Paper 068-31

# SAS/OR®: Making Sense of Network Data with NV Workshop

Ed Hughes and Phil Meanor, SAS Institute Inc., Cary, NC

## ABSTRACT

Network data, structured into nodes and links joining the nodes, is being used to describe more and more enterprise systems and their associated planning problems. Examples include supply chains, communications networks, Web sites, database designs, and credit card associations. Interest in this area has been so intense that several promising new fields of study have arisen, among them "The Science of Networks" and "Small World Networks."

Just as network data is being used more frequently and studied more intensely, the size of typical network data is increasing also. The growing size of network data creates a problem: How can you identify patterns and relationships and extract other useful information from data that can feature thousands of nodes and even more links? One answer is visual analysis. SAS/OR's Network Visualization Workshop (NV Workshop) is designed to enable you to visualize and investigate large networks. Data tables, statistical graphs, and network plots can be used separately or can be linked together as you explore your network data and uncover hidden information.

## INTRODUCTION

*"No man is an Island, entire of itself; every man is a piece of the Continent, a part of the main."* —John Donne

### THE IMPORTANCE OF NETWORKS

Networks—biological, social, technological, business, and otherwise—are a fact of life in an increasingly interconnected world. In a very real sense networks have always been with us, since there have always been food chains in nature, alliances and partnerships date back to the dawn of civilization, and supply chains have existed for centuries if not millennia.

Nevertheless, networks today are far more prominent in everyday life, in business, and in academic research than they were 20, 10, or even 5 years ago. In all but the most private types of decision-making today, networks—manifested as the World Wide Web, supply chains, ecological webs, social networks, or any number of other forms—play a leading or supporting role.

In part this is due simply to greater emphasis on and awareness of preexisting links between organisms, individuals, organizations, and businesses. However, the rise of networks is also due to technological advances and the changes that they have brought about. Better and less expensive communication technology (Internet, cellular and satellite telephony, video conferencing, etc.) engenders interaction between individuals and organizations that would otherwise have interacted only occasionally, if ever. Faster and more efficient transportation removes barriers to in-person interaction imposed by geographic distances. Increased credit-card purchasing, fueled by advances in retail point-of-sale technology, creates networks of merchants and customers, linked by transactions.

In all of these cases the growing importance of networks is also attributable to continuing advances in data technology, enabling us to gather more data and to store and use it more easily and more economically than ever before. Thus, even as we become more deeply interconnected we also have access to more and more data describing the details and the extent of our interconnectedness. The challenge is one of utilizing network data in the most effective manner.

### INCLUDED IN THIS PAPER

In this paper we begin by describing the nature and structure of network data, with a focus on the data formats with which NV Workshop works. We discuss the types of network visualization and network data exploration and analysis supported by NV Workshop, placing special emphasis on how these various visualization techniques, explorations methods, and analyses can be coupled together in order to deliver much greater insights into the information, patterns, and relationships hidden within network data.

We conclude with two examples of the application of NV Workshop to analyze real-world network data. In each case NV Workshop enables us to analyze a large network data set and to uncover links and relationships that are only brought to light readily by visual analysis.

## NV WORKSHOP

NV Workshop is an interactive, graphics-oriented application for visualizing and investigating network or graph data. The application is Microsoft Windows-based and implements network visualization techniques that have proven useful for extracting hidden information and patterns in network data, particularly for large networks. Using a combination of data tables, statistical graphs, and network plots, you can selectively view and filter your network data to help uncover relationships that might be concealed by the volume of the data.

The following sections provide details on the format of the data expected by NV Workshop as well as an overview of the functionality provided in the application. We conclude with examples illustrating how you can use NV Workshop to investigate actual network data.

### NETWORK DATA: SOURCES AND FORMAT

A network can be roughly defined as an interconnected group or system; therefore, *network data* is the information that describes or defines such a system. This information typically consists of two parts—details on the items being connected and information on the connections between these items. Different disciplines have different names for these components. For instance, in graph theory the components are sometimes referred to as vertices and edges, while in other contexts they are called nodes and arcs. For the purposes of this application we call an item at the end of a connection a *node* and the connection between two nodes a *link*.

Many real-world problems can be represented by using a collection of nodes and links. Common examples include supply chains, Web sites, database schema, communication networks, and software module dependencies. For a supply chain the nodes might represent manufacturing plants, warehouses, and customer locations while the links might represent the flow of goods or products between the locations. For a communications network the nodes could be switches, routers, and other hardware devices with attributes such as capacity, device type, traffic volume, number of dropped packets, etc. The links could represent transmission facilities or media connecting the nodes, and the data associated with the links might be failure rates, error rates, traffic volume, etc.

To create a network plot NV Workshop requires two data sets: a link data set and a node data set. Together these data sets constitute the network data. Each row in the link data set describes one link between nodes in the network and must contain at least two variables (or columns) to identify the link. The values of these variables must be node identifiers. The FROM variable lists the node at which the link originates and the TO variable lists the node at which the link terminates. Thus, a link for which the corresponding row in the link data set lists FROM="A" and TO="B" begins at node "A" and ends at node "B." Other variables in the link data set can be used to store attributes related to the specific links. You use the **Data: Edit Attributes** dialog box from the NV Workshop pull-down menus to identify the FROM and TO variables in your link data set. Figure 1 shows part of a sample link data set.

bdlinks.sas7bdat

| | INDIVIDUAL | CORPORATION | BOARD_CHAIRMAN | CEO | NUMBER_OF_BOARDS | BOARD_SIZE |
|---|---|---|---|---|---|---|
| | Nom | Nom | Int | Int | Int | Int |
| 1 | Breyer | Walmart | 0 | 0 | 1 | 16 |
| 2 | Chambers | Walmart | 0 | 0 | 2 | 16 |
| 3 | Coughlin | Walmart | 0 | 0 | 1 | 16 |
| 4 | Friedman_S | Walmart | 0 | 0 | 2 | 16 |
| 5 | Gault | Walmart | 0 | 0 | 1 | 16 |
| 6 | Glass | Walmart | 0 | 0 | 1 | 16 |
| 7 | Hernandez | Walmart | 0 | 0 | 1 | 16 |
| 8 | Lepore | Walmart | 0 | 0 | 1 | 16 |
| 9 | Reason | Walmart | 0 | 0 | 1 | 16 |
| 10 | Sanders | Walmart | 0 | 0 | 1 | 16 |
| 11 | Scott_H | Walmart | 0 | 1 | 1 | 16 |
| 12 | Shewmaker | Walmart | 0 | 0 | 1 | 16 |
| 13 | Soderquist | Walmart | 0 | 0 | 1 | 16 |
| 14 | Villarreal | Walmart | 0 | 0 | 1 | 16 |
| 15 | Walton_J | Walmart | 0 | 0 | 1 | 16 |
| 16 | Walton_S | Walmart | 1 | 0 | 1 | 16 |
| 17 | Boskin | ExxonMobil | 0 | 0 | 1 | 13 |
| 18 | Dahan | ExxonMobil | 0 | 0 | 1 | 13 |
| 19 | Esrey | ExxonMobil | 0 | 0 | 3 | 13 |

Figure 1: Excerpt from a Link Data Set

In this link data set the variable labeled INDIVIDUAL represents the FROM node and the variable labeled CORPORATION contains the TO node. Row 1 designates a link between node Breyer and node Walmart. (Example 2 later in this paper explores this network in greater detail.)

