

&Degrees. of Separation: Social Network Analysis Using The SAS® System

Shane Hornibrook, Charlotte, NC

ABSTRACT

Social Network Analysis, also known as Link Analysis, is a mathematical and graphical analysis highlighting the linkages between "persons of interest". This analytic approach has immense practical importance in fields such as epidemiology and fraud analysis. Social Network Analysis (SNA) tools provide spider web-like graphs indicating the strength and type of connections between entities. The SAS® System combines the data extraction, manipulation, analytic, and visualization tools needed to distill massive databases into a visual representation of the most unusual set of linkages. This paper presents a brief overview of the SAS tools and methods the author has found useful in changing a sea of data into a graphical Social Network Analysis.

SAS's parallel-processing methods (MP Connect) can greatly speed up elapsed time when extracting linkage data from large Database Management Systems, especially when data are pulled from separate databases. Delivering the reports using SAS/IntrNet® allows for interactive exploration, filtering, and prioritization of the linkage data. The Treeview and Constellation Applets are two of SAS's methods of examining link data. MP Connect, SAS/IntrNet, and SAS's Java Applets are a perfect fit when analysts have a time-sensitive need to analyze and report on Social Networks.

INTRODUCTION

Social Network Analysis is a method of analyzing and presenting data that contains link information, such as "who knows whom," "who calls whom," "who does business with whom": information linking individuals or entities together. In a graphical context, each point or "node" is connected to other nodes on the graph by lines, called "edges" or "links" (**Figure 1**). These node and edge graphs neatly summarize relationships between entities. Data that would otherwise occupy thousands of lines in a database can easily be represented by a graph that can be understood by the eye. Without Social Network Analysis methodologies, the full extent of the interconnections and structure of the network would be lost in the sea of data.

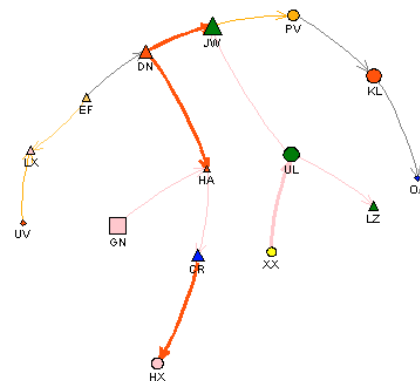


Figure 1: Node and edge diagram

The SAS Institute has created a wide selection of tools for analysis and display of link data to suit varying needs for Social Network Analysis methods. If your data has a hierarchical, or "tree" format, with parent→child relationships, the *TREE* procedure (part of **SAS/STAT®**) provides basic graphical display of data organized in a hierarchy (tree). **SAS/OR®** (Operations Research) contains the mathematical and graphical procedures *NETFLOW* and *NETDRAW*, as well as the interactive application *Network Visualization Workshop*, or "NV Workshop". **SAS/Enterprise Miner™** provides a "Link Analysis" node (**Figure 2**), as part of its Explore category. The Link Analysis node generates statistical summaries related to network analysis, and also provides graphical display of link data. The SAS Institute has also created two Java display and query tools: the **Treeview Applet**, and the **Constellation Applet**. These two applets are generally utilized in conjunction with one of SAS's XML generating macros: %DS2CONST or %TREEVIEW, and are accessible to Base SAS licensees.



Link Analysis

Figure 2: SAS Enterprise Miner™ Link Analysis node

The graphical presentation of link data is not unique to SAS. Other programs provide methods to analyze and browse link data. SAS's advantage is its capability to extract data from distant or complex data sources. Often the data that describes an entity is not held in one table, let alone one database. Comprehensive Social Network Analysis requires data to be drawn from as many relevant sources as possible; SAS is an excellent tool for collating data from disparate data sources and integrating this data, quickly.

