ARIZONA STATE UNIVERSITY

CSE 310, SLN 70793 — **Data Structures and Algorithms** — Fall 2016

Project #3

Available 10/31/2016; milestone due 11/14/2016; complete project due 11/28/2016

In recent years there has been a great deal of interest in so-called "small-world" graphs [7]. These types of graphs are characterized by a high degree of local clustering and a small number of long-range connections, making them very efficient at transferring information. This "small-world" architecture underlies the well-known "six degrees of separation" phenomenon, in which it is believed that a connection can be found between any two people on the planet requiring no more than six intermediate links.

In this project, we represent climatological data as a graph and compute some characteristics of it to determine if it is a small-world graph.

Note: This project is to be completed **individually**. Your implementation **must** use C/C++ and your code **must** compile and run on the Linux machine `general.asu.edu`.

All dynamic memory allocation **must** be done yourself, i.e., using either `malloc()` and `free()`, or `new()` and `delete()`. You may not use any external libraries to implement any part of this project.

As always, a script will be used to check the correctness of your program. Therefore, absolutely no changes to these project requirements are permitted.

Use a version control system as you develop your solution to this project, e.g., Dropbox, Bitbucket, or GitHub. Your code repository should be *private* to prevent anyone from plagiarizing your work.

1 Arctic Sea Ice

Sea ice covers most of the Arctic Ocean and plays a significant role in the global water cycle and the global energy balance. It is also considered to be a sensitive indicator of climate change. Thus, any changes in the Earth's climate are likely to first be seen in areas such as the high Arctic.

Since the 1970s, the areal extent of sea ice has been shrinking. In September of 2007, the mean sea ice extent was 1.65 million square miles, which was the lowest ever recorded for the month of September, shattering the previous record in 2005 by 23%. Current climate model projections indicate that the Arctic could be seasonally ice-free by 2050–2100, which will significantly impact the global climate [2].

Because of its importance as a proxy indicator of climate change, a great deal of research is conducted on Arctic sea ice. Data acquired by meteorological satellites provides one of the most effective ways to study large-scale changes in sea ice conditions in the Arctic. The longest continuous satellite record of sea ice comes from the Nimbus-7 Scanning Multi-channel Microwave Radiometer (SMMR) and Defense Meteorological Satellite Program Special Sensor Microwave/Imager (DMSP SSM/I) series of meteorological satellites. Data acquisition started in late 1978, with the first full year of data in 1979, and is ongoing. This data is maintained for the Arctic and Antarctic by the National Snow & Ice Data Center (NSIDC); see `http://nsidc.org/data/seaice_index/`.

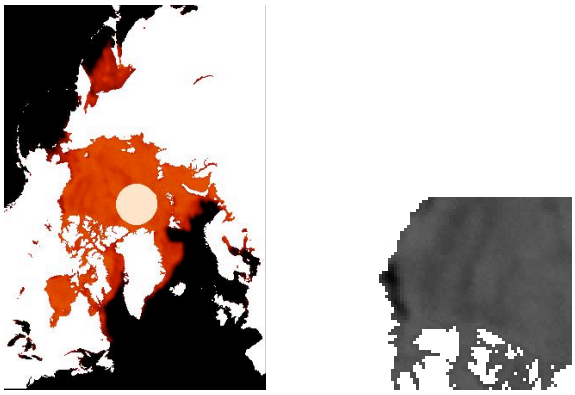
The sea ice concentration (SIC) anomaly data set that we will use consists of 27 years (1979–2005) of weekly SIC anomaly data derived from the SMMR-SSM/I passive microwave data set. An *anomaly data set* is when the long-term average is subtracted from the data, to remove seasonal trends, making the data more amenable to statistical analysis.

The data for each week is for a 304×448 floating point array representing the northern hemisphere. The data value at each cell (x, y) in the array represents the percentage of deviation in ice concentration from the 27-year average for a given week. For example looking at the values of the array for week 30 of 1990, at cell $(100, 200)$, the value is -4.5. This means that at cell $(100, 200)$ for week 30 of 1990, the sea ice concentration was 4.5% lower than the 27-year average value for week 30 for that cell.