A node data set, defining the nodes in the network, is also required. In this data set each row represents one node in the network. While NV Workshop provides the capability to automatically create a minimal node data set from the link data set you provide, you will typically want to provide your own node data set to the application. The node data set must contain at least one variable, the NODE ID variable, collectively containing all of the node identifiers for the network. Note that these must be the same node identifiers used as values of the FROM and TO variables in the link data set. The node data set usually also includes many variables containing attribute information for the nodes in the network. Figure 10 contains a portion of the node data set associated with the link data set from Figure 1. The NODE ID variable is labeled NAME in this data set; you declare the NODE ID variable using the **Data: Edit Attributes** dialog box from the NV Workshop pull-down menus.

bdnodes.sas7bdat

| | CATEGORY | NAME | FULL_NAME | TYPE | NUMBER_OF_BOARDS | REVENUES |
|---|---|---|---|---|---|---|
| | Nom | Nom | Nom | om | Int | Int |
| 1071 | Individual | Coan | Gaylord_Coan | I | 1 | 15000 |
| 1072 | Individual | De Boon | Herman_De Boon | I | 1 | 15000 |
| 1073 | Individual | Mimran | David_Mimran | I | 1 | 15000 |
| 1074 | Individual | Mulroney_M | M_Mulroney | I | 1 | 15000 |
| 1075 | Individual | Strauss | Robert_Strauss | I | 1 | 15000 |
| 1076 | Individual | Vanier | John_Vanier | I | 1 | 15000 |
| 1077 | Individual | Webb_O | O_Webb | I | 1 | 15000 |
| 1078 | Individual | Young_A | Andrew_Young | I | 1 | 15000 |
| 1079 | Retail Trade | Walmart | Wal-Mart_Stores | C | 16 | 219812 |
| 1080 | Energy | ExxonMobil | Exxon_Mobil | C | 13 | 191581 |
| 1081 | Manufacturing | GM | General_Motors | C | 14 | 177260 |
| 1082 | Manufacturing | Ford | Ford_Motor | C | 14 | 162412 |
| 1083 | Energy | Enron | Enron | C | 17 | 138718 |
| 1084 | Manufacturing | GE | General_Electric | C | 16 | 125913 |
| 1085 | Finance And Insurance | Citigroup | Citigroup | C | 17 | 112022 |
| 1086 | Energy | ChevronTexaco | Chevron_Texaco | C | 15 | 99699 |
| 1087 | Technology | IBM | International_Business_Machines | C | 14 | 85866 |

Figure 2: Excerpt from a Node Data Set

**OVERVIEW OF NV WORKSHOP FUNCTIONALITY**

Using NV Workshop is straightforward. A typical NV Workshop session might consist of the following steps:

1. Load your network data into NV Workshop.
2. Use the Attributes dialog box to identify the FROM and TO variables in the link data set and the NODE variable in the node data set.
3. Create a combination of statistical and network plots to investigate the data.

**LOADING DATA INTO A DATA TABLE**

Before any analysis can begin you must load your data into the application. As noted earlier, NV Workshop requires two pieces of input information in the form of SAS data sets. You load data by selecting either **File: Open** or **File: Recent Data Sets** from the NV Workshop pull-down menus. Selecting **File: Open** opens the standard Microsoft **Open File** dialog box, with a few additional fields; see Figure 3.

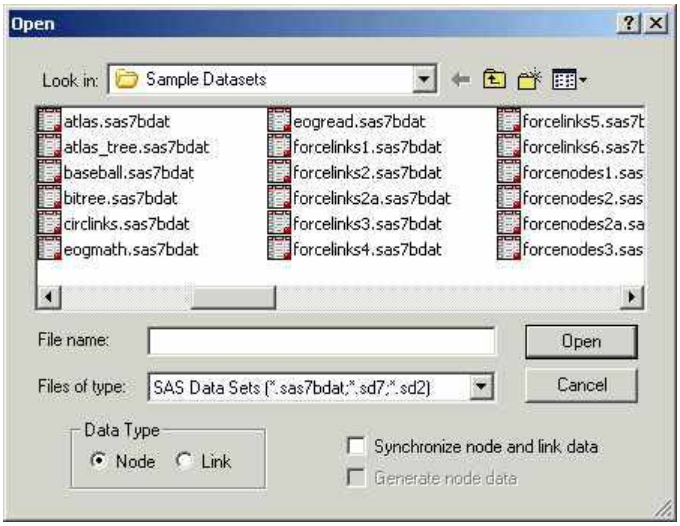


Figure 3: NV Workshop's Open File Dialog Box

In addition to the file name selection fields, the **Open File** dialog box supplies a radio button for you to identify the type of data (node or link) being loaded. When loading link data, you can choose to have the node data set created by the application (from the link data) by selecting the **Generate node data** check box in this dialog box or by selecting **Data: Create Node Data** from the NV Workshop pull-down menus.

The last eight link and node data sets read into NV Workshop are saved in the **Recent Data Sets** lists for quick future access. The names of these data sets are displayed in the pull-down menu under **File: Recent Data Sets**, and selecting one of these items loads the associated data set.

Note that you can have at most one link data set and one node data set loaded in NV Workshop at any time.

**DATA TABLE**

Whenever a data set is loaded into NV Workshop a data table displaying the data set values is automatically created. Figures 1 and 2 above are examples of NV Workshop data tables. The data table operates somewhat like a typical spreadsheet. You can edit data and sort, insert, and delete columns.

The first column in the data table contains the observation label, which is used to identify observation markers in plots. The second column contains icons indicating whether the observation is used in plots or analyses. A marker icon appears if the observation appears in any plots and indicates the marker color and shape associated with the observation. A $\chi^2$ icon appears if the observation is used in any analyses or calculations.

All views of the data in NV Workshop, that is, all data tables and plots, are linked through an underlying data model. Therefore any changes or selections made in a data table will also be reflected in any other views of that same data.

**STATISTICAL PLOTS**

After you have loaded link and/or node data into NV Workshop, you can begin to analyze or investigate the data with a variety of statistical plots. These plots include histograms, bar charts, box plots, and scatter plots, and serve two distinct roles in NV Workshop. First, they enable exploratory data analyses on either the link or the node data. Distributional patterns or correlations in the data can often be detected using these plots.

