Project Report CEN 502 (Project 1)

Submitted by:

Group 4

Darshit Kandpal

Sugam Kapoor

Haoran Lyu

1.

Representation of the graph

In the graph simulated, an object oriented representation of linked list using Java has been used. Each node denotes a cell in the given Arctic region and each edge between two nodes denote that the correlation coefficient between them is greater than 0.9.

In this we have considered the graph to be all those points which have an edge to atleast 1 more vertex. For example: If a vertex has no neighbours it is not part of our graph. In order to calculate the value of the gamma ie the Clustering coefficient we consider only the nodes which have more than one neighbour since a neighbourhood of 1 cannot be a graph.

Similarly to calculate the value of characteristic path length we take only the connected nodes.

a) Following histograms show the degree distributions of the nodes for 27 years' data taking correlation threshold as 0.9.

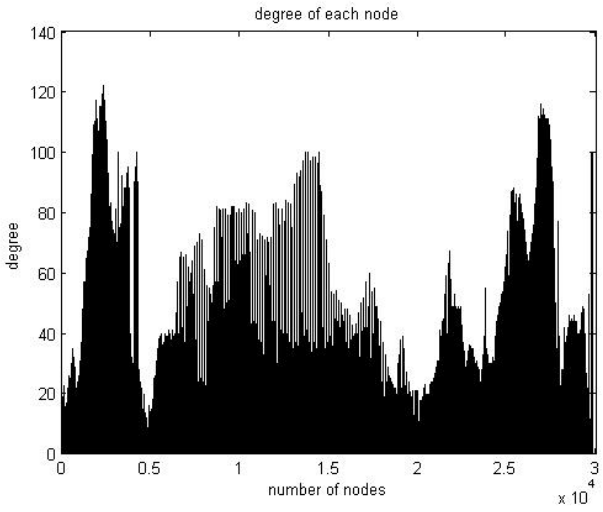


Fig. 2. Number of Nodes vs Degree

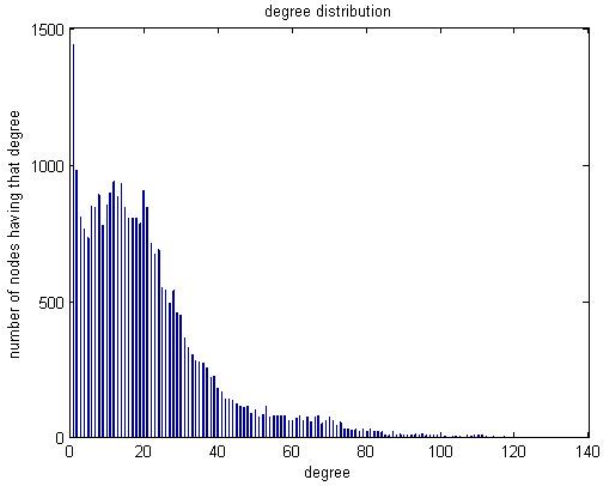


Fig. 1. Degree vs Number of Nodes

To calculate the supernodes, we have assumed the following condition:-

If the degree of a node is greater than or equal to N_{mean} + (5*Standard Deviation), then it is considered to be a supernode. The following supernodes have been identified:-

Node Index	4968	5074	5515	5624	5626	5739	5740	5741	5858	5859	5860
Degree	117	114	115	114	115	119	113	117	114	118	119

Node Index	5861	5980	5981	5982	6107	6108	6109	41333	41334	41335	41527
Degree	117	116	117	112	114	117	113	116	113	115	114

b) Clustering Coefficient $\gamma(G_r)$ = 0.5766881096865856
Characteristic Path Length L (G_r) = 36.80721377523421

c) Clustering Coefficient for random graph $\gamma(G_{random})$ = 0.00072019
Characteristic Path Length for random graph L (G_{random}) = 3.3613

On comparing the values for the generated graph and the random graph, we observe that:-

1. The clustering coefficient of the generated graph is much higher than that of the random graph. This is because there is strong correlation between arctic ice densities as they are uniform in most places. Whereas, the data is random for a random graph showing negligible correlation.
2. The characteristic path length of the generated graph is around 11 times longer than that of the random graph as the graph for ice concentration is strongly connected to other nodes, both near and far.

