 2005)	$S = 1$	$S = 2$	$S = 3$	$S = 4$
Degree Mean	21.406266	25.79877	20.758305	24.91398	8.080232	3.465716	5.3646	3.767331
Clustering Coefficient ( $G_r$ )	0.538669	0.58906	0.562845	0.569309	0.156625	0.00292	0	0
Clustering Coefficient ( $G_{rand}$ )	0.000719	0.000715	0.000655	0.000744	0.002465	0.002353	0.004376	0.003002
Ratio of Clustering Coeff ( $G_r / G_{rand}$ )	749.1919332	823.8601399	859.3053435	765.2002688	63.5395538	1.24096898	0	0
Characteristic Path Length ( $L_r$ )	62.748812	65.181935	57.741008	67.722916	4.87253	3.053042	3.246193	2.571854
Characteristic Path Length ( $L_{rand}$ )	3.362497	3.22834	3.417132	3.24012	3.874274	5.869292	4.233492	5.379275
Ratio of ( $L_r / L_{rand}$ )	18.66137338	20.1905422	16.89750586	20.90136044	1.25766273	0.52017211	0.7667885	0.47810421

We observe that for Task 1 and Task 2, when we do not take a time lag, the ratio of clustering coefficient ( $G_r$ ) to the random graph ( $G_{random}$ ) is very high (closer to or higher than 750) in each case, while when we consider the time lag, it significantly reduces. In fact, for  $s = 3$  and  $s = 4$ , the ratio is zero. This means that no 3 vertices in the graph form a triangle, which implies no 3 data points are strongly correlated. In addition, we find that the characteristic path length of the correlation graph ( $L_r$ ) is similar when we consider the 27 years data set and the 9 years data set.

From the table, it is evident that the statistics obtained from Task 1 and Task 2 are not that different. However the characteristic path length is higher for the random graph than for the correlation graph  $G_r$  in Task 3.

The ratios ( $G_r / G_{random}$ ) and ( $L_r / L_{random}$ ) represent the characteristics of a small world graph. These ratios are not as high in the case of lagged correlation as they are in the 27-year period and the 9-year periods' correlation. Hence, we conclude that the small world graphs (27 year period, 9 year periods) are highly connected. With the exception of time lag  $s = 1$ , where the graphs seem to be highly connected as well, other lagged correlations seem to be poor in connectivity.

## 8. Conclusion

We analyzed the anomaly data set representing the sea ice concentration around the Arctic Circle. We computed the Pearson correlation coefficient for the data set. We then represented these graphs in pseudo color. There were three scenarios to compare: 1) 27 years of data 2) three sets of 9-year data each, and 3) a lag of  $s = \{1, 2, 3, 4\}$  weeks. We used threshold =  $\{0.9, 0.95\}$  in all of these cases and have compared the results of graph statistics. We learnt that optimizing code is important when handling huge data sets.

## Acknowledgment

We would like to thank Dr. Violet Syrotiuk for providing the opportunity to analyze this dataset and for encouraging us to use various optimization techniques to make our code run faster while consuming less memory.