Furthermore, the plots can provide "data filters" for network plots. Since all views (tables or plots) of a data set are linked, observation selections in one plot or view are reflected in all plots using that data set. Using the statistical graphs, you can selectively filter the observations displayed in the network plots to uncover important relationships between nodes. (More detailed information on the usefulness of filtering observations appears later in this paper.)

A statistical plot is created using the **Plots** pull-down menu on the main NV Workshop window, which includes a menu entry for each plot type. Selecting one of these plot menu entries opens a dialog box in which you can select the data variables to be used to create the plot. Once the plot is created, you can use the left mouse button to select individual observations (or groups of observations), or use the right mouse button to invoke other features of the plot. A sample box plot is show in Figure 4 below.

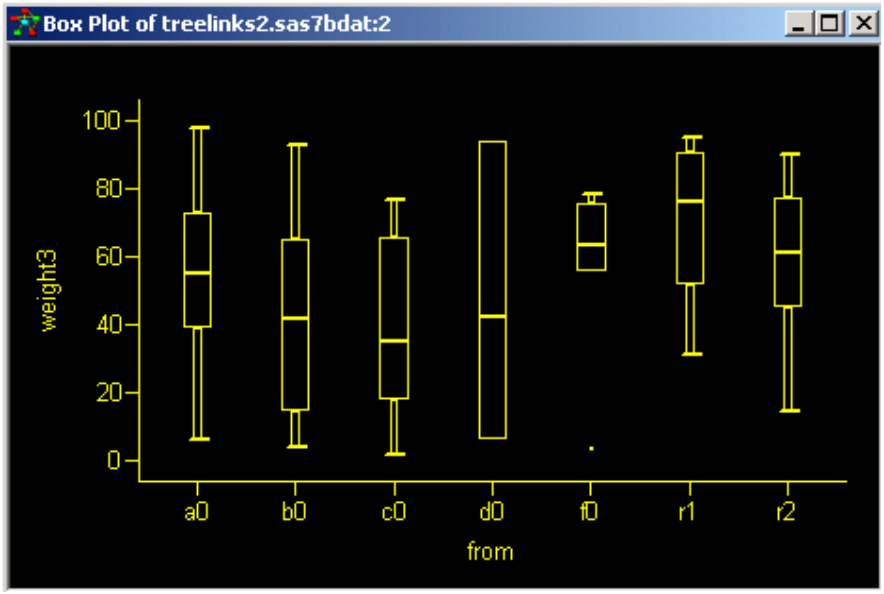


Figure 4: Box Plot

**NETWORK PLOTS**

While statistical plots use data from either the link or the node data set, a network plot uses information from both the link and node data sets to generate a graphical depiction of the network. As with the statistical plots, a network plot is created through menu entries in the **Plots** pull-down menu under the **Visualize Network** heading.

The nodes in a network plot can be arranged in a variety of patterns, and NV Workshop offers five node layout choices for the nodes in a network plot:

- Circular: nodes are arranged in a circle and links are shown within the circle
- Hierarchical: a treelike plot, placing nodes with more connecting arcs closer to the center
- Hexagonal: nodes are evenly distributed across one or more hexagons
- Multi-level force: uses graph partitioning and force-directed heuristics to position nodes
- Fixed position: nodes are arranged according to X and Y positioning data that you supply

Depending on the nature of the data, your choice of layout algorithm can greatly affect the insight you can glean from the visualization of your network. Since you are typically not aware a priori of any underlying structure in your data, it is often useful to try multiple layout algorithms to determine which are the most useful for the network being investigated. An example of a hierarchical network plot is presented in Figure 5.

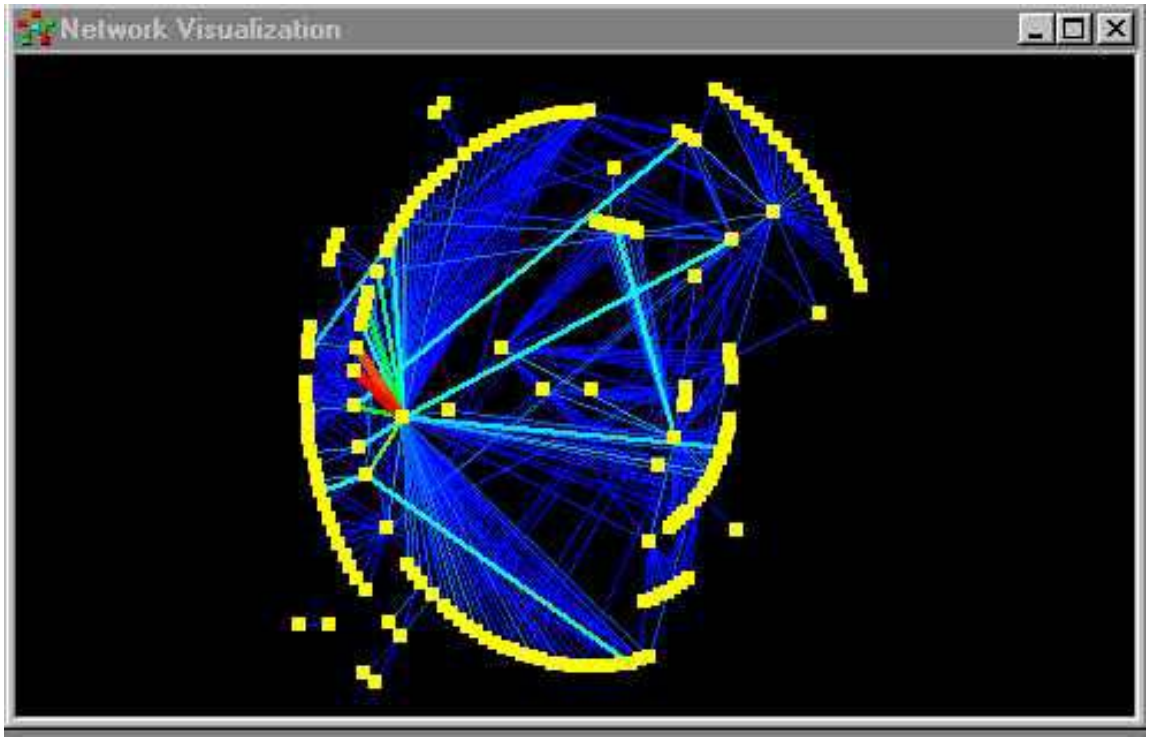


Figure 5: Hierarchical Network Plot

Once the network plot is created, you can use the left mouse button to select individual observations (or groups of observations), or use the right mouse button to invoke other features of the plot. For example, a number of "tools" may be invoked to change the function of the cursor within the network plot. Choices include:

- Selection Tool: selecting one or more nodes displays the nodes and their connected arcs, hiding all others
- Highlight Tool: similar to the selection tool, but "grays out" the nonselected nodes and arcs
- Label Tool: displays the label information for the selected node(s)
- Pan Tool: moves the entire plot in the direction of the cursor; useful for exploring very large plots
- Zoom Tool: produces an enlarged view of the selected area
- Lens Tool: magnifies the region of the plot near the cursor (see below)

"Lens Tool" is one of the more interesting features available on the network plot. Essentially the Lens Tool applies a transformation to the location of the nodes in the network plot relative to the location of the cursor. Move the cursor and the lens moves. Network plots tend to be very dense and the use of a lens enables you to focus in on the detail in a particular region of the plot. There are multiple lenses available in a network plot (accessible via the "Graph Properties" selection from the network plot pop-up menu) and each has various options. Figure 6 shows the use of a radial lens applied to an entire network plot.