Since there are 52 weeks per year and 27 years of data, there are $52 \times 27 = 1,404$ sea ice concentration readings for each position (x, y) over the years. Therefore, for each of the $304 \times 448 = 136,192$ positions there is a time series $[x, y, t]$, $1 \leq t \leq 1,404$, of SIC data with 1,404 values, starting at week 1 of 1979.

The data set is given as 1,404 files each containing a 304×448 32-bit floating point array (little-endian byte order). The format of the filenames is: `diffwNNyYYYY+landmask`, where `wNN` denotes week `NN` and `yYYYY` denotes the year. For example, `diffw31y1983+landmask` is the file for week 31 of 1983. The "`+landmask`" part of the name indicates that a landmask was applied to the data.

Since we are dealing with sea ice, land masses can be ignored; these constitute approximately half of the cells in each of the arrays. Land is denoted by the value 168. Missing data is denoted by the value of 157. Figure 1(a) is a sample SSM/I sea ice concentration image, which has been pseudo-coloured to make it easier to view. Each pixel corresponds to a nominal physical area of 25 square kilometers. There is a large circular disk over the North Pole, an area of missing data due to the satellite's orbit. The satellite orbits from pole to pole (i.e., longitudinally), but at an incline, so there is a circular area that is not covered. Hence, the only missing data is in the circular region over the North Pole.



(a) Full Arctic region

(b) Beaufort Sea subregion

Figure 1: Sample SSM/I total sea ice concentration image.

Figure 1(b) shows a subregion near the Beaufort Sea. The corresponding data set has only $63 \times 63 = 3969$ cells for each of the 27 years. This smaller data set is otherwise identical to the full data set and should be used for code development and testing.

1.1 A Graph Representation for the Climatological Data Set

Recall that our data set consists of arrays representing sea ice concentration for each week for the 27 years 1979–2005. The data set can be thought of as a stack of two-dimensional (304×448) arrays, ordered according to time. Since there are 27 years in the data set and the data was weekly, there are $52 \times 27 = 1,404$ arrays in the stack. The data set can therefore be visualized as a three-dimensional array consisting of a stack of two dimensional arrays. For each cell (x, y), there is a time series $[x, y, t]$, $1 \leq t \leq 1,404$, of values in the stack.

How is a graph constructed from this type of climatological data? Tsonis et al. derived a correlation-based graph $G = (V, E)$ [1, 6]. The vertex set V corresponds to the cells, i.e., for the full SIC data set, the graph has 302×448 vertices. To determine the edge set, the Pearson correlation coefficient (see §1.1.1) is calculated between all pairs of cells (x, y) and (x', y'), $1 \leq x, x' \leq 304$, $1 \leq y, y' \leq 448$, $(x, y) \neq (x', y')$, of time series vectors. That is, the correlation coefficient is computed between $[x, y, t]$ and $[x', y', t]$, $1 \leq t \leq 1,404$, for each possible pair of cells. Since there are $n = 136,192$ cells, there are $n(n-1)/2$ pairs of cells, and so the correlation coefficient is calculated for 3,547,116 pairs of time series. If the correlation coefficient for a pair of cells (x, y) and (x', y') of time series, i.e., $[x, y, t]$ and $[x', y', t]$, $1 \leq t \leq 1,404$, is greater than some

threshold r_{thresh}, then an edge is inserted between cells (x, y) and (x', y'). The final result is a graph with edges between all cells having a correlation greater than the threshold r_{thresh}.

Use an *adjacency list* of represent the graph because it is expected to be reasonably sparse.

1.1.1 Pearson Correlation Coefficient

To get a measure of how strongly two vectors X and Y, of length n, are related, we use the correlation coefficient. Correlation is concerned with trends: if X increases, does Y tend to increase or decrease? How much? How strong is this tendency?

The Pearson correlation coefficient measures the strength and direction of a linear relationship between X and Y. The formula for the *sample correlation coefficient*, denoted by r is:

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

where,

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

In our case the vector X corresponds to the time series of length 1,404 for cell (x, y), whereas the vector Y corresponds to the time series of the same length for cell (x', y'). You will need to compute the sample correlation coefficient between all pairs of cells (x, y) and (x', y'), $(x, y) \neq (x', y')$. For a given pairs of cells (x, y) and (x', y'), if $|r| \geq r_{thresh}$ then you should insert an edge in the graph between the vertex corresponding to (x, y) and the vertex corresponding to (x', y'). (The absolute value $|r|$ of the correlation coefficient should be used, since r can be positive or negative.)