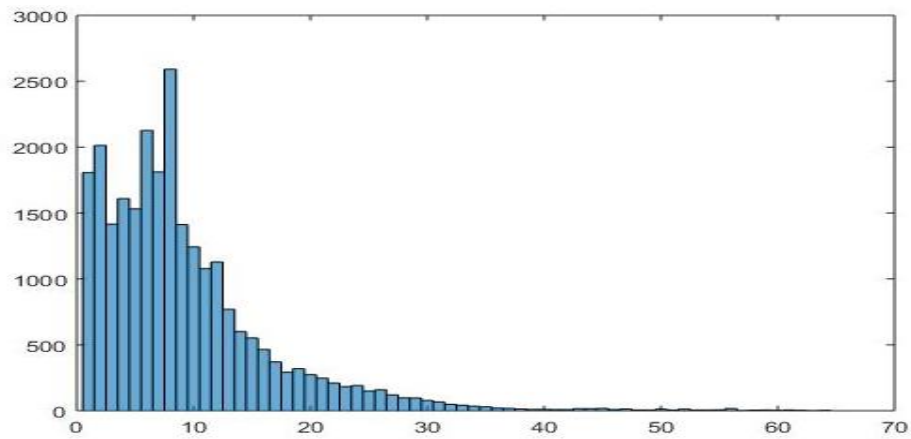
## References

- [1] "Documentation," Linear or rank correlation. [Online]. Available at: <http://www.mathworks.com/help/stats/corr.html>. [Accessed: 26-Sep-2015].
- [2] A. Tsonis, K. L. Swanson, and P. J. Roebber, "What do networks have to do with climate?" Bulletin of the American Meteorological Society, pp. 585–595, May 2006.
- [3] [https://en.wikipedia.org/wiki/Small-world\\_network#Applications](https://en.wikipedia.org/wiki/Small-world_network#Applications)
- [4] <http://www.mathworks.com/matlabcentral/fileexchange/10922-matlabbgf>
- [5] [https://www.cs.purdue.edu/homes/dgleich/packages/matlab\\_bgl/](https://www.cs.purdue.edu/homes/dgleich/packages/matlab_bgl/)
- [6] [https://www.cs.purdue.edu/homes/dgleich/packages/matlab\\_bgl/matlab\\_bgl\\_v2.1.pdf](https://www.cs.purdue.edu/homes/dgleich/packages/matlab_bgl/matlab_bgl_v2.1.pdf)