Social Network Analysis methods are an excellent investigative technique. In the context of an ongoing fraud or criminal investigation, any turn-around time can have serious repercussions. Often, network analysis requests are performed with what could be described as “shooting into the dark.” An investigator may request a report on a particular entity and want to see their connections to other entities. While this ad hoc reporting is parsimonious, it does not provide the investigator the flexibility to follow a trail of links that may not be immediately apparent. An interactive, on-demand reporting system enables the investigator to query the data and search for “interesting” connections between entities. To this end, the system presented in this paper shows the advantage of delivering Social Network Analyses using SAS/IntrNet.

DATA LAYOUT

Regardless of your chosen display method, the data for a Social Network Analysis must be gathered efficiently. Depending on the methodology used, Social Network link data is usually laid out in one of two formats, though other formats exist. The first is an n-by-n matrix with all entities listed across the top and down the side. This format is used in many matrix-based network analysis methodologies, such as those that use PROC IML. The second common format is “dyadic” data: a three-variable data set listing node one, node two, and a count of connections between node one and two.

From a data management perspective, creating the dyadic data is easier. An advantage to selecting software that utilizes the dyadic data format is data set size. An n-by-n format table can obviously not be created for data sets having millions of nodes, for storage reasons alone. In large network analyses many cells in an n-by-n table would be empty, with links being sparse between row and column nodes. In this context a dyad table occupies less space as a data line only exists when there is a link between two nodes.

Explaining how to create a link dyad is very straightforward: “Just join your data to itself and see what links!” In practice, if your data is of any substantial size, you will need to think about how to do this without filling up disk space, or creating a hopelessly long running query.

The SQL block above shows the most basic method for building a dyad for an undirected graph. In this query, a link is built where one credit card (CC) has been used at the same merchant (merchID). The *additional qualifiers* portion of the SQL statement can be used to filter the data prior to joining.

For data that has a time component, one can build a data set for a directed graph, known as a *digraph*. The *additional qualifiers* statement would then contain time filtering options such as *where a.datetime<b.datetime* or date windowing: *where a.date between b.date-3 and b.date*.

GATHERING THE DATA

MP CONNECT: HARNESSING MULTIPROCESSOR POWER!

Now that we have the basic SQL needed to build a dyad a framework for submitting such SQL is required.

In environments with multi-million row tables containing linkage data for thousands or millions of entities, the key to success is to off-load the processing of these data sets to their systems of origin, when possible. Also advantageous would be to work in an environment with well-indexed tables and an extraordinarily fast relational database.

Often, for any analysis beyond the most straightforward exploratory analysis, linkage data must be pulled from multiple source systems. For timely return of results

```

/** Typical dyad
    construction via SQL */
create table dyad as
select a.cc as node1,
       b.cc as node2,
       count(distinct a.merchID)
          as count
from transactions as a,
     transactions as b
where a.cc^=b.cc
     and a.merchID =b.merchID
     and additional qualifiers
group by a.cc, b.cc;

```

```

/** MP Connect remote submit */
rsubmit process= teradata wait=no;
proc sql;
  connect to teradata (&teradb.);
  create table links_cc as
  select ...
  ;
quit;
endrsubmit;

rsubmit process=db2 wait=no;
proc sql;
  connect to db2 (&db2db.);
  create table links_doc as
  select ...
  ;
quit;
endrsubmit;
waitfor _all_ db2 teradata;
libname terawork slibref=work
          server=teradata;
libname db2work slibref=work
          server=db2;

data links_all;
  merge db2work.links_doc
        terawork.links_cc;
  by node1 node2;
run;

```

we can convert this requirement into an advantage. We can place the data processing burden on these source databases using remote submits and MP Connect. MP Connect is part of SAS/CONNECT® and is SAS's solution to spawning multiple processes in a multiprocessor environment. With MP Connect methods we can start several local or remote SAS sessions and drive our data extraction through these sessions. The second SQL on page 2 contains an example of two remote submit sessions that, if executed on SMP hardware, would execute in parallel. By executing in parallel, the elapsed time for the return of these two queries is only the length of time for the longest query execution, not the sum of the execution time of all queries.