2 Analyses of the Climatological Graph

We are interested in whether the graph derived from the SIC climatological data is a small-world graph. First, construct a correlation-based graph $G_r = (V_r, E_r)$ for the sea ice anomaly dataset for each correlation threshold $r_{thresh} \in \{0.95, 0.925, 0.9\}$. (Start with the largest threshold as it will yield the sparsest graph.)

Now, for each correlation-based graph G_r:

1. Plot the histogram of the degree distribution of G_r. If $|V_r| = n$ vertices, the degree of a vertex can range from zero (the vertex is isolated) to $n - 1$ (the vertex has an edge to every other vertex in the graph). A histogram of the degree distribution for G_r therefore plots a count of the number of vertices of each degree d, $0 \leq d \leq n - 1$.

   Recall that a small-world graph is characterized by a high degree of local clustering and a small number of long-range connections. As a result, we expect degree distribution of a small-world graph to have a long-tailed distribution.

2. Compute the number of connected components in G_r and their size (i.e., number of vertices in each component). G_r with n vertices may consist of from one to n connected components. A connected component, $C = (V_C, E_C)$, is a subgraph of G_r such that $V_C \subseteq V_r$ and $E_C \subseteq E_r$. For any pair of vertices $u, v \in V_C$, there exists a path from u to v consisting of a finite number of edges in E_C. Similarly, every

edge in E_C has endpoints in the vertices of C. Therefore, every vertex within a component is reachable from any other vertex in that component, but there are no paths between components. Discovery of connected components is implemented in this project using a recursive depth-first traversal of G_r.

For a small-world graph, what do you think the component structure should look like?

3. Another way to determine if the graph G_r is a small-world graph is to calculate the mean *clustering coefficient* $\gamma(G)$ and the *characteristic path length* (or diameter), $L(G)$, of G_r and compare it to a random graph G_{random} of the same size (i.e., the same number of vertices). (These characteristics are defined in §2.1 and in §2.2, respectively.) For a small-world graph, $\gamma(G_r) \gg \gamma(G_{random})$ and $L(G_r) \geq L(G_{random})$ [5, 7].

   Compute the clustering coefficient, $\gamma(G_r)$, and the characteristic path length, $L(G_r)$, of the graph G_r.

4. Compare $\gamma(G_r)$ and $L(G_r)$ to $\gamma(G_{random})$ and $L(G_{random})$ for the random graph, G_{random}, of comparable size. (See §2.3 for details.)

For the milestone an full project deadline, you will need to write a report that summarizes your findings for the 27-year period. Be sure to work on the "small" Beaufort Sea data set first.

2.1 Clustering Coefficient

The neighbourhood $N(v)$ of a vertex v consists of all the vertices adjacent to v. The graph generated by $N(v)$, $\langle N(v) \rangle$ has vertex set $N(v)$ and its edges are all edges of the graph with both endpoints in $N(v)$. Use $k(v)$ and $e(v)$ to denote the numbers of vertices and edges in $\langle N(v) \rangle$. The *clustering coefficient* γ_v of v is:

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

In words, γ_v for a vertex v is the proportion of edges between the vertices within its neighbourhood divided by the number of edges that could possibly exist between them.

The *clustering coefficient of a graph* G is the mean of the clustering coefficient of all vertices of G and is denoted $\gamma(G)$.

2.2 Characteristic Path Length

Let $d_{i,j}$ be the length of the shortest path between the vertices i and j. Then the *characteristic path length* $L(G)$ for the graph $G = (V, E)$, is $d_{i,j}$ averaged over all $\binom{n}{2}$ pairs of vertices, where $n = |V|$. The Floyd-Warshall all-pairs shortest paths algorithm may be useful here.