## Appendix

### Task 1

For Threshold ( $r$ ) = 0.95



Graph showing the degree distribution of  $G_r$

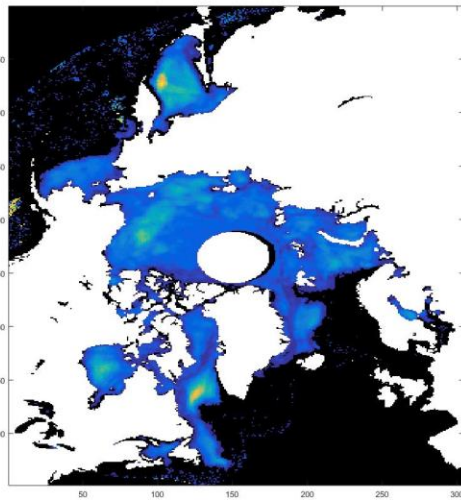


Figure : All nodes obtained from the graph  $G_r$

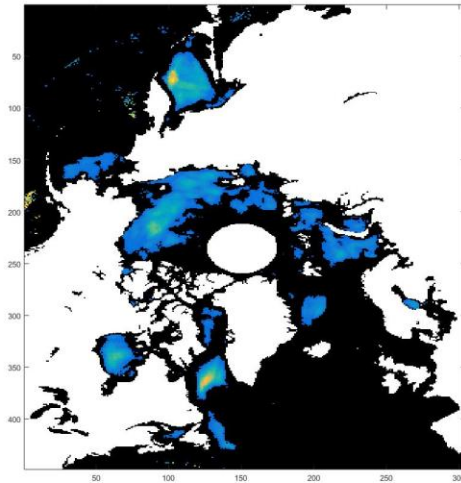
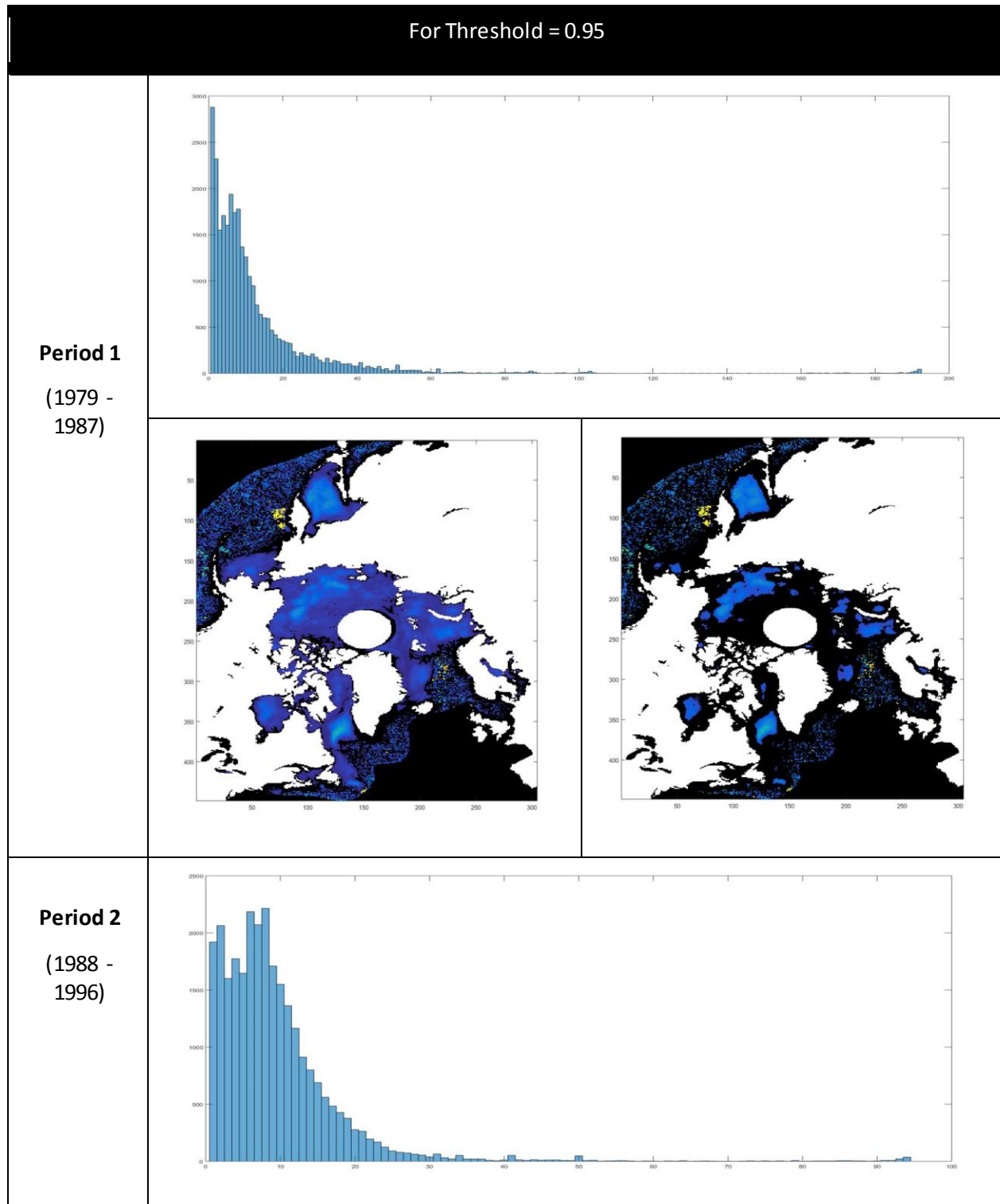
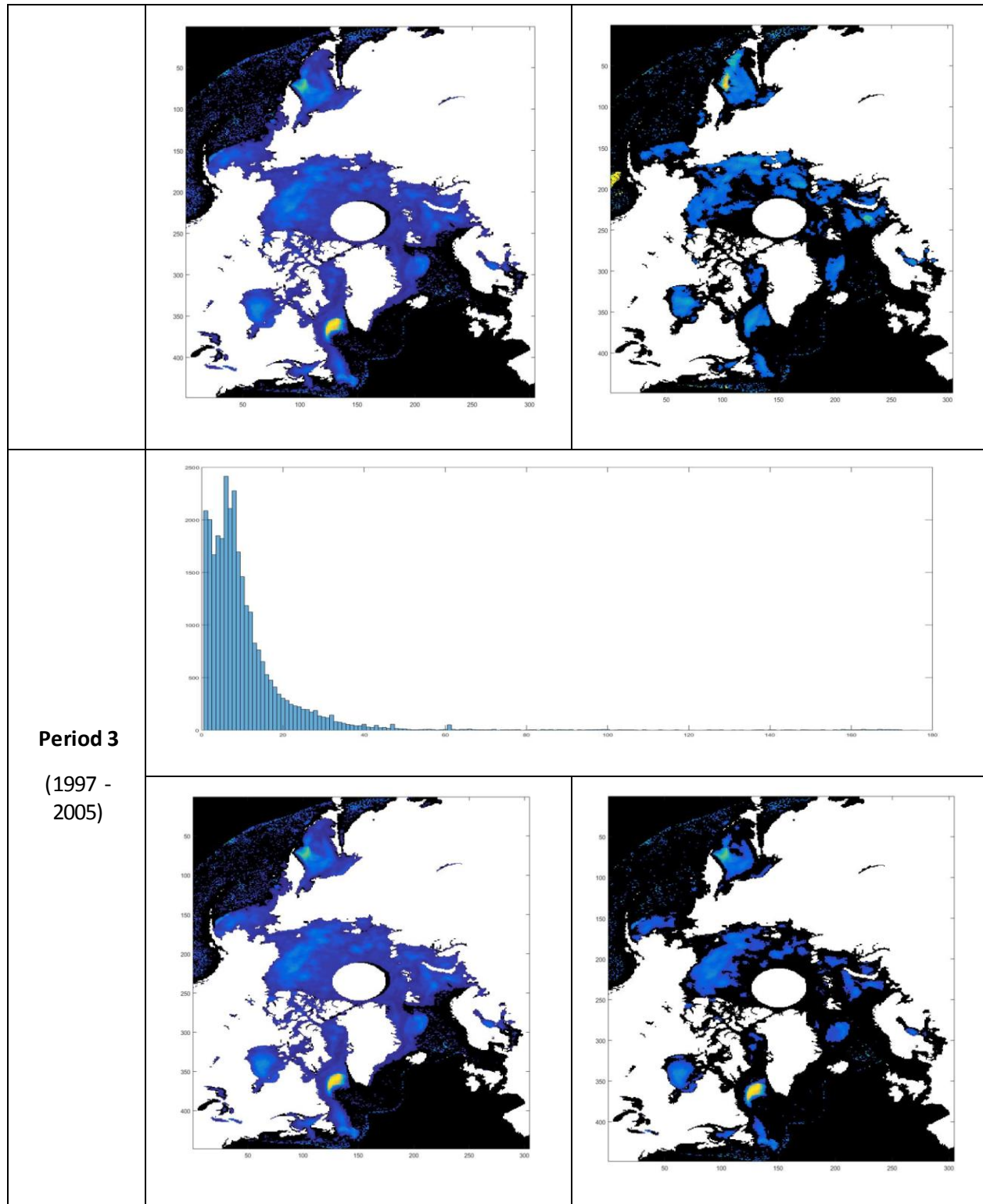