2. Now the dataset has been split into 3 equal parts of 9 years each.

Analysis for the first part

a) The following histograms show the degree distributions of the nodes for the first 9 year period

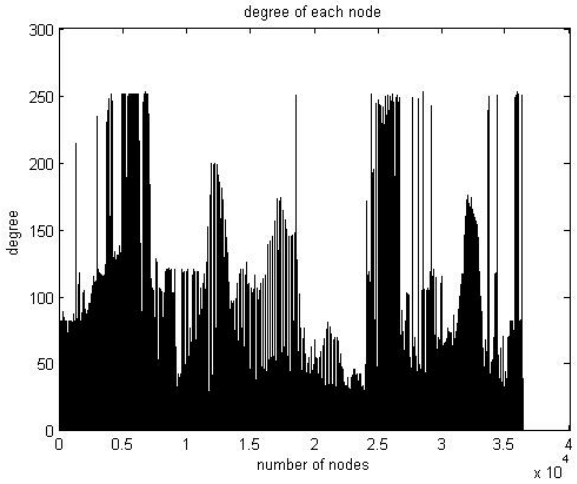


Fig. 1. Degree vs Number of Nodes

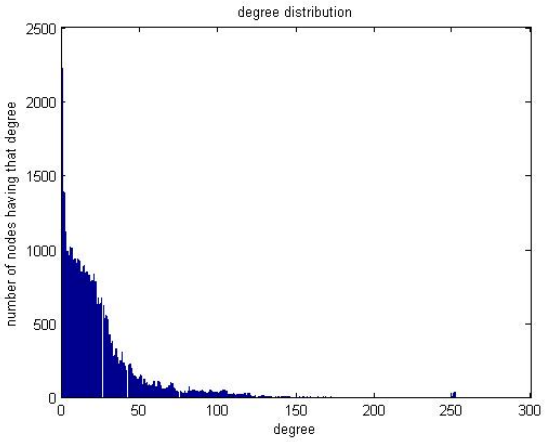


Fig. 2. Number of Nodes vs Degree

Following are the supernodes presented in the format (node index, degree):-

(2524,214),(4683,234),(5662,230),(5789,238),(5901,237),(5902,237),(5911,247),(6204,251),(6335,246),(7264,202),(7265,251),(7267,201),(7269,213),(7392,193),(7394,223),(7396,251),(7397,251),(7398,250),(7399,233),(7524,248),(7525,250),(7526,250),(7528,251),(7529,251),(7530,250),(7531,233),(7532,230),(7533,231),(7655,251),(7658,251),(7659,218),(7660,205),(7662,188),(7663,240),(7665,189),(7782,239),(7783,249),(7788,207),(7789,216),(7791,245),(7792,230),(7793,236),(7794,188),(7913,251),(7914,242),(7916,251),(7917,250),(7918,251),(7921,250),(7922,250),(7923,200),(8045,250),(8046,249),(8048,251),(8050,224),(8051,232),(8170,251),(8171,251),(8172,251),(8173,251),(8174,250),(8296,250),(8297,243),(8298,249),(8299,251),(8301,250),(8302,246),(8418,250),(8419,251),(8420,236),(8421,251),(8422,202),(8424,251),(8425,250),(8426,251),(8541,251),(8542,223),(8543,251),(8544,192),(8545,251),(8546,251),(8547,251),(8549,251),(8550,225),(8663,251),(8664,243),(8665,251),(8666,236),(8667,251),(8668,251),(8669,242),(8670,251),(8671,214),(8672,234),(8783,251),(8785,209),(8786,250),(8787,230),(8789,251),(8791,194),(8792,218),(8906,250),(8907,251),(9035,188),(9036,216),(9487,245),(9490,244),(9590,251),(9591,222),(9593,250),(9694,234),(9695,251),(9696,214),(9698,237),(9790,243),(9791,250),(9792,250),(9793,250),(9794,250),(9795,253),(9796,238),(9797,204),(9887,250),(9888,235),(9889,251),(9890,250),(9892,238),(9893,229),(9984,251),(9985,251),(9986,207),(9987,249),(9