2.3 Metrics for Random Graphs

A random graph G_{random} corresponding to G_r has the same number of vertices as G_r, namely $n = |V_r|$. However, edges are inserted into G_{random} at random such that the edge density of G_{random} matches the edge density of G_r, i.e., the probability p that edges are present in G_{random} should match that of G_r. This edge probability p is calculated from G_r having n vertices and mean vertex degree k as:

$$p = \frac{\frac{nk}{2}}{\binom{n}{2}} = \frac{\frac{nk}{2}}{\frac{n(n-1)}{2}} = \frac{k}{n-1}.$$

This calculation is derived intuitively by reasoning that the fraction of actual edges $\frac{nk}{2}$ out of the total number of possible edges $\binom{n}{2}$ should approximate the edge probability of the graph.

The clustering coefficient of a random graph G_{random} is $\gamma(G_{random}) = \frac{k}{n}$. Similarly, the characteristic path length of a random graph G_{random} is $L(G_{random}) = \frac{\log n}{\log k}$. Here, n is the total number of vertices in the correlation graph G_r, and k is the mean vertex degree of G_r.

Given the degree distribution of G_r it is straightforward to compute k.

2.4 Submission Instructions

All submissions are electronic. This project has two submission deadline dates.

1. The milestone deadline date is Monday, 11/14/2016.
2. The final project deadline date is on Monday, 11/28/2016.

It is your responsibility to submit your project well before the time deadline!!! Late projects will not be accepted. Do not expect the clock on your machine to be synchronized with the one on Blackboard!

Multiple submissions are allowed. The last submission will be graded.

2.5 Requirements for Milestone Deadline

By the milestone deadline, your project must compute the degree distribution and number and size of connected components in a correlation based graph derived from the "small" Beaufort Sea SIC data set.

Submit electronically, on Monday, 11/14/2016 using the submission link on Blackboard for the Project #3 milestone, a zip[1] file named `yourFirstName-yourLastName.zip` containing the following:

Project State (5%): In a folder (directory) named `State` provide a brief report (.txt, .doc, .docx, .pdf) that addresses the following:

1. Explain all design decisions. Discuss your representation of the graph, and any optimizations you made in computations. (Depending on the order in which you calculate the statistics, you can likely save and make use of previous results to cut down on the computation time.) Can you compute a worst-case bound on the time and/or space of your algorithms?
2. Describe any problems encountered in your implementation for this project milestone.
3. Describe any known bugs in your project milestone.
4. While this project is to be completed individually, describe any significant interactions with anyone (peers or otherwise) that may have occurred.
5. Cite any external code bases, books, and/or websites used or referenced.

Implementation (50%): In a folder (directory) named `Code` provide:

1. In one or more files, your well documented C/C++ source code implementing this project milestone.
2. A `Makefile` that compiles your program to an executable named `p3` on the Linux machine `general.asu.edu`. Our TA will write a script to compile and run all student submissions on `general.asu.edu`; therefore executing the command `make p3` in the `Code` directory must produce the executable `p3` also located in the `Code` directory.

Report (20%): In a folder (directory) named `Report` provide a report (.doc, .docx, .pdf) containing the following:

1. For the "small" data set, i.e., the subregion in the Beaufort Sea, for each correlation threshold $r_{thresh} \in \{0.95, 0.925, 0.9\}$, plot the degree distribution.
   (a) Do you think the degree distribution is consistent with that of a small-world graph? Why or why not?

[1] **Do not** use any other archiving program except `zip`.

(b) Identify any *supernodes*, i.e., vertices with significantly higher vertex degree than the average, and where they occur. Describe your determination of supernode.

2. For the "small" data set, i.e., the subregion in the Beaufort Sea, for each correlation threshold $r_{thresh} \in \{0.95, 0.925, 0.9\}$, compute the number of connected components in G_r and their size (i.e., number of vertices).
   (a) For a small-world graph, how do you think the component structure should look?
   (b) Do your results support your hypothesis?

Correctness (25%): The correctness of your program will be determined by running your program on the "small" data set for a given threshold.

The milestone is worth 30% of the total project grade.

2.6 Requirements for Final Project Deadline

For the full project deadline, your project must perform all the analyses described in §2 a correlation based graphs derived from the "small" Beaufort Sea SIC data set. If you can run the analyses on the full data set, a bonus of up to 10% is available.