Figure : Supernodes represented by density.

## Task 2

Construction of a correlation based graph for 9-year periods.

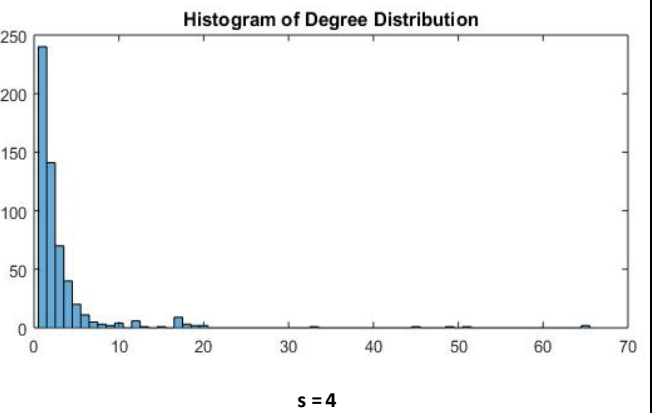
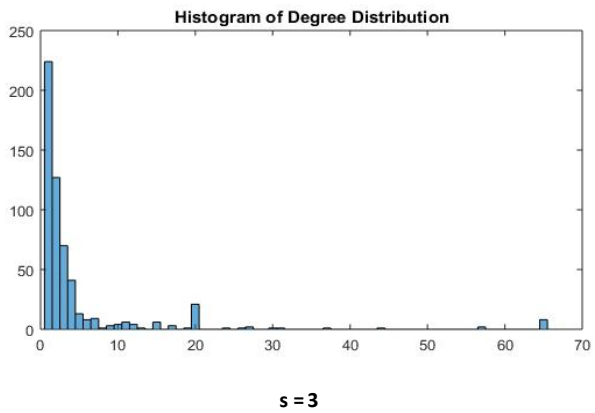
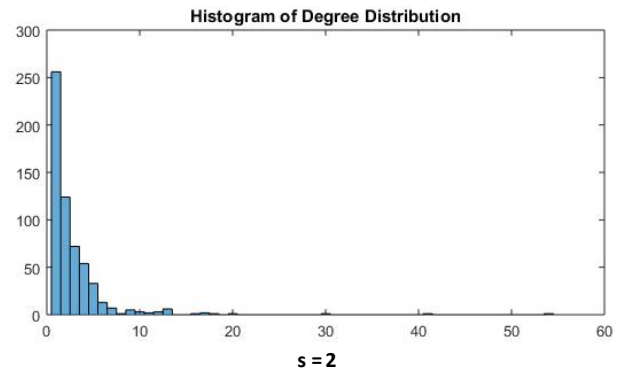
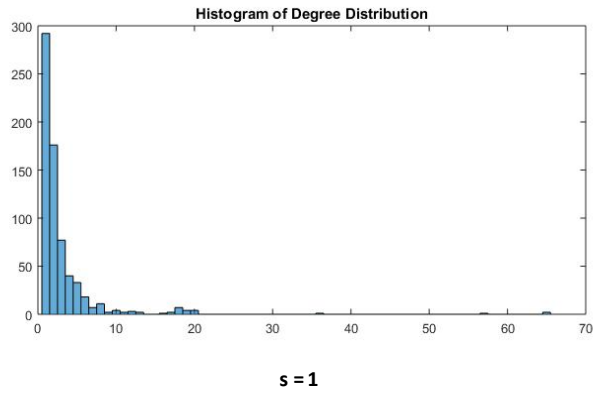






