An Introduction to and Tutorial for BioGrapher (Workbook for Excel 2010+)

September 2014

 First download the Excel Workbook and associated help files and sample graphs from http://www.bioquest.org/esteem/ and open the workbook in Excel and read this document as a description and tutorial while locating menu items in the workbook.

*******USE AT YOUR OWN RISK******

----->This workbook uses MACROS, whose use MUST be enabled via the standard TOOLS--->MACRO--->SECURITY Excel menu item or through the menu bar/ribbon alert that is displayed when the workbook is first launched. In addition, the user must select "Enable Macros" from the security dialog box that pops up each time the workbook is opened when the "medium" security setting is specified. Finally, note that the custom menu items installed for this workbook will not work after the user enters a value into a cell until the user hits the enter key or clicks on a different blank cell.

This work was originally funded in part by the BioQUEST Consortium (project BEDROCK) at Beloit College (PI Professor John Jungck) and NSF-DUE-CCLI Grant #0633651.

PURPOSE: A simple Excel-based workbook with worksheets as a front end for the AT&T GraphViz Graph Layout software suite. BioGrapher enhances Excel-based tools developed in the Chemistry and Biology Departments at Beloit College to allow for convenient visualization of graphs and graphical connections that are important in systems and computational biology, molecular biology, biochemistry, and genetics.

What is graphical layout and visualization? The GraphViz software library uses complex algorithms to provide different types of visualizations of network graphs. The objective is to present a visual rendering of nodes and links in a fashion that maximizes visual impact, and allows the human viewer to instantly discern patterns, e.g., clustering, and identify key nodes (with large in/out degree) that are essential for effective communication and signaling in the network. Finally, insights into graphical parameters like the diameter and mean free path (average shortest path) can also be intuited. The essential basis of all the algorithms implemented in GraphViz is to minimize edge-crossings and clutter. This can be done in a number of different ways, including representing the graph using a "**circular**" visual layout where the nodes are arranged along the circumference of a circle, a "**spring model**" visual layout based on a simulation that minimizes the energy of the entire system, treating the links as mechanical springs, and a "**radial**" layout where a (large degree) node is positioned at the center, and spokes consisting of links radiate from the center. Finally, for some graphs without closed circuits, a "**tree**" (hierarchical) layout may also be appropriate. For complex graphs, the user should experiment with the different visualization choices, which often provide insights that complement and supplement each other.

DESCRIPTION: The Excel spreadsheet allows for UNDIRECTED (or DIRECTED) graphs and trees to be specified in the standard adjacency matrix notation. Diagonal elements of a square matrix represent nodes; non-zero [usually 1] off-diagonal elements represent connected edges. Since undirected graphs can be represented by a symmetric matrix, only the upper half of the matrix needs to be entered as 1's and 0's for undirected graphs. The rest of the graph can be automatically filled in by invoking 'Complete Undirected' (see below). Once data are entered or imported (see below) into the spreadsheet, visualization of the graph in a number of different layout styles, each of which can provide unique insights into the nature of the graph and connections between nodes, is achieved using a form window that provides access to some very powerful graphical layout routines (GraphViz software) available from AT&T in the public domain under the Common Public License (CPL).

ORGANIZATION of Excel Workbook: The workbook has seven worksheets, whose functions are described below. In addition, it has two special menus (CALCULATE and GRAPHICAL VISUALIZATION) that are added to the Excel menu bar and must be used for all operations on the spreadsheets, INCLUDING LOADING AND SAVING DATA in comma-separated value (CSV) Excel and text-compatible format.

1. *DATA MATRIX*: FIRST ENTER THE ORDER OF THE MATRIX in cell (1, 1) of WORKSHEET "INFO," and then enter the adjacency matrix here as 1's and 0's, starting with cell (1, 1). Immediately after the last column for a particular row (node), it is possible to specify a label for the node (optional, else row numbers are used as labels by default), followed by another optional column where a background COLOR (red, blue, green, yellow, cyan, orange, or magenta) for a node label text box can be specified. Only half the symmetric matrix for undirected graphs needs be entered, and the rest of the graph can be completed by invoking the ‘Complete Matrix (Undirected)' menu item of the CALCULATE menu. I

2. *GRAPH*: This worksheet displays the graph, laid out using the options invoked in the SHOW GRAPH menu item of the GRAPHICAL VISUALIZATION menu. The graph is a vector object and can be displayed and printed on any scale without the "jaggies." It can also be copied and pasted into other worksheets, spreadsheets, or MSWord. Note that you can zoom in and out using CONTROL Z and entering a value!

3. *GRAPH PROPERTY*: This worksheet is reserved for display of numerical and graphical results of the computation of various mathematical properties of the graph using the CALCULATE menu. See the description of the CALCULATE menu for more details. Note that the name of this worksheet will change depending on properties computed. These calculations work only with adjacency matrices.

4. *INFO*: This special worksheet is reserved for saving temporary global data and for use in future versions of BioGrapher.

5. *PIXEL MATRIX*: This feature shows the ones and zeroes of the adjacency matrix as white and black squares respectively, and is a very useful visualization that aids in rearrangement of graphs that are in the Shkurba form, etc.

6. *FOR USER* (1): Can be used be used as a scratch worksheet.

7. *FOR USER* (2): Can be used as a scratch worksheet.