NETWORK VISUALIZATION AND ANALYSIS TOOLS

SAS has many useful tools for displaying linkage structure in your data. SAS/STAT® cluster analysis tools such as the *TREE* procedure provide adequate display of data with a hierarchical structure. However, with a great number of nodes, the tree can become difficult to interpret. The following is a sample of some of the other techniques available to the Social Network Analyst.

ENTERPRISE MINER™: LINK ANALYSIS NODE

Link Analysis is a natural component of the broader suite of “explore” tools found in SAS Enterprise Miner™. The ready access to network analysis metrics and ease of graphing make Link Analysis available to non-programming analysts.

After running the link analysis node, a variety of standard graphs can be displayed and saved. Graphs include histograms, link Chi-squared distribution, first- and second- order centrality measures including unweighted and weighted measures.

Of course, the Link Analysis node allows analysts the ability to specify differing node and link properties such as color and size. Nodes may also be repositioned for more attractive and informative graphs.

Layouts include Circle, Grid, Multidimensional

```

/** An example of portions
    of the XML
    generated by
    SAS/Enterprise Miner™ **/
<?xml version="1.0"
encoding="Windows-1252"?>
<NodeLinkDiagram>
<Nodes>
  <Node>
    <ID>N1</ID>
    <X>250</X>
    <Y>309</Y>
    <Color>0x274CB5</Color>
    <label> </label>
    <tip>COLR=0.227</tip>
    <size>11</size>
  </Node>
  ...
<Links>
  <Link>
    <fromNode>N8</fromNode>
    <toNode>N10</toNode>
    <width>1</width>
    <color>0x55AA55</color>
    <value>9</value>
  </Link>
  ...

```

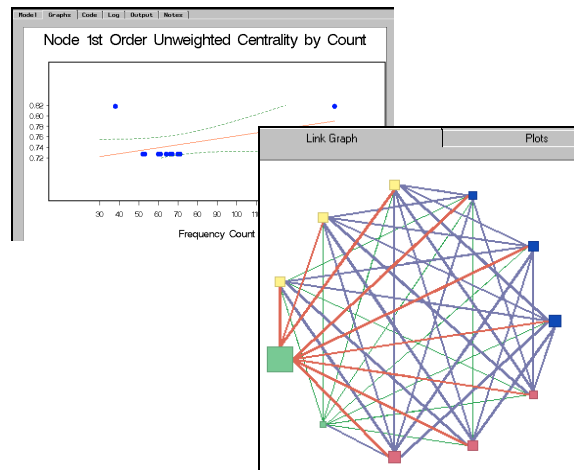


Figure 3: Screenshots of two of the many pieces of output created by SAS Enterprise Miner™ -- Node centrality measures (upper) and circular layout (lower)

Scaling (MDS), Parallel Axis, and Tree. The MDS layout utilizes the MDS procedure, the TREE layout uses the NETFLOW procedure. A layout modifier named “Swap”, energizes any of the previous layouts by essentially “jiggling” the graph, moving nodes with higher link weights closer together. The “Swap” layout when run iteratively, moves related nodes closer to each other, usually beautifying the graph each time