988,208),(10086,246),(10087,242),(10191,236),(10192,232),(10194,191),(16725,199),(16726,189),(16868,198),(16869,198),(16870,188),(17006,199),(17007,197),(17145,190),(17146,197),(17147,198),(17148,195),(17283,189),(17284,190),(23927,250),(23928,250),(30842,251),(30953,192),(31055,195),(31370,244),(31373,236),(31377,196),(31570,246),(31572,189),(31575,231),(31664,213),(31665,239),(31667,243),(31668,235),(31669,241),(31670,237),(31767,242),(31875,227),(31876,229),(31988,199),(31989,228),(31991,241),(32113,228),(32116,202),(32242,249),(32364,235),(32490,248),(32491,239),(32495,250),(32496,250),(32620,239),(32738,249),(32739,196),(32852,246),(32978,233),(32981,251),(33101,245),(33107,190),(33337,248),(33350,192),(33443,250),(33445,249),(33450,249),(33451,249),(33555,243),(33556,245),(33575,241),(33674,248),(35097,248),(35738,247),(36188,252),(37010,242),(43615,239),(43862,233),(43863,249),(44865,250),(47011,248),(47210,220),(47211,204),(47212,250),(47213,208),(47318,249),(47319,252),(47419,250),(47420,241),(47421,251),(47515,249),(47831,250),(47858,250)

b) Clustering Coefficient $\gamma(G_r)$ = 0.6435962873637511
Characteristic Path Length L (G_r) = 24.64048754531346

c) Clustering Coefficient for random graph $\gamma(G_{random})$ = 0.00074609
Characteristic Path Length for random graph L (G_{random}) = 3.1840

On comparing the values for the generated graph and the random graph, we observe that:-

1. The clustering coefficient of the generated graph is much higher than that of the random graph. This is because there is strong correlation between arctic ice densities as they are uniform in most places. Whereas, the data is random for a random graph showing negligible correlation.

2. The characteristic path length of the generated graph is around 8 times longer than that of the random graph as the graph for ice concentration is strongly connected to other nodes, both near and far.

Analysis for the second part

a) The following histograms show the degree distributions of the nodes for the second 9 year period