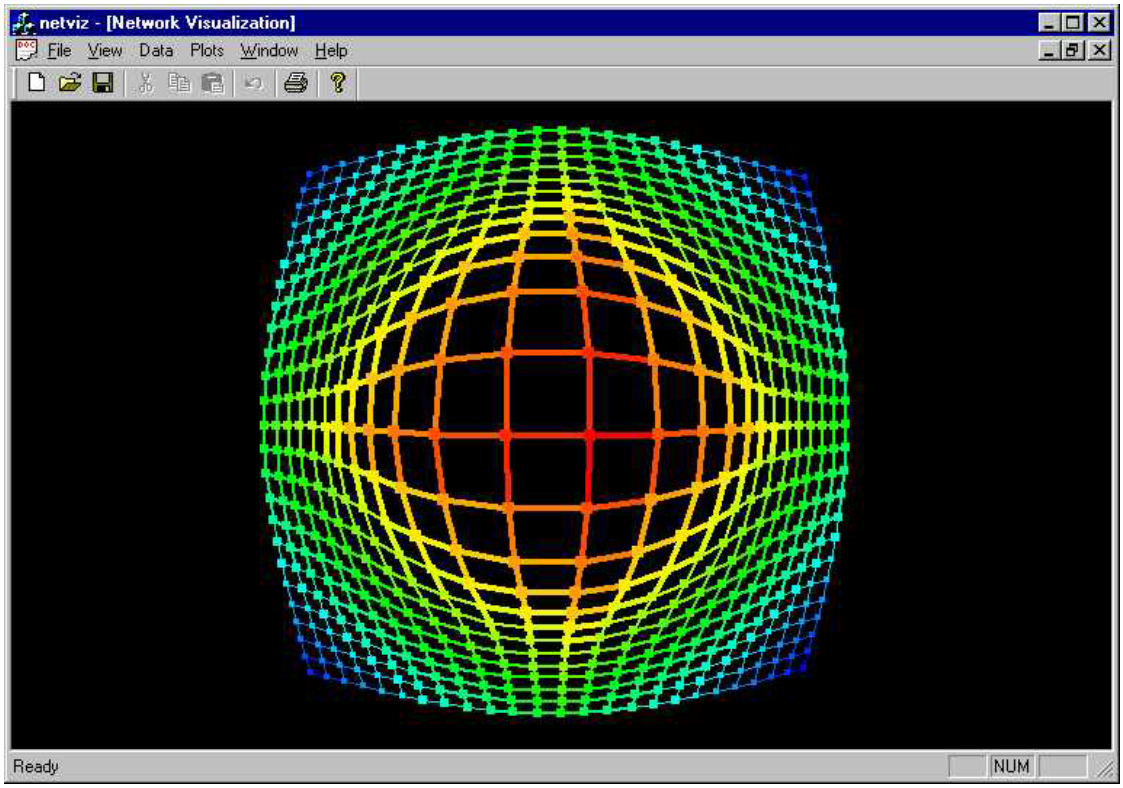


Figure 6: Radial Lens, Applied to a Fixed Position Network Plot

You might have noticed in the previous network plot samples that the nodes and links had colors assigned to them. NV Workshop provides a facility to assign colors to nodes and links based on user-designated variables in the respective node and link data sets. In addition, you can assign different shapes to the nodes based on a user-designated variable in the node data set. The **Data: Edit Attributes** dialog box from the NV Workshop pull-down menus enables you to assign these attributes.

There are many applications available in the marketplace for visualizing networks or drawing graphs, and there are even more for creating statistical plots for performing exploratory data analysis. What differentiates NV Workshop from these other packages is the use of statistical plots in conjunction with the network plots to uncover patterns buried in the data. A plot feature unique to NV Workshop that makes this combination particularly powerful is the ability to put the plots into what we refer to as Selection Modes.

**SELECTION MODE**

Selection Mode refers to the technique or algorithm the graphics will use to display selected observations in their plots. Recall that all graphics in NV Workshop are linked through a common underlying data model and that all graphics are notified when observations are selected in any graphic. NV Workshop's graphics support two types of linked selection: Global Selection Mode and Local Selection Mode.

*Global Selection Mode* is the traditional linked-selection technique used in SAS/INSIGHT and other products. In Global Selection Mode, all data views (plots and data tables) share a common observation selection state. When you select an observation in one view, that observation is treated as selected in all other views. Global Selection Mode, in essence, enables you to graphically subset data at a single level. In a situation in which there are three plots, A, B, and C, selecting a subset of the observations in plot A causes plots B and C to treat that same subset as selected.

*Local Selection Mode* is a new graphical data analysis technique developed at SAS Institute. In Local Selection Mode, each data view has a private observation selection state. Local Selection Mode enables you to graphically subset data at multiple levels. In the aforementioned situation of three plots, A, B, and C, Local Selection Mode enables you to configure plot C to display either the union or the intersection of the selections made in plots A and B.

While in Local Selection Mode, a data view operates in one of two roles: as a Selector view or as an Observer view. A Selector view enables you to manually select observations. Unlike Global Selection Mode, however, when you select an observation in a Selector view the observation is marked as selected only for that view. A Selector view

treats an observation as selected only if you selected the observation in that view. In the example cited above, plots A and B are Selector views.

An *Observer* view does not allow you to manually select observations. An Observer view treats an observation as selected based on the observation's selection state in the Selector views and on the Observer view's *scheme*. There are two Observer view schemes: Union and Intersection. The Union scheme treats an observation as selected if the observation is selected in *any* of the Selector views. The Intersection scheme treats an observation as selected if the observation is selected in *all* of the Selector views. In the example situation, plot C is an Observer view in the Intersection scheme.

## NV WORKSHOP EXAMPLES

In this section we present two examples of how you can use the features available in NV Workshop to investigate real network data.

### EXAMPLE 1: CREDIT CARD FRAUD

This example describes the use of NV Workshop to investigate patterns of credit card fraud. The data used here represents a sampling from a large credit card transaction database. All transactions (both valid and fraudulent) related to customers with at least one fraudulent credit card transaction were extracted from the database. A visualization of this data is shown in Figure 7. The yellow nodes in the network represent merchants, and the light blue nodes represent customers. Each link represents a transaction between a customer and a merchant; the blue links correspond to valid transactions while the red links correspond to fraudulent transactions.