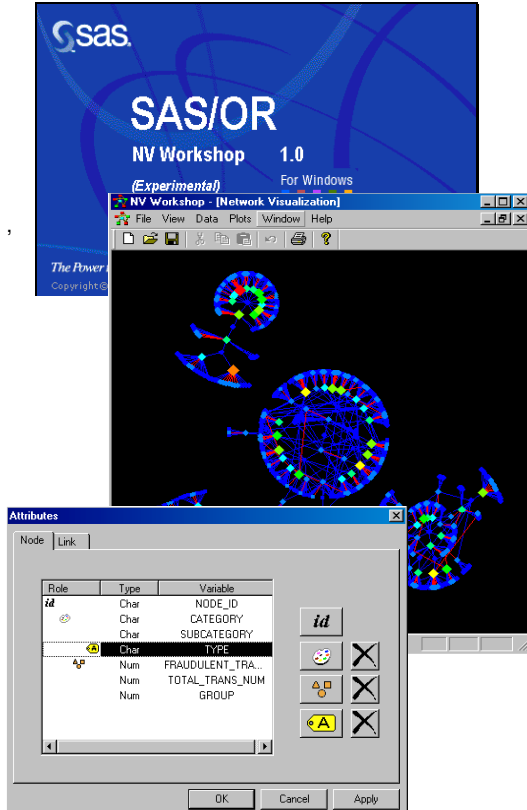
Unweighted- and weighted- **centrality** measures included are first-order undirected centrality and second-order undirected centrality. The centrality measures can be described as a basic metric counting the number of nodes to which each node connects (first order centrality), and to which those nodes connect (second order centrality) relative to the total number of nodes. **Clustering** of nodes can be accomplished by using kernel density or nearest neighbor algorithms. **Scoring** can be based on links or nodes.

The node and link data can be saved in XML format (left), or as a SAS data set (Netdata, or Matrix). The XML format is useful as a format for conversion into other packages.

The entire Enterprise Miner report can be saved to a catalog, and revisited as needed.

SAS/OR: NETWORK VISUALIZATION WORKSHOP

NV Workshop (*Experimental*) is a standalone point-n-click tool available to Microsoft Windows SAS/OR users (**Figure 4**). The application allows browsing of link data, coloring nodes, as well as presenting data using four different layout types: Hierarchical, Circular, Hexagonal, and Fixed Layout. NV Workshop readily reads and writes data as regular SAS data sets.



In addition to network visualizations, NV Workshop also allows you to create histograms, boxplots, and scatterplots.

Node ID's, labels, colors, and shapes are readily specified (**Figure 4**). You can select link variables as well as the link color and label.

This application is classed as *Experimental*, (**Figure 4**) and you should use it accordingly: as a data exploration tool.

Caveats: Batch processing of data is not possible, to the author's knowledge. Images can only be saved by screen captures.

SAS/OR: NETDRAW, NETFLOW

While the visualizations created in NV Workshop are extremely useful, SAS/OR users have other tools available. If you require analytical tools to determine flow through a network, such as shortest path, SAS'S PROC NETFLOW and PROC NETDRAW are the solution.

PROC NETFLOW solves useful network analysis questions such as the Shortest Flow problem, and Minimum Cost problem.

A natural display mechanism for NETFLOW output is the NETDRAW procedure. The two procedures are often used together, but NETDRAW can create graphs from any appropriately formatted data.

Figure 4: NV Workshop, Part of SAS/OR® -- Easily specify ID, color, shape and label

The general layouts used in NETDRAW are orthogonal layouts: edges are drawn along vertical or horizontal axis, and nodes are aligned.

PROC NETDRAW can also be utilized with the Annotate facility to place annotations, including images. Interactive mode allows you to adjust node locations using the MOVE command.

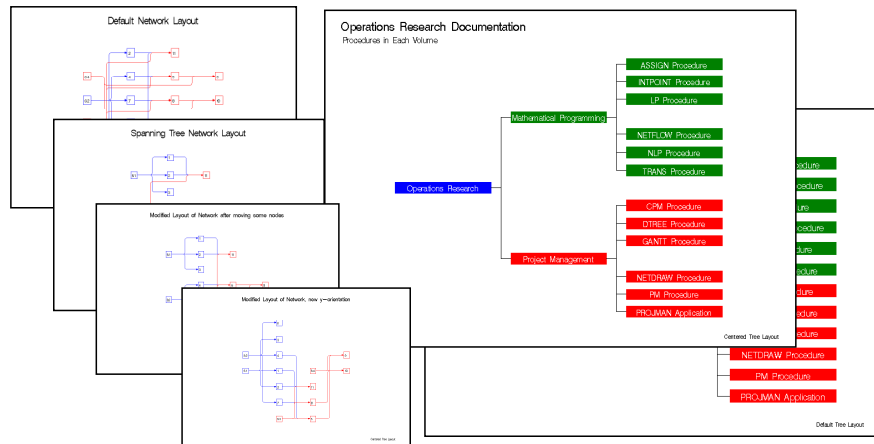


Figure 5: NETDRAW Layouts -- SAS (2004, 2006)