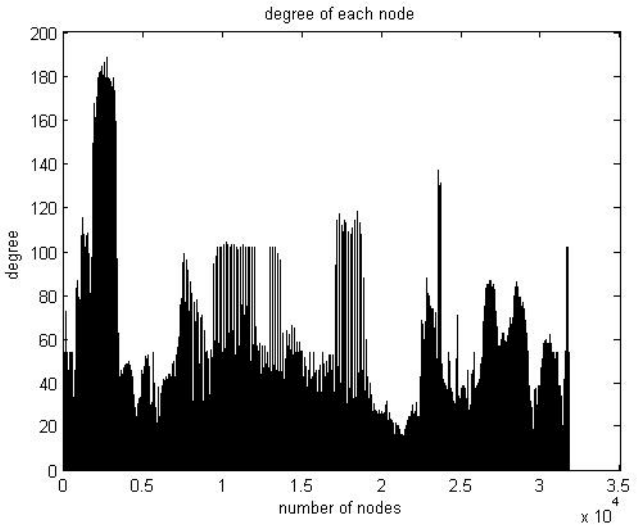


Fig. 1. Degree vs Number of Nodes

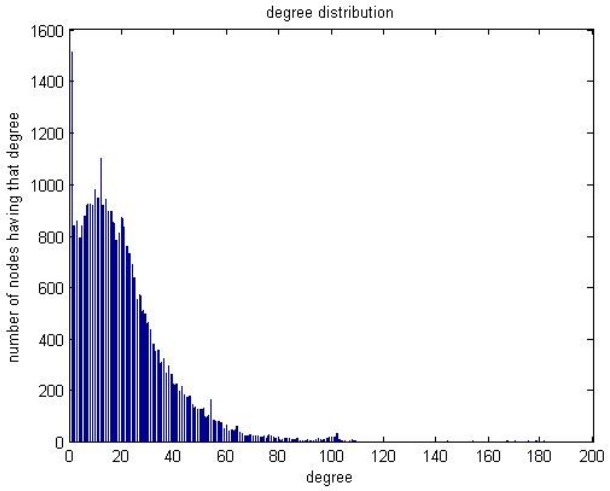


Fig. 2. Number of Nodes vs Degree

Following are the supernodes presented in the format (node index, degree):-

(3835,123),(3836,126),(3837,135),(3838,124),(3931,138),(3932,147),(3933,131),(3934,149),(3935,125),(4030,167),(4031,149),(4032,148),(4033,147),(4034,127),(4129,128),(4130,136),(4131,150),(4132,151),(4133,152),(4134,127),(4231,144),(4232,139),(4233,160),(4234,159),(4235,161),(4236,122),(4335,140),(4336,157),(4337,167),(4338,166),(4339,142),(4439,145),(4440,142),(4441,170),(4442,168),(4443,160),(4543,128),(4544,152),(4545,154),(4546,179),(4547,176),(4548,153),(4549,133),(4648,144),(4649,174),(4650,164),(4651,180),(4652,178),(4653,181),(4654,132),(4754,125),(4755,169),(4756,175),(4757,177),(4758,181),(4759,180),(4760,131),(4859,150),(4860,175),(4861,167),(4862,175),(4863,170),(4864,181),(4964,121),(4965,147),(4966,169),(4967,173),(4968,182),(4969,179),(4970,144),(5070,139),(5071,144),(5072,165),(5073,181),(5074,184),(5075,178),(5178,143),(5179,144),(5180,171),(5181,176),(5182,173),(5183,166),(5287,132),(5288,153),(5289,168),(5290,178),(5291,180),(5292,158),(5398,141),(5399,156),(5400,167),(5401,172),(5402,186),(5510,137),(5511,154),(5512,172),(5513,176),(5514,179),(5515,124),(5621,133),(5622,164),(5623,181),(5624,178),(5625,177),(5626,133),(5735,137),(5736,147),(5737,171),(5738,177),(5739,171),(5740,188),(5741,155),(5855,136),(5856,154),(5857,173),(5858,175),(5859,170),(5860,176),(5861,179),(5862,135),(5977,133),(5978,158),(5979,167),(5980,176),(5981,173),(5982,178),(5983,178),(5984,167),(6104,132),(6105,148),(6106,159),(6107,165),(6108,170),(6109,175),(6110,177),(6111,168),(6233,142),(6234,159),(6235,168),(6236,166),(6237,175),(6238,172),(6239,155),(6366,154),(6367,163),(6368,162),(6369,179),(6370,170),(6499,154),(6500,162),(6501,178),(6502,127),(6630,149

b) Clustering Coefficient $\gamma(G_r)$ = 0.6022431400549386
Characteristic Path Length L (G_r) = 44.32846852676166

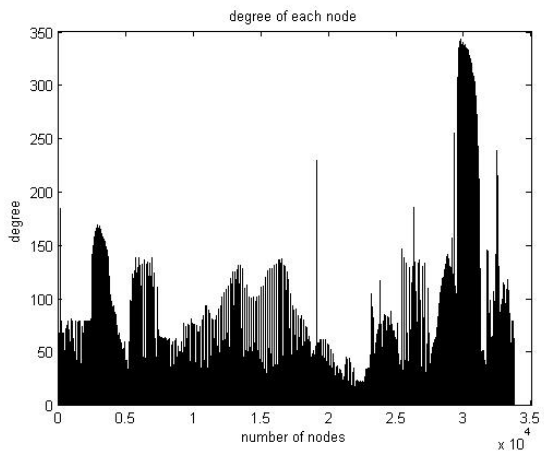
c) Clustering Coefficient for random graph $\gamma(G_{random})$ = 0.00068602
Characteristic Path Length for random graph L (G_{random}) = 3.3646
On comparing the values for the generated graph and the random graph, we observe that:-

1. The clustering coefficient of the generated graph is much higher than that of the random graph. This is because there is strong correlation between arctic ice densities as they are uniform in most places. Whereas, the data is random for a random graph showing negligible correlation.

2. The characteristic path length of the generated graph is around 14 times longer than that of the random graph as the graph for ice concentration is strongly connected to other nodes, both near and far.

Analysis for the third part

a) Following histograms show the degree distributions of the second nine-year period