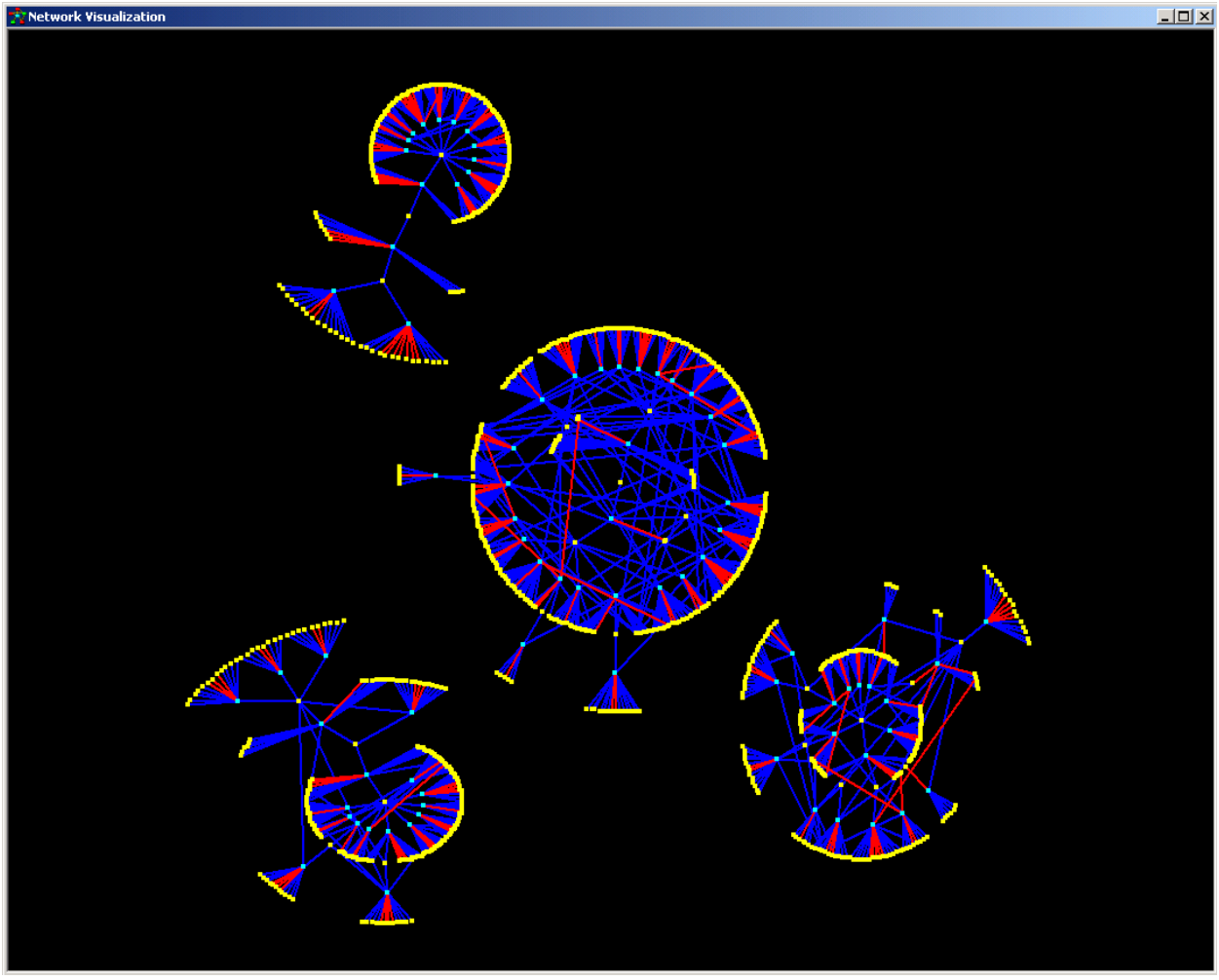


Figure 7: Credit Card Transaction Network

The node data set for this example, CCNODES, contains a total of 80 customers and 985 merchants, and the NODE_ID variable lists the customer and merchant names. The merchants and customers are classified by categories and subcategories given by the CATEGORY and SUBCATEGORY variables in the node data set. The TYPE variable indicates whether the node is a customer (C) or merchant (M). The FRAUDULENT_TRANS_NUM

and TOTAL_TRANS_NUM variables give the numbers of fraudulent transactions and total transactions, respectively. A portion of the node data set, CCNODES, is shown in Figure 8.

ccnodes.sas7bdat

| | NODE_ID | CATEGORY | SUBCATEGORY | TYPE | FRAUDULENT_TRANS_NUM | TOTAL_TRANS_NUM | GROUP |
|---|---|---|---|---|---|---|---|
| | Nom | Nom | Nom | om | Int | Int | Int |
| 70 | Customer070 | Customer | 35to44 | c | 1 | 17 | 4 |
| 71 | Customer071 | Customer | 55to64 | c | 3 | 12 | 4 |
| 72 | Customer072 | Customer | 65andOver | c | 2 | 8 | 4 |
| 73 | Customer073 | Customer | 65andOver | c | 1 | 8 | 4 |
| 74 | Customer074 | Customer | 25to34 | c | 3 | 13 | 4 |
| 75 | Customer075 | Customer | 25to34 | c | 3 | 15 | 4 |
| 76 | Customer076 | Customer | 25to34 | c | 1 | 16 | 4 |
| 77 | Customer077 | Customer | Under25 | c | 2 | 11 | 4 |
| 78 | Customer078 | Customer | 45to54 | c | 4 | 17 | 4 |
| 79 | Customer079 | Customer | 65andOver | c | 2 | 9 | 4 |
| 80 | Customer080 | Customer | 25to34 | c | 4 | 13 | 4 |
| 81 | Merchant0001 | Retail | DrugStores | m | 0 | 1 | 1 |
| 82 | Merchant0002 | Retail | FoodStore | m | 0 | 1 | 1 |
| 83 | Merchant0003 | Services | Restaurants | m | 1 | 1 | 1 |
| 84 | Merchant0004 | Services | Restaurants | m | 0 | 1 | 1 |
| 85 | Merchant0005 | Services | OtherServices | m | 0 | 1 | 1 |
| 86 | Merchant0006 | Services | OtherServices | m | 0 | 1 | 1 |
| 87 | Merchant0007 | Retail | General | m | 1 | 1 | 1 |
| 88 | Merchant0008 | Services | OtherServices | m | 0 | 1 | 1 |
| 89 | Merchant0009 | Retail | GasStation | m | 1 | 1 | 1 |