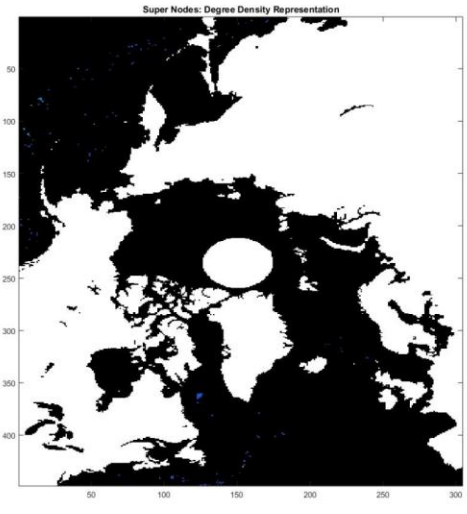
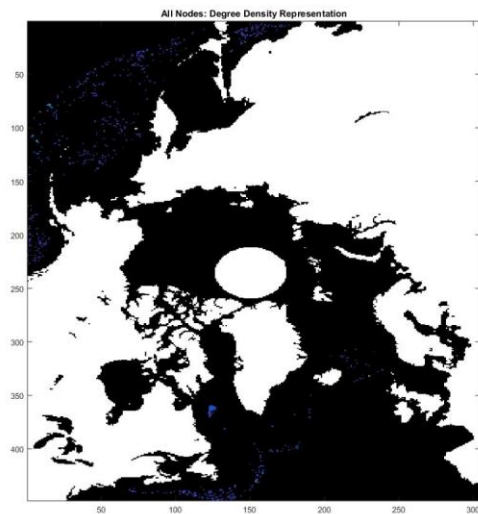
### Task 3