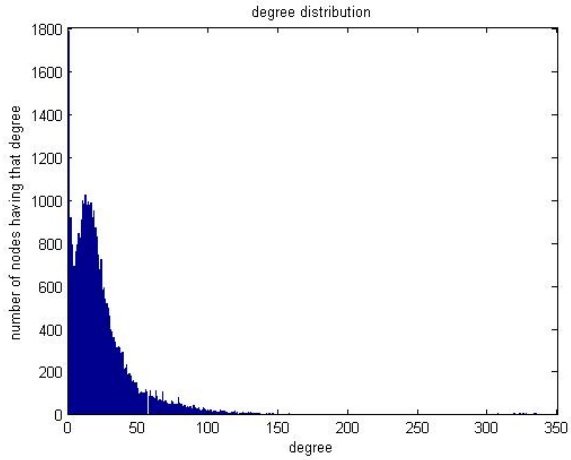


Fig. 1. Degree vs Number of Nodes

Fig. 2. Number of Nodes vs Degree

Following are the supernodes presented in the format (node index, degree):-

(26378,229),(40142,254),(40144,217),(40145,255),(40431,223),(40432,206),(40528,201),(40529,261),(40530,293),(40531,307),(40532,283),(40533,247),(40626,248),(40627,282),(40628,307),(40629,325),(40630,335),(40631,322),(40632,302),(40633,272),(40634,246),(40726,217),(40727,279),(40728,295),(40729,313),(40730,324),(40731,335),(40732,341),(40733,335),(40734,327),(40735,312),(40736,291),(40737,250),(40828,230),(40829,284),(40830,307),(40831,325),(40832,321),(40833,327),(40834,334),(40835,342),(40836,326),(40837,326),(40838,315),(40839,288),(40840,258),(40933,289),(40934,318),(40935,323),(40936,330),(40937,328),(40938,333),(40939,337),(40940,334),(40941,324),(40942,321),(40943,307),(40944,294),(40945,260),(41034,264),(41035,314),(41036,319),(41037,330),(41038,332),(41039,329),(41040,335),(41041,339),(41042,338),(41043,325),(41044,319),(41045,306),(41046,295),(41047,252),(41136,212),(41137,283),(41138,320),(41139,323),(41140,330),(41141,329),(41142,333),(41143,336),(41144,335),(41145,333),(41146,323),(41147,319),(41148,310),(41149,280),(41150,213),(41236,235),(41237,307),(41238,326),(41239,331),(41240,327),(41241,334),(41242,333),(41243,338),(41244,337),(41245,333),(41246,324),(41247,312),(41248,305),(41249,275),(41331,263),(41332,318),(41333,326),(41334,325),(41335,332),(41336,334),(41337,334),(41338,334),(41339,333),(41340,322),(41341,321),(41342,310),(41343,297),(41344,219),(41427,285),(41428,323),(41429,327),(41430,328),(41431,330),(41432,334),(41433,327),(41434,327),(41435,323),(41436,321),(41437,311),(41438,295),(41439,273),(41440,237),(41524,224),(41525,305),(41526,325),(41527,327),(41528,332),(41529,333),(41530,325),(41531,317),(41532,318),(41533,319),(41534,318),(41535,301),(41536,289),(41537,278),(41538,244),(41614,257),(41615,313),(41616,324),(41617,330),(41618,333),(41619,325),(41620,321),(41621,314),(41622,312),(41623,312),(41624,311),(41625,305),(41626,290),(41627,271),(41628,251),(41629,206),(41701,272),(41702,325),(41703,326),(41704,328),(41705,329),(41706,329),(41707,323),(41708,302),(41709,318),(41710,308),(41711,311),(41712,292),(41713,286),(41714,267),(41715,251),(41716,203),(41780,281),(41781,322),(41782,324),(41783,327),(41784,327),(41785,320),(41786,324),(41787,311),(41788,308),(41789,312),(41790,303),(41791,284),(41792,280),(41793,263),(41794,263),(41795,228),(41862,287),(41863,327),(41864,325),(41865,323),(41866,319),(41867,317),(41868,316),(41869,307),(41870,297),(41871,298),(41872,295),(41873,287),(41874,276),(41875,262),(41876,245),(41877,247),(41946,295),(41947,323),(41948,319),(41949,319),(41950,315),(41951,310),(41952,296),(41953,284),(41954,284),(41955,281),(41956,283),(41957,265),(41958,274),(41959,250),(41960,249),(41961,238),(41962,202),(42029,299),(42030,318),(42031,317),(42032,309),(42033,304),(42034,267),(42035,273),(42036,247),(42037,250),(42038,236),(42039,231),(42040,288),(42041,270),(42042,229),(42043,218),(42120,296),(42121,320),(42122,315),(42123,301),(42124,278),(42125,240),(42126,206),(42129,204),(42130,204),(42131,255),(42132,210),(42216,298),(42217,311),(42218,308),(42219,300),(42220,263),(42221,223),(42316,286),(42317,308),(42318,302),(42319,280),(42320,227),(42417,274),(42418,303),(42419,292),(42420,266),(42517,253),(42518,294),(42519,287),(42520,247),(42616,248),(42617,289),(42618,268),(42619,230),(42710,206),(42711,272),(42712,261),(42713,225),(42803,267),(42804,267),(42805,213),(42894,233),(42895,258),(42896,228),(42981,202),(42982,243),(42983,202),(43069,215),(43070,213),(43244,212),(45605,200),(45687,210),(45688,239),(45766,215

b) Clustering Coefficient $\gamma(G_r)$ = 0.6128214862035625
Characteristic Path Length L (G_r) = 38.179834999569245

c) Clustering Coefficient for random graph $\gamma(G_{random})$ = 0.00080388
Characteristic Path Length for random graph L (G_{random}) = 3.1606