Figure 8: CCNODES Data Set Excerpt

The link data set (CCLINKS) contains a total of 1,187 credit card transactions. The CUST_ID and MERCH_ID variables correspond to the customer and merchant NODE_ID variable from the node data set. The categories and subcategories from the node data set are carried through to the variables CUST_CAT, MERCH_CAT and MERCH_SUBCAT in the link data set. The FRAUD variable indicates whether the transaction is fraudulent (1) or not (0). A portion of the link data set, CCLINKS, is shown in Figure 9.

cclinks.sas7bdat

| | CUST_ID | CUST_CAT | MERCH_ID | MERCH_CAT | MERCH_SUBCAT | FRAUD | GROUP |
|---|---|---|---|---|---|---|---|
| | Nom | Nom | Nom | Nom | Nom | Int | Int |
| 1 | Customer001 | 65andOver | Merchant0001 | Retail | DrugStores | 0 | 1 |
| 2 | Customer001 | 65andOver | Merchant0002 | Retail | FoodStore | 0 | 1 |
| 3 | Customer001 | 65andOver | Merchant0003 | Services | Restaurants | 1 | 1 |
| 4 | Customer001 | 65andOver | Merchant0004 | Services | Restaurants | 0 | 1 |
| 5 | Customer001 | 65andOver | Merchant0005 | Services | OtherServices | 0 | 1 |
| 6 | Customer001 | 65andOver | Merchant0006 | Services | OtherServices | 0 | 1 |
| 7 | Customer001 | 65andOver | Merchant0007 | Retail | General | 1 | 1 |
| 8 | Customer002 | 35to44 | Merchant0008 | Services | OtherServices | 0 | 1 |
| 9 | Customer002 | 35to44 | Merchant0009 | Retail | GasStation | 1 | 1 |
| 10 | Customer002 | 35to44 | Merchant0010 | Retail | OtherRetail | 0 | 1 |
| 11 | Customer002 | 35to44 | Merchant0011 | Services | OtherServices | 0 | 1 |
| 12 | Customer002 | 35to44 | Merchant0012 | Retail | NonStore | 1 | 1 |
| 13 | Customer002 | 35to44 | Merchant0013 | Services | OtherServices | 0 | 1 |
| 14 | Customer002 | 35to44 | Merchant0014 | Retail | FoodStore | 0 | 1 |
| 15 | Customer002 | 35to44 | Merchant0015 | Other | OtherMerchants | 1 | 1 |
| 16 | Customer002 | 35to44 | Merchant0016 | Retail | GasStation | 0 | 1 |
| 17 | Customer002 | 35to44 | Merchant0017 | Services | OtherServices | 1 | 1 |
| 18 | Customer002 | 35to44 | Merchant0018 | Retail | OtherRetail | 0 | 1 |
| 19 | Customer002 | 35to44 | Merchant0019 | Services | Entertainment | 0 | 1 |

Figure 9: CCLINKS Data Set Excerpt

The CCNODES and CCLINKS data sets were used to produce the transaction network shown in Figure 7 above. After loading the data sets into NV Workshop, the data attributes were assigned by selecting **Data**: **Edit Attributes** from the pull-down menu. For the node data set, the variable NODE_ID is used for the ID and Label attributes, and the TYPE variable is used for the Color attribute. In the link data set the CUST_ID variable is used for the FROM attribute while the TO attribute uses the MERCH_ID variable. The variable named FRAUD is used to color the links. The network plot is created by selecting **Plots**: **Visualize Network**: **Hierarchical** from the pull-down menu.

One form of credit card fraud involves an employee at a merchant stealing customers' credit card numbers and either selling or using them. Under this scenario the fraudulent transaction typically is not directly associated with the merchant where the employee in question works. Instead, the fraudulent transaction is connected to another merchant through a common customer with the original merchant. In a network plot, the fraudulent links (depicted in red here) are typically one link removed from the problem merchant. Although this method is not proof that such a merchant is guilty of committing fraud, it identifies specific merchants that had access to customers' credit card numbers that have been used in fraudulent transactions. These merchants may warrant further investigation.

The *degree* of a network node is the number of links having that node as an endpoint. In this example the degree of a merchant is the number of customers who completed a transaction with the merchant. Recall that this data is limited to customers who had at least one fraudulent transaction associated with their credit card number. Therefore, a merchant with high degree has connections with many customers in the transaction network, thus raising suspicions of problems at this merchant. The goal, then, is to identify merchants with high degree, which is defined in this case as being greater than or equal to three.

## IDENTIFYING SUSPICIOUS MERCHANTS

We use the Zoom Tool to select and magnify the subnetwork (the upper-left subnetwork from the network plot in Figure 7), shown in Figure 10. The pattern in this subnetwork enables the merchants with high degree in group 1 to be detected by direct observation. By using the Label Tool with the network plot, it is easy to see that the merchants with high degree are Merchant0192 and Merchant0193, with degree of 11 and 3, respectively.