#### Histogram for threshold ( $r$ ) = 0.95

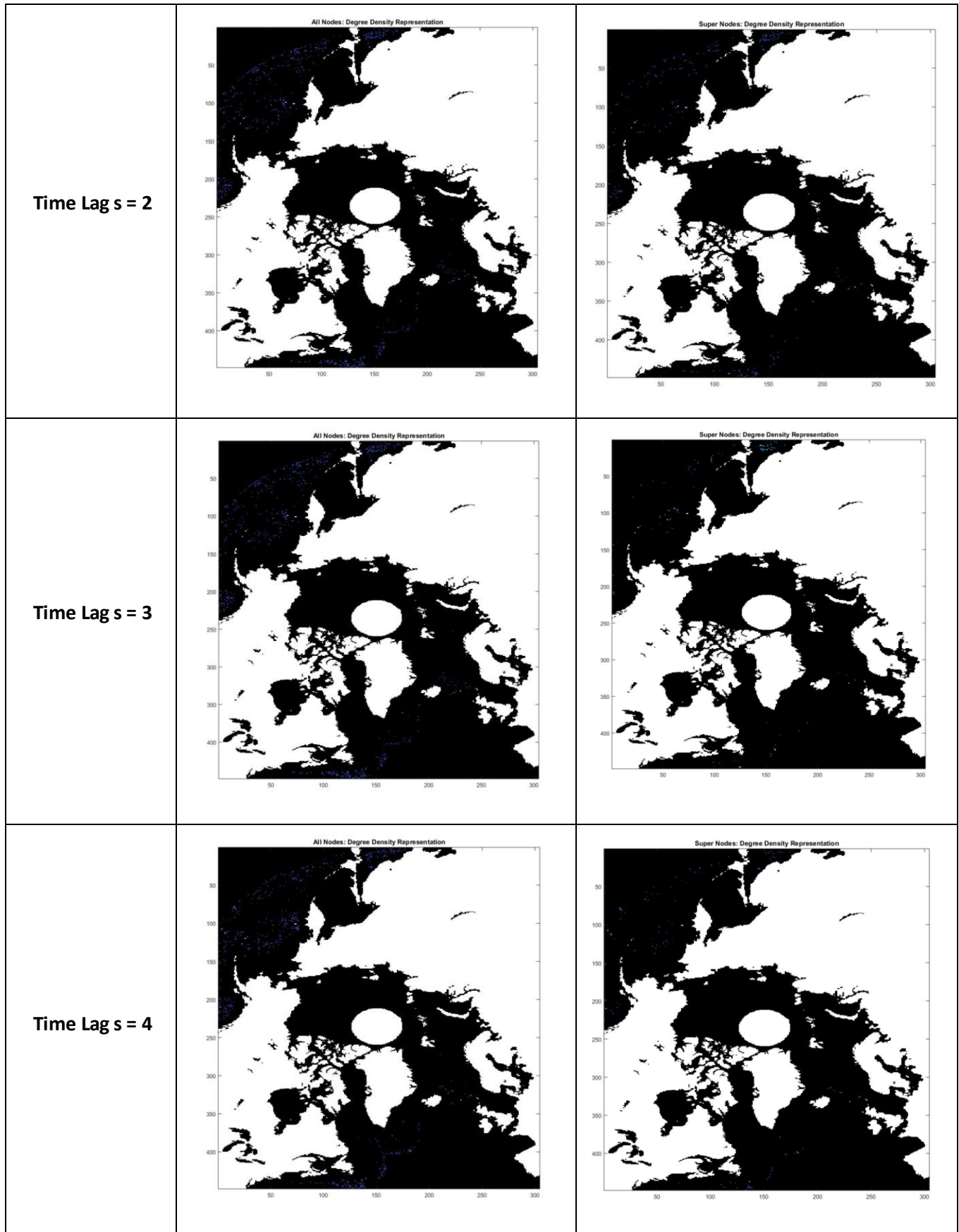


#### Threshold $r = 0.95$

Time Lag  $s = 1$







The various parameters that we obtained for the three tasks (Threshold  $r = 0.95$ ) are listed in the table below.

Parameters	Threshold of $R = 0.95$							
	Task 1	Task 2			Task 3			
		Period 1 (1979 - 1987)	Period 2 (1988 - 1996)	Period 3 (1997 - 2005)	S = 1	S = 2	S = 3	S = 4
Degree Mean	9.238238	13.0716	9.519464	11.253684	3.108853	2.827939	4.803571	3.300353
Clustering Coefficient ( $G_r$ )	0.407065	0.4916	0.436278	0.452136	0.008891	0	0	0
Clustering Coefficient ( $G_{rand}$ )	0.000363	0.000448	0.000345	0.000382	0.004512	0.004818	0.008578	0.005831
Ratio of Clustering Coeff ( $G_r / G_{rand}$ )	1121.391185	1097.321429	1264.573913	1183.602094	1.97052305	0	0	0
Characteristic Path Length ( $L_r$ )	55.285335	31.3653	55.575915	62.785484	3.81083	1.986259	2.261122	2.093545
Characteristic Path Length ( $L_{rand}$ )	4.562164	3.9996	4.537804	4.251086	5.761709	6.132496	4.032177	5.308574
Ratio of ( $L_r / L_{rand}$ )	12.11822613	7.842109211	12.247315	14.76928107	0.66140619	0.32389079	0.56076953	0.3943705