On comparing the values for the generated graph and the random graph, we observe that:-

1. The clustering coefficient of the generated graph is much higher than that of the random graph. This is because there is strong correlation between arctic ice densities as they are uniform in most places. Whereas, the data is random for a random graph showing negligible correlation.

2. The characteristic path length of the generated graph is around 12 times longer than that of the random graph as the graph for ice concentration is strongly connected to other nodes, both near and far.

3. Now the data is analyzed by considering a time lag of 1, 2, 3 and 4 weeks

Time lag of 1 week

a)

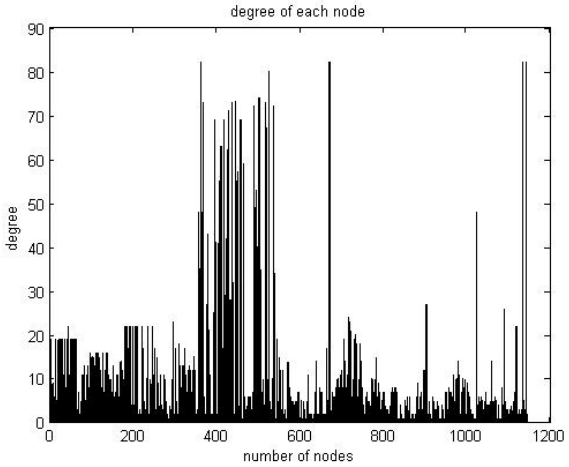


Fig. 1. Degree vs Number of Nodes

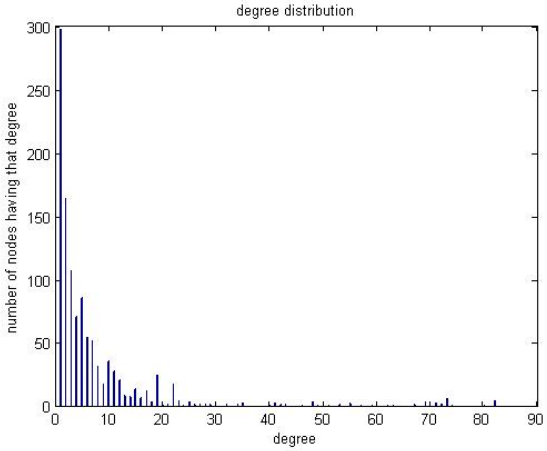


Fig. 2. Number of Nodes vs Degree

Following are the supernodes presented in the format (node index, degree):-

(7525,82),(9791,74),(9985,80),(23927,82),(23928,82),(47831,82),(47858,82)

b) Clustering Coefficient $\gamma(G_r)$ = 0.8124893400650662

Characteristic Path Length L (G_r) = 0.08109516082818895

c) Clustering Coefficient for random graph $\gamma(G_{random})$ = 0.0070

Characteristic Path Length for random graph L (G_{random}) = 3.3745

On comparing the values for the generated graph and the random graph, we observe that:-

1. The clustering coefficient of the generated graph is much higher than that of the random graph. This is because there is strong correlation between arctic ice densities as they are uniform in most places. Whereas, the data is random for a random graph showing negligible correlation.

2. The characteristic path length of the random graph is much longer than that of the generated graph as the graph for ice concentration is strongly connected to other nodes, both near and far.