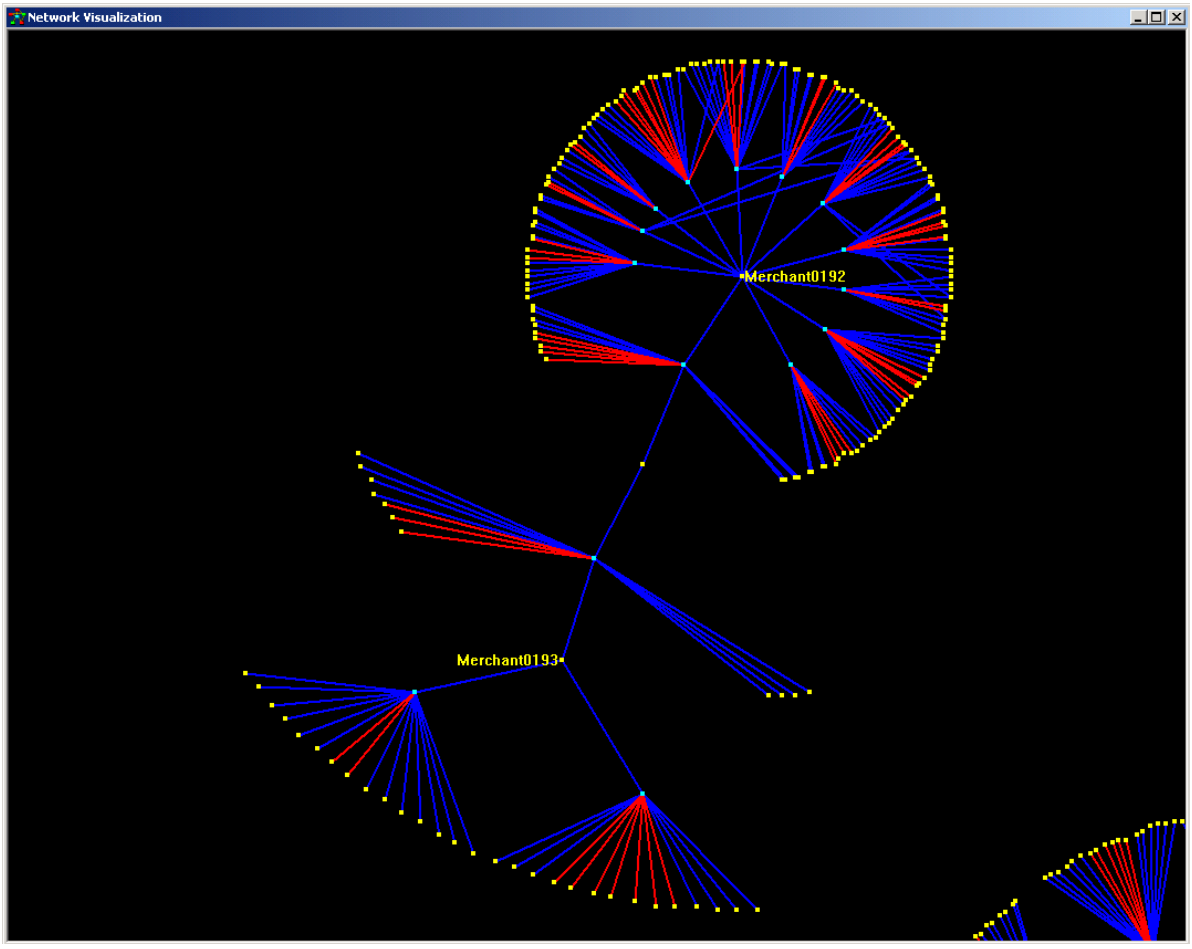


Figure 10: Subnetwork of Group 1

For the remaining three clusters in the network plot, the density of the subnetwork makes it difficult to detect all the merchants with high degree by direct observation. Fortunately, statistical plots and the Local Selection feature of NV Workshop can be used to filter the data in the network plot. Figure 11 shows the result of using a scatter plot with local selection to parse the visualized data. In this plot, the yellow nodes, which can be easily seen, are all the merchants with high degree.

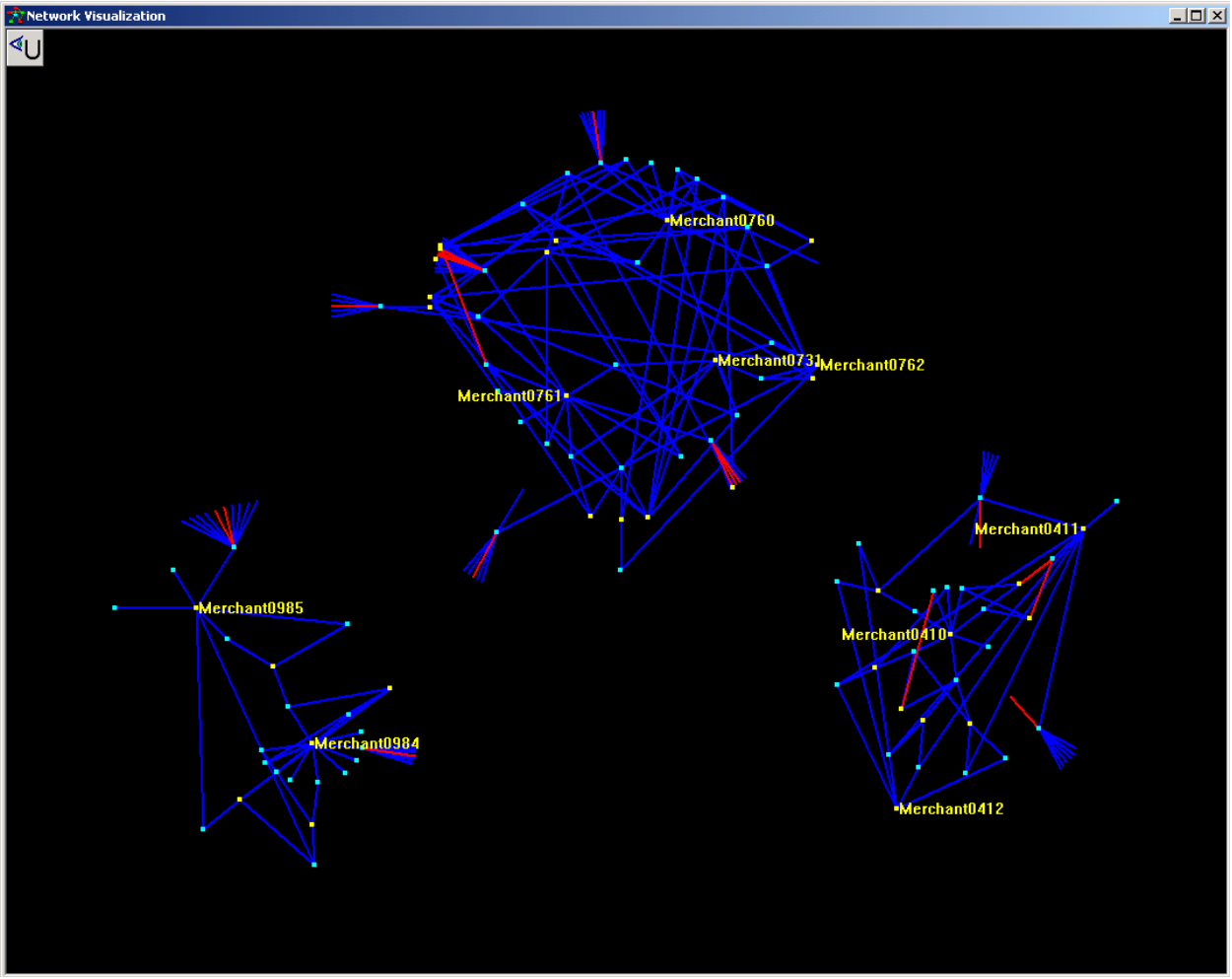


Figure 11: Subnetwork of Merchants with High Degree

The following steps describe how to create the display shown in Figure 11, starting from the network shown in Figure 7:

1. Activate the node data set and create a scatter plot using the **Plots**: **Create Scatterplo**t menu. Select TOTAL_TRANS_NUM as the X variable and GROUP as the Y variable.
2. In the scatter plot, right-click on the horizontal axis to open the **Axis Properties** window. Here you can change the axis scale, as shown in Figure 12. Similarly, you can change the scale of the vertical axis. The scatter plot is shown in Figure 13.
3. In the scatter plot, select the merchants with high degree in groups 2, 3, and 4; i.e., select all yellow nodes with X coordinate greater than or equal to 3 and Y coordinate greater than or equal to 2.
4. Right-click on the graph area of the network plot and choose **Selection Mode**. Select **Local selection mode: Observer View**:**Union**.
5. In the scatter plot, select all customers (i.e., all blue nodes) with the **Shift** key depressed.
6. Right-click on the graph area of the network plot and choose the **Zoom Tool** to magnify the plot.
7. Right-click on the graph area of the network plot and choose the **Label Tool** to add labels.