All manipulations of data are carried out using TWO ENHANCED AND SPECIAL MENUs that has been added to the standard Excel menu bar on the Mac. [For PCs use the ADD INs standard menu tab to display these menus.] Note that the menus are visible only when you click on a blank cell and NOT when you have a chart or drawing/shape selected! Here is a brief description of the relevant menus and various menu items:

+++GRAPHICAL VISUALIZATION+++

DRAW GRAPH will draw a graph of the data using a specified layout (method of representing the graph's topology) from the ATT GraphViz suite. The default layout is the circular layout. The spreadsheet remembers the last layout style selected and uses it for subsequent displays. Invoking this menu item displays a form with the various choices, and choices for font sizes for node (vertex) labels. Note that computation of the layout may take some time (even minutes!) depending on the size and complexity (connectivity) of the graph to be displayed, especially for circular layout. The form displayed by invoking DRAW GRAPH also has radio buttons that allow a choice between drawing UNDIRECTED or DIRECTED graphs. This feature allows asymmetrical adjacency matrices must be set up correctly for DIRECTED graphs!

LOAD MATRIX DATA allows for data representing a graph in the adjacency matrix notation described above to be imported into the spreadsheet. The data have usually been SAVED (see below) in CSV format previously after data entry in BioGrapher, although the data could also have been generated by another program or have been manually typed in as a standard file.

SAVE MATRIX DATA saves previously loaded or entered data in a CSV format compatible with BioGrapher and readable by the LOAD DATA menu item.

+++CALCULATE MENU+++

Shortest Path: Computes the shortest path between nodes using a version of Floyd's algorithm, implemented in VBA by Michael H. Cole, (mcole@ie.montana.edu, http://logistica-design.blogspot.com/2006/02/floyds-shortest-path-algorithm.html).

Also computes the DIAMETER AND THE CHARACTERISTIC PATH LENGTH (as defined by Watts.)

The results are displayed in the GRAPH PROPERTY worksheet, whose name changes to SHORTEST PATH.

Clustering Coefficient: Calculates the value for each node as well as the mean value for the graph, as defined by Watts. It displays the results in columns showing the node ID number, neighborhood size (how many other nodes directly connected), and clustering coefficient for each node. A header shows the mean clustering coefficient for the graph.

In addition, graphs showing the neighborhood distribution (number of nodes for a given neighborhood number v/s neighborhood number) and the log-log plot of this distribution (with a fitted trendline), useful for "SMALL WORLD" connectivity tests!

Complete Matrix (Undirected): Automatically completes the lower half of the (symmetric) adjacency matrix if you enter only half the matrix, i.e., a set of connections in one direction.

Flip 1s and 0s: Flips the graph, yielding the complementary graph, which must first be applied to adjacency matrices imported from javaBenzer.

+++SMALL WORLD+++

Beta Graph: Generates adjacency matrix of a small world beta graph model as specified in Page 67 of "Small Worlds" book. The code is based on the sample code from the article "Simulating Small-World Networks" by Mary Lynn Reed.

You can input Number of nodes, Adjacency neighbors, and Rewiring Probability. Number of nodes specifies the total number of nodes in the graph. Beta graph model arranges nodes in a circle, and nodes next to each other are regarded as neighbors. Adjacency neighbors specify how many neighbors each node is connected to. If the number is 4, a node is connected to 2 nearest neighbors on the left and 2, on the right. The input for this field should be even integers. The code goes through each connection in the graph, and with the probability specified by the rewiring probability input, breaks the connection and randomly creates another one.

Barabasi-Albert Model: Generates adjacency matrix of a model for small world network by Barabasi-Albert algorithm based on the paper "Emergence of Scaling in Random Networks" by Barabasi, A.L. and Albert, R. The code is based on the article "Simulating Small-World Networks" by Mary Lynn Reed.

You can input Initial number of nodes, Final number of nodes, and Number of edges added with each node. The model starts with Initial number of nodes specified and grows by one node at each time step to reach the Final number of nodes. With each new node, new edges, which number is specified by Number of edges added with each node, are also added. These connect the new node with random nodes from the graph. Nodes with higher connections have higher probability to receive new connections.

The initial nodes were made to form a one-dimensional lattice with each node having degree 2. This makes it easy to execute the model of nodes with higher degree having higher probability to get new connections. It also avoids the graph being disconnected.

BIBLIOGRAPHY

1. "Small Worlds: The Dynamics of Networks between Order and Randomness." Duncan J. Watts, Princeton University Press (1999).

2. For a description of the DOT graphical layout language and implementation of the AT&T GraphViz* package, see http://www.graphviz.org/Documentation.php.

3. For a mathematical definition of the clustering coefficient, see http://en.wikipedia.org/wiki/Clustering_coefficient. For small world networks, also see http://en.wikipedia.org/wiki/Small-world_network.

4. For a good discussion of shortest path algorithms, see http://www-unix.mcs.anl.gov/dbpp/text/node35.html

5. "Simulating Small World Networks." Mary Lynn Reed. http://www.drdobbs.com/architecture-and-design/simulating-small-world-networks/184405611

6. "Emergence of Scaling in Small World Networks." A.L. Barabasi, and R. Albert. http://www.sciencemag.org/content/286/5439/509.full