Time lag of 2 weeks

b)

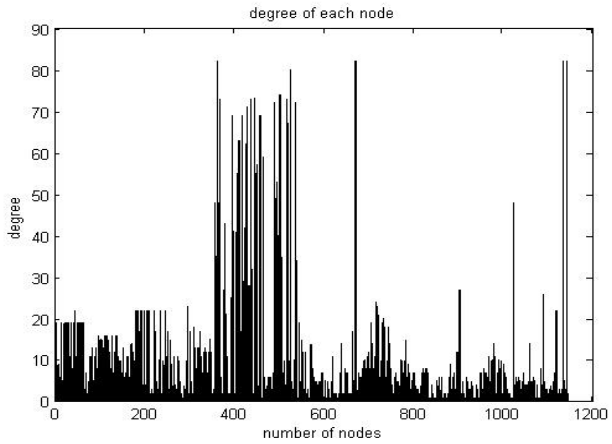


Fig. 1. Degree vs Number of Nodes

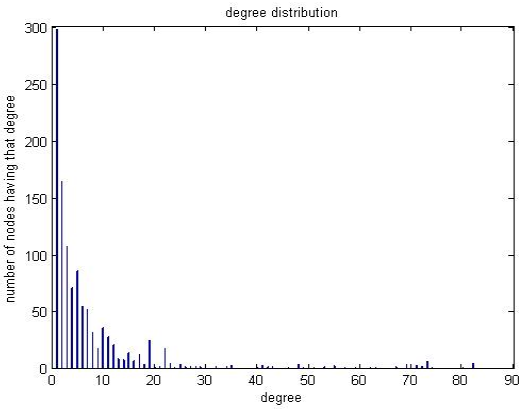


Fig. 2. Number of Nodes vs Degree

Following are the supernodes presented in the format (node index, degree):-

(7525,82),(9791,74),(9985,80),(23927,82),(23928,82),(47831,82),(47858,82)

b) Clustering Coefficient $\gamma(G_r)$ = 0.2344

Characteristic Path Length L (G_r) =0.0410166

c) Clustering Coefficient for random graph $\gamma(G_{random})$ = 0.0025

Characteristic Path Length for random graph L (G_{random}) =5.6214

On comparing the values for the generated graph and the random graph, we observe that:-

1. The clustering coefficient of the generated graph is much higher than that of the random graph. This is because there is strong correlation between arctic ice densities as they are uniform in most places. Whereas, the data is random for a random graph showing negligible correlation.

2. The characteristic path length of the random graph is much longer than that of the generated graph as the graph for ice concentration is strongly connected to other nodes, both near and far.

Time lag of 3 weeks

c)

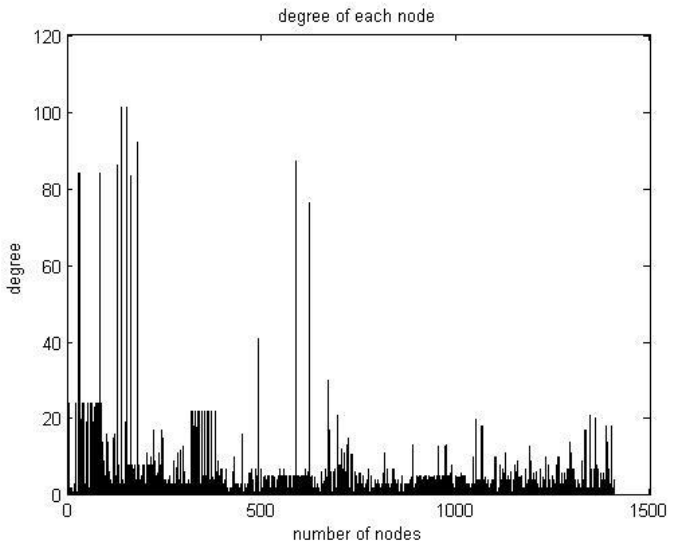


Fig. 1. Degree vs Number of Nodes

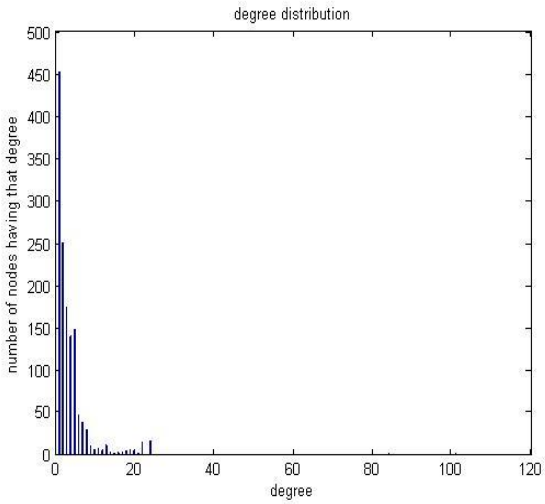


Fig. 2. Number of Nodes vs Degree

Following are the supernodes presented in the format (node index, degree):-

(315,84),(711,84),(1252,86),(1359,101),(1476,101),(1544,83),(1691,92),(8410,87),(8769,76)

b) Clustering Coefficient $\gamma(G_r)$ = 0.4939

Characteristic Path Length L (G_r) = 0.05372

c) Clustering Coefficient for random graph $\gamma(G_{random})$ = 0.0032

Characteristic Path Length for random graph L (G_{random}) = 4.8177

On comparing the values for the generated graph and the random graph, we observe that:-

1. The clustering coefficient of the generated graph is much higher than that of the random graph. This is because there is strong correlation between arctic ice densities as they are uniform in most places. Whereas, the data is random for a random graph showing negligible correlation.

