ARIZONA STATE UNIVERSITY

CEN 502, SLN 79367 — **Computer Systems II** — Fall 2015

Dr. Violet R. Syrotiuk

Project #1: Algorithm Design

Available 08/24/2015 due 09/18/2015; Group formation by 08/28/2015 (see §3)

1 What do Graphs have to do with Climate?

Tsonis et al. [1, 2, 3] are among the first to apply the theory of graphs to climate data, specifically to a National Centers for Environmental Protection (NCEP) wind anomaly gridded dataset. An anomaly or deviation dataset is when the long-term average is subtracted from the data, giving the deviation from the long-term average.

The dataset used consisted of monthly arrays representing the height of the 500 hPa (hectopascal) pressure level for 55 years (1950-2004). Roughly speaking, the height of the 500 hPa pressure level is a representation of the general wind circulation of the atmosphere. The 500 hPa dataset can be thought of as a stack of two-dimensional (72×37) arrays, ordered according to time. Since there are 55 years in the dataset and the data was monthly, there are 55 years $\times$ 12 months per year = 660 arrays in the stack.

The dataset 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 $[x, y, t]$, $1 \leq t \leq 660$, of values in the stack.

The question is: how are graphs used with this type of data? Tsonis et al. derived a correlation-based graph $G = (V, E)$ [4, 5] from this dataset. The vertex set V corresponds to the cells. To determine the edge set, the Pearson linear correlation coefficient is calculated between all pairs of cells (x, y) and (x', y'), $1 \leq x, x' \leq 72$, $1 \leq y, y' \leq 37$, $(x, y) \neq (x', y')$, of time series. That is, the correlation coefficient is computed between $[x, y, t]$ and $[x', y', t]$, $1 \leq t \leq 660$, for each possible pair of cells. Since there are $n = 2664$ cells, there are $n(n-1)/2$ pairs of cells, and so the correlation coefficient was 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 $[x', y', t]$, $1 \leq t \leq 660$, is greater than some threshold, 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.

In recent years, there has been a great deal of interest in so-called "small-world" graphs, pioneered by Watts and Strogatz [6]. 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. (See `http://www.columbia.edu/cu/news/media/01/duncanWatts/` for more information about small-world graphs.)

To determine whether the graph G is a small-world graph, the mean clustering coefficient $\gamma(G)$ and the characteristic path length (or diameter), $L(G)$, are calculated and compared to a random graph G_{random} of the same size. For a small-world graph, $\gamma(G) \gg \gamma(G_{random})$ and $L(G) \geq L(G_{random})$ [6]. Tsonis et al. applied these statistics to the derived climate graph and found that it had a small-world architecture [3].

For this project you will do similar analysis, but apply it to a different type of climatological data — namely, to a sea ice concentration anomaly dataset.

2 Sea Ice Concentration Anomaly Data

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 Earths 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 is the lowest ever recorded for the month of September, shattering

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

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 [7]. (For a NASA animation of annual sea ice extent, see `http://www.nasa.gov/vision/earth/environment/seaice_meltdown.html`.)

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.

The sea ice concentration (SIC) anomaly dataset that we will use consists of 27 years (1979–2005) of weekly SIC anomaly data derived from the SMMR-SSM/I passive microwave dataset. An anomaly dataset 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 a 304×448 floating point array representing the Northern Hemisphere. The data value at each cell (x, y) in an array represents the percentage of deviation in ice concentration from the 27-year average for a given week. For example, suppose we look at 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, 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 $\times$ 27 years, there are 1,404 arrays in the data stack. Each array has 304 columns and 448 rows for a total of 136,192 cells. Therefore, there are 136,192 time series, and each time series $[x, y, t]$, $1 \leq t \leq 1,404$, and each time series contains 1,404 values, starting at week 1 of 1979.

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 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 sq. km. There is a large circular disk over the North Pole, an area of missing data due to the satellites 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.

The dataset consists of 1,404 data 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.

3 Group Formation

All projects in CEN 502 are **group** projects. Form a group of three (3) and submit by e-mail the names of the group members to our TA, Tarun Chanana by Friday 08/28/2015. He will assign a group number to each group; this group number will be used in the submission process.