If your data has natural groupings and you want to separate your nodes by these groupings, NETDRAW can easily position nodes into differing areas of the graph, using ALIGN (horizontal) and ZONE (vertical). This is useful for general classification of nodes.

DO IT YOURSELF: THE ANNOTATE FACILITY

Just as the annotate facility proves itself useful for NETDRAW graph annotation, it can also be used as a stand-alone graph drawing facility (Fairfield-Carter, 2004)

Social network analysis layout algorithms are freely available, and, for an enterprising programmer the ability to “do it yourself” is appealing. Using Social Network Analysis layout algorithms, such as force directed placement algorithms, optimized X and Y locations for each node can be determined. The Annotate facility allows precise placement of graphical elements, such as primitive shapes and lines. The Annotate facility also allows placement of images, opening up the possibility of going beyond the limited “Circle, Square or Diamond” choice for node shapes using other methodologies.

JAVA APPLETS: %TREEVIEW AND THE TREEVIEW APPLET

The %TREEVIEW macro and Treeview Applet allow presentation of data sets that have a tree form; that is, there is only one path from any one node to another.

In the context of Social Network Analysis, data that has a one-way relationship, with no back routing, could be hierarchical data such as ownership, or data with a time component. An example is the ubiquitous organizational chart (Figure 6).

Using a fish-eye layout with the tree view applet is one of the more pleasing properties of this graph (Figure 7). In a dense graph the fish-eye layout allows detailed examination of one portion of the graph, such as one branch of the tree. The amount of fish-eye is controllable using keyboard shortcuts (control key), clicking on the diagram and moving the mouse up or down. As an interesting branch of the tree is discovered, you can pan down the tree to view the details.

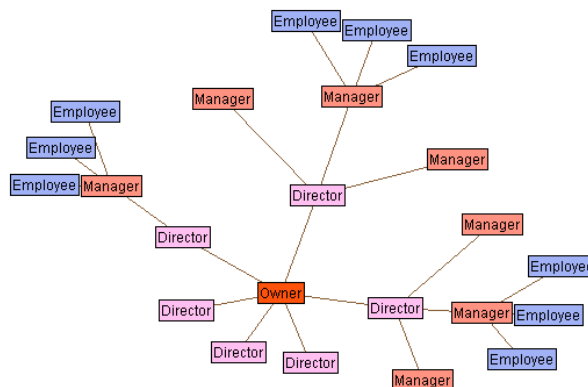


Figure 6: Treeview Applet

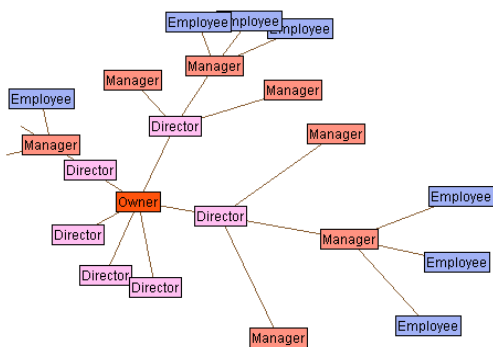


Figure 7: Treeview Applet with “fish-eye” effect

JAVA APPLETS: %DS2CONST AND THE CONSTELLATION APPLET

While uni-directional or hierarchical graphs are useful, much of the most interesting Social Network Analysis data contains bi-directional linkages, not just hierarchical. “Entity A” may have a relationship with “Entity B” and “Entity B’s” relationship with “Entity A” may or may not be the same strength. For example, “Entity A” may phone “Entity B” twice as much as “Entity B” phones “Entity A”. This bi-directional but unequal strength relationship is important to note, but not possible to display in typical hierarchical graphs such as %TREEVIEW, due to their nature.