2. The characteristic path length of the random graph is much longer than that of the generated graph as the graph for ice concentration is strongly connected to other nodes, both near and far.

Time lag of 4 weeks

d)

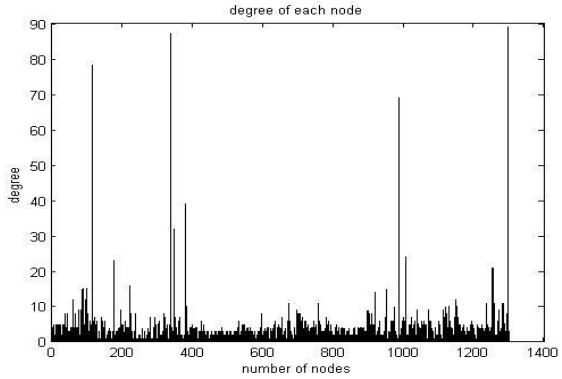


Fig. 1. Degree vs Number of Nodes

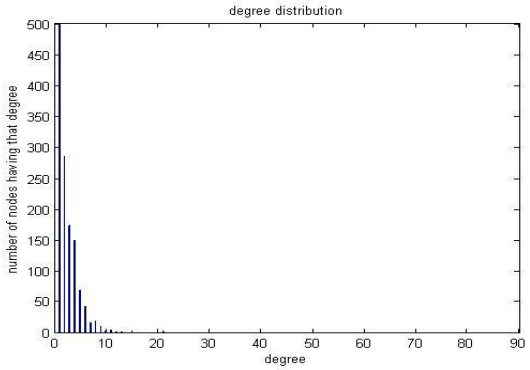


Fig. 2. Number of Nodes vs Degree

Following are the supernodes presented in the format (node index, degree):-

(1352,78),(5011,87),(5114,32),(5961,39),(33683,69),(47869,89)

b) Clustering Coefficient $\gamma(G_r)$ = 0.3926

Characteristic Path Length L (G_r) = 0.05372

c) Clustering Coefficient for random graph $\gamma(G_{random})$ = 0.0023

Characteristic Path Length for random graph L (G_{random}) = 6.4813

On comparing the values for the generated graph and the random graph, we observe that:-

1. The clustering coefficient of the generated graph is much higher than that of the random graph. This is because there is strong correlation between arctic ice densities as they are uniform in most places. Whereas, the data is random for a random graph showing negligible correlation.

2. The characteristic path length of the random graph is much longer than that of the generated graph as the graph for ice concentration is strongly connected to other nodes, both near and far.

Optimizations

The following optimizations have been performed in the project:-

1. While comparing various cells for calculating S_{xy}, comparisons have only been made in the upper diagonal of the matrix representing them. This eliminates redundancy by not considering repetitive pairs. For instance, the comparison between (cell 3, cell 5) is same as that between (cell 5, cell 3). Only one of these is considered.
2. The graph itself uses adjacency list instead of adjacency matrix leading to better space complexity.
3. The nodes having 0 values have been ignored in calculating S_{xy} which reduces computations.
4. While calculating the shortest path, Breadth First Search has been used Instead of running the BFS across 2 vertices at a time. We have only considered one vertex at a time and each BFS calculates its shortest distance from all other nodes.

Parallelization

Multithreading in Java has been implemented so that the code can run on multiple threads simultaneously, thereby greatly decreasing computation time. In this project, all programs have been run on 16 threads simultaneously. The data has been divided into 16 portions by breadth (448/16).

Complexity Calculation:

The major space is consumed by the data itself which is represented as a 302x448x1404 array therefore. The space complexity is therefore O(row x col x weeks) as all the other data structures are negligible compared to the data.

The time complexity is O((row x col)^2 x weeks) in the formations of the graph.

The complexity of gamma calculation and length calculation is negligible compared to this.