Submit electronically, on Monday, 11/28/2016 using the submission link on Blackboard for the complete Project #3, a zip[2] file named `yourFirstName-yourLastName.zip` containing the following:

Project State (5%): Follow the same instructions for Project State as in §2.5.

Implementation (40%): Follow the same instructions for Implementation as in §2.5.

Report (25%): In addition to the analyses run for the Report for the milestone deadline (§2.5), you must also run the following analyses:

1. For the "small" data set, i.e., the subregion in the Beaufort Sea, for each correlation threshold $r_{thresh} \in \{0.95, 0.925, 0.9\}$, compute the clustering coefficient, $\gamma(G_r)$, and the characteristic path length, $L(G_r)$, of the graph G_r.
2. Compute the clustering coefficient, $\gamma(G_{random})$, and the characteristic path length, $L(G_{random})$, of a random graph G_{random} corresponding to each G_r.
   (a) Compare $\gamma(G_r)$ and $L(G_r)$ to $\gamma(G_{random})$ and $L(G_{random})$ for the random graph, G_{random}, of comparable size.
   (b) What conclusion can you draw from your results?

Correctness (30%): The same instructions for Correctness as in §2.5 apply.

Bonus (10%): Repeat the analyses on the full data set. (I know that this is *a lot* of extra work for a bonus; it is a good indicator of the scalability of your code.)

3 Marking Guide

The project milestone is out of 100 marks.

Project State (5%): Summary of project state, use of a zip file, and directory structure required (i.e., a folder/directory named `State`, `Code`, and `Report` is provided).

Implementation (50%): 40% for the quality of implementation in your code including proper memory management, 10% for a correct `Makefile`.

[2] **Do not** use any other archiving program except `zip`.

Report (20%): 20% for the requested plots, and answers to questions asked in §2.5.

Correctness (25%): 20% for correct output on several files of sample input, 5% for correct processing of files.

The full project is out of 100 marks.

Project State (5%): Summary of project state, use of a zip file, and directory structure required (i.e., a folder/directory named `State`, `Code`, and `Report` is provided).

Implementation (40%): 30% for the quality of implementation in your code, 10% for a correct `Makefile`.

Report (25%): 25% for the requested plots, and answers to the questions asked in §2.5 and 2.6.

Correctness (30%): 25% for correct output on the "small" Beaufort Sea data set..

Bonus (10%): A repeat of the analyses on the full data set.

Acknowledgements

Thanks to my former student Wayne S. Chan who motivated this project.

References

[1] J. P. Onnela, K. Kaski, and J. Kertész. Clustering and information in correlation based financial networks. *European Physical Journal B*, 38(2):353–362, March 2004.

[2] J. T. Overpeck, M. Sturm, J. A. Francis, D. K. Perovich, M. C. Serreze, R. Benner, E. C. Carmack, F. S. Chapin III, S. C. Gerlach, L. C. Hamilton, L. D. Hinzman, M. Holland, H. P. Huntington, J. R. Key, A. H. Lloyd, G. M. MacDonald, J. McFadden, D. Noone, T. D. Prowse, P. Schlosser, and C. Vorosmarty. Arctic system on trajectory to new, seasonally ice-free state. *Earth Observation Science*, 86(34):309–316, August 2005.

[3] A. A. Tsonis. Is global warming injecting randomness into the climate system? *Earth Observation Science*, 85(38):361–364, September 2004.

[4] A. A. Tsonis and P. J. Roebber. The architecture of the climate network. *Physica A*, 333:497–504, 2004.

[5] A. A. Tsonis, K. L. Swanson, and P. J. Roebber. What do networks have to do with climate? *Bulletin of the American Meteorological Society*, pages 585–595, May 2006.

[6] M. Tumminello, D. Matteo, T. Aste, and R. N. Mantegna. Correlation based networks of equity returns sampled at different time horizons. *European Physical Journal B*, 55:209–217, 2007.

[7] D. J. Watts and S. H. Strogatz. Collective dynamics of "small-world" networks. *Nature*, 393:440–442, 1998.