The %DS2CONST macro and Constellation applet are an excellent tool for representing bi-directional relationships. The most interesting graph type available in the Constellation applet is the Associative graph. Associative graphs display nodes and edges based on their weighted values; node size and edge width can represent the relative size of the node and edge values. The graph has curved lines, ending in an arrow, pointing from the source node to the sink node. If relationships are reciprocated, there will be an additional arrow pointing back to the first node (Figure 8).

If you are interested in only the links between one node and the rest of the network, the right-click menu provides the option to select all links, only links into your selection, only links out of your selection, and links into or out of your selection (Figure 12, Figure 13).

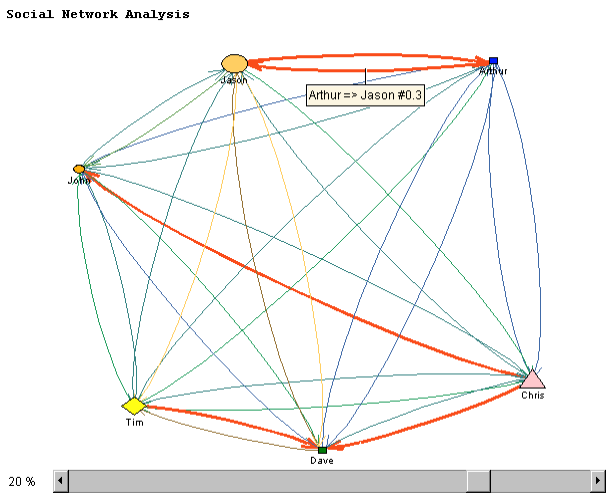


Figure 8: Weighted links and nodes

JAVA APPLET FUNCTIONALITY

The Constellation Applet and Treeview Applet share a great deal of functionality. You can select single or multiple nodes and highlight connections, query data (Figure 10) use right click functionality for resetting the view (Figure 9, Figure 11), opening URLs, zooming, and panning.

INTERACTIVE SOCIAL NETWORK ANALYSIS USING SAS/INTRNET AND JAVA APPLETS

SAS/IntrNet is not required to use either the Constellation or Treeview Applet. However, integrating the Constellation Applet and/or the Treeview Applet with SAS/IntrNet opens up a virtually unlimited analytical world for your users that would not otherwise be accessible. For example: clicking on nodes in the Java Applet can be integrated with custom-built JavaScripts and additional web queries. JavaScript methods can be passed collated lists of selected nodes (e.g. node identifiers are passed in a delimited string: 4;5;6;7;8). URL's can open new windows, with further SAS/IntrNet queries or detail data for the node selected.

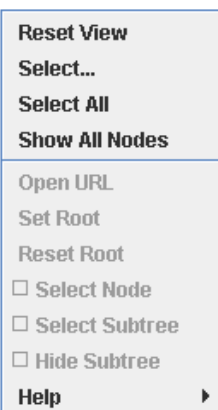


Figure 11: Treeview Applet right-click menu

Integrating a Constellation Applet or Treeview Applet "as is" in your own web page requires minor effort. To specify that the %DS2CONST or %TREEVIEW macros generate the appropriate header for the Application Dispatcher you can specify runmode=S for *Server*. If you would like to save a copy of the results returned you can specify runmode=B, for *Batch*. It is this latter method that can be used to provide an audit trail and replicability required in fields such as fraud analysis and law enforcement.

With batch output you have the option of saving the output to a specified location with a unique identifier. Subsequently you can then embed and serve the graph up to your users as part of a larger online report.

The Constellation Applet also provides a slider bar that allows you to interactively determine the minimal value of the edge weights to display (Figure 8). Edges below the threshold are hidden.