In forming your group consider the following. Your friends probably have a similar background to you. This may not be advantageous to you for working on projects in this course. I suggest to find group members of complementary strength in each of the three areas of this course, namely algorithm design, computer networks, and computer architecture. This may ease project development.

It is expected that group members contribute **equally** to each project. You will be asked to *honestly* assess the contribution of members to me. If at any time you consider the group dysfunctional, inform me immediately. I can solve the problem only if I know about it.

4 Tasks

You may use a programming language of your choice to solve this project.

1. Similar to the NCEP wind anomaly gridded dataset described in §1, construct a correlation-based graph $G_r = (V_r, E_r)$ for the complete sea ice anomaly dataset for each correlation threshold $r \in \{0.9, 0.95\}$. (The absolute value of the correlation coefficient $|r|$ should be used, since r can be positive or negative.)

   Now, for each correlation-based graph G_r:

   (a) Plot the histogram of the degree distribution of G_r. Identify any *supernodes*, i.e., vertices with significantly higher vertex degree than the average, and where they occur. (You might project them onto a map of the Earth to better visualize the supernodes.)

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

   (c) 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.

2. Now, split the dataset into three equal parts of nine years each. Repeat Task 1 for each of these nine-year periods.

3. Finally, repeat Task 1, except that when computing the correlation-based graph, consider a time lag of $s \in \{1, 2, 3, 4\}$ weeks.

4. Write a report that summarizes your findings for the 27-year period, and the three 9-year periods. Compare and contrast the differences in the graph statistics between the 27-year period, the 9-year periods, and the lagged correlation. Add a section to the report that discusses 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. Given time, it is even possible to parallelize some of the computation (say, using `OpenMP`). Can you compute a worst-case bound on the time and/or space of your algorithms?

5. Submit electronically before midnight of Friday, 09/18/2015 a zip file[1] named `Groupx.zip` where `x` is your group number. This zip file should contain two directories, one named `Code` and one named `Report`. The `Code` directory must contain all of the code implementing the solution for this project. In addition, it must include a `README` describing how to compile and run the code (i.e., details of platform, compiler, etc.). You may be asked to demo your code. The `Report` directory must contain your report (in .pdf format) addressing task 4.

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

4.1 Pearson Correlation Coefficient

To get a measure of how strongly X and Y values are related, we will 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 the X and Y variables. Like other numerical measures, the population correlation coefficient is ρ and the sample correlation coefficient is denoted by r. The formula for the sample correlation coefficient is:

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

where,

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

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

$$\begin{aligned} {s_x}^2 &= \frac{S_{xx}}{n-1}, \text{and} \\ {s_y}^2 &= \frac{S_{yy}}{n-1} \end{aligned}$$

respectively, and the quantity called the *sample covariance* as $s_{xy} = \frac{S_{xy}}{n-1}$.

4.2 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, respectively, in the graph $\langle N(v) \rangle$. The *clustering coefficient* γ_v of v is:

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

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 $\gamma(G)$.

4.3 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 $n = |V|$.

4.4 Metrics for Random Graphs

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.

Acknowledgements

Sincere thanks to Wayne S. Chan who is with the Centre for Earth Observation Science (CEOS) at the University of Manitoba, Canada (see `http://www.umanitoba.ca/ceos/`), for his help in designing this project.

References

[1] A. A. Tsonis, "Is global warming injecting randomness into the climate system?" *Earth Observation Science*, vol. 85, no. 38, pp. 361–364, September 2004.

[2] A. A. Tsonis and P. J. Roebber, "The architecture of the climate network," *Physica A*, vol. 333, pp. 497–504, 2004.

[3] A. 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.

[4] J. P. Onnela, K. Kaski, and J. Kertész, "Clustering and information in correlation based financial networks," *European Physical Journal B*, vol. 38, no. 2, pp. 353–362, March 2004.

[5] 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*, vol. 55, pp. 209–217, 2007.

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

[7] 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*, vol. 86, no. 34, pp. 309–316, August 2005.