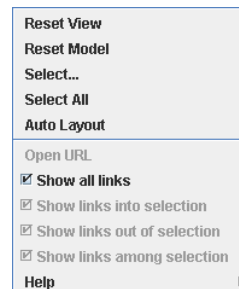
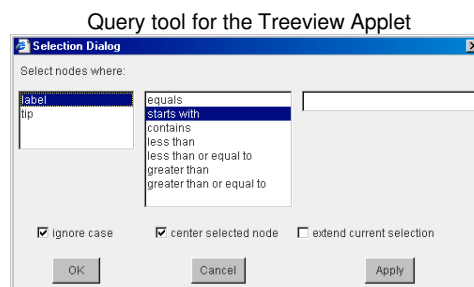
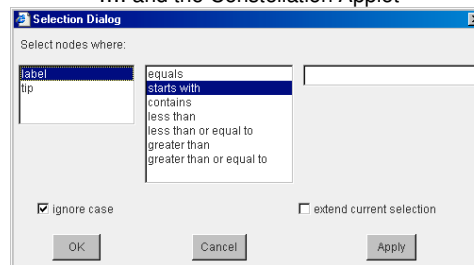


Figure 9: Constellation Applet right-click menu



Query tool for the Treeview Applet



... the notable difference being the ability to center a node in the former query tool

Figure 10: Java Applets Query Tools

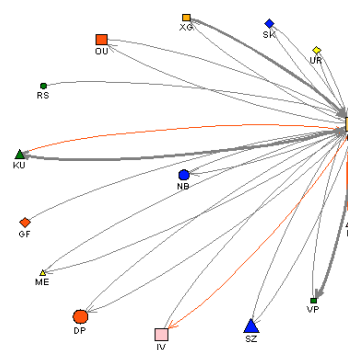


Figure 12: links among selection, Constellation Applet

Developing a SAS/IntrNet report that integrates the Constellation and/or Treeview Applets removes significant workload from data analysts. Primary investigators are able to query the data sources directly and retrieve Social Network Analyses. If the analysis is not what they are interested in they can immediately re-query for more specific data. If a single node or a pattern of node to node linkages is "interesting" the investigator can select one or more nodes for further reporting. If the thresholds are not set correctly, the investigator can re-query with adjusted thresholds.

SAS/INTRNET QUERIES

From the initial SAS/IntrNet query page (Figure 14), users can select the nodes of interest, by whichever properties they choose. This initial query page can default to a generic Social Network Analysis graph. This generic graph should give a census of nearest neighbor nodes. The default query behavior ensures that investigators can immediately see the connections between the target and other nodes. For example, businesses they own or manage and other businesses to which their co-owners and managers are connected.

There are many user-modifiable properties set by the %DS2CONST macros. However, the most important variables for the display of networks are the edge thickness, edge color, node colors, node sizes, and node shapes. By creating simple query pages that provide options for setting these macro variables users can specify the entities they are searching for as well as set the graph properties (Figure 15). When the request is submitted, SAS/IntrNet, utilizing MP Connect, extracts and manipulates the data requested and serves up another page with the embedded Java Applet.

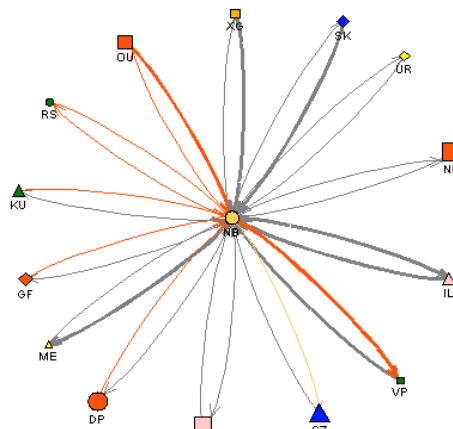


Figure 13: Links among selection, Constellation Applet

Social Network Analysis Query

Name: ?

SSN: 000-00-0000 ?

Address: Street: City: State: ZIP:

ID Number: ? default graph modify graph

Select ID Network ID default graph modify graph

Previous Projects Select Project default graph modify graph

Figure 14: SAS/IntrNet Initial Query Page

CONSTELLATION MACRO: THE SKY'S THE LIMIT

The %DS2CONST macro and Constellation Applet's are able to hyperlink from a selected node or pass node ID's as delimited lists to a JavaScript. Practically, this means that these IDs can be the basis for additional reports, queries, and ultimately storage in a database of "interesting" linkages.

Social Network Analysis Report Screen

Select Node 1: All Nodes Select Node 2: All Nodes

Select Linkage Variable: Analysis Select Values: All Values

From Date: 01/01/1991 To Date: 09/01/2006 Select Threshold: 99th Percentile

Display Properties

Node Color: Select Variable Node Size: Select Variable Node Shape: Select Variable

Edge Color: Select Variable Edge Thickness: Select Variable

Figure 15: SAS/IntrNet Sub-Query Page

Integration of the Constellation macro into a suite of on-line investigative tools is a bridge to an unlimited number of options. No longer do investigators have to wait for the outcome of a linkage request: They can query the data and receive easy to interpret graphical output. The query, SAS log, and output data can be stored, and therefore an audit trail can automatically be built. Clicking on interesting nodes within the graph can spawn additional drill-down reports of any type, and the nodes of interest can be stored in an investigator's online file.

CONCLUSION

SAS provides a wide variety of Social Network Analysis capable tools. SAS/IntrNet based reports integrating the Constellation Applet and utilizing MP Connect are just one of SAS's many powerful Social Network Analysis tools.

REFERENCES

Fairfield-Carter, Brian (2004), "A Stand-Alone SAS® Annotate System for Figure Generation"
<http://www2.sas.com/proceedings/sugi29/061-29.pdf> , SAS® Users Group International Proceedings SUGI29

SAS® Institute (2004), "Example 5.15: Organizational Charts with PROC NETDRAW" SAS Help and Documentation, SAS Institute Inc, Cary, NC

SAS® Institute (2006), "Alternate Network Arc Routing program" (NETWORK.SAS):
<http://support.sas.com/rnd/app/examples/exNDR.sas.html>

RECOMMENDED READING

The SAS Institute, "Macro Arguments for the DS2CONST, DS2TREE, DS2CSF, and META2HTM Macros"
<http://support.sas.com/91doc/getDoc/graphref.hlp/a002606467.htm> , SAS Institute Inc, Cary, NC

The SAS Institute, "Sample 1168: Constellation applet example with DS2CONST macro"
<http://support.sas.com/ctx/samples/index.jsp?sid=1168&tab=code> , SAS Institute Inc, Cary, NC

The SAS Institute, "How to Use This Constellation"
http://support.sas.com/rnd/datavisualization/webgraphs/v9_1/en/constchartapplet/mainmenu.htm , SAS Institute Inc, Cary, NC

Scott, 2000, "Social Network Analysis: A Handbook" Sage Publications Inc.

Sparrow, MK, 2000, "License to Steal: How Fraud Bleeds America's Health Care System" Westview Press. *Commentary: Contains a description of a Social Network Analysis performed in Florida in 1993, related to Medicare Fraud detection (Page 243). In addition, a book review has been published in the FBI Law Enforcement Bulletin - January 2002 (http://www.fbi.gov/publications/leb/2002/jan02leb.pdf).*

Wasserman and Faust, 1994, "Social Network Analysis: Methods and Applications" Cambridge University Press.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Shane Hornibrook
Charlotte, NC
Phone: (407) 744-4387
E-mail: nesug_paper @ ShaneHornibrook.com
Web: <http://www.ShaneHornibrook.com/nesug2007/>

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are registered trademarks or trademarks of their respective companies.