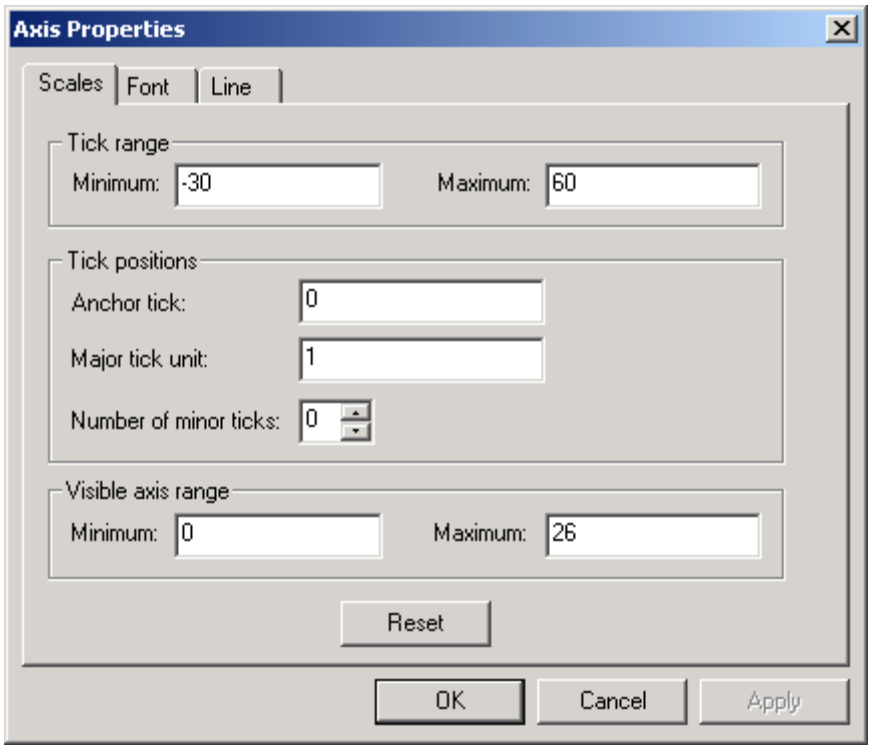


Figure 12: Horizontal Axis Properties

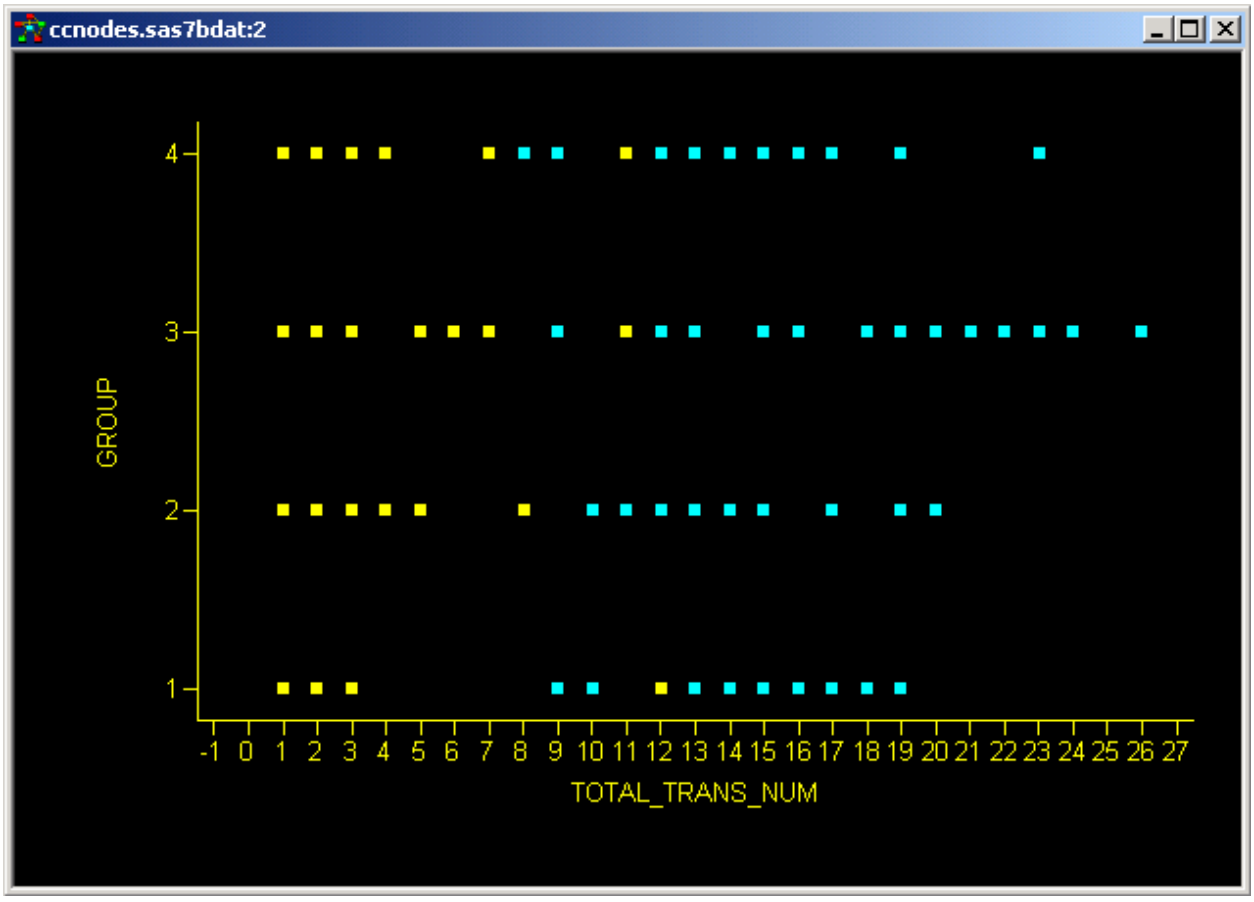


Figure 13: Scatter Plot

In summary, this example shows how to use NV Workshop to investigate credit card fraud. By using the visualization features and observation-filtering capabilities of NV Workshop, you can identify merchants who warrant additional attention with regard to fraudulent credit card transactions.

**EXAMPLE 2: FORTUNE 100 BOARDS OF DIRECTORS**

This example describes the use of NV Workshop to investigate relationships among boards of directors of Fortune 100 companies circa the year 2001, with a goal of identifying especially influential board members and uncovering company-to-company relationships created by sharing common directors. The network of Fortune 100 boards of directors is shown in Figure 14. The yellow nodes in the network represent directors and the light blue nodes represent corporations. Each arc links a director and a corporation; the red arcs link CEOs and their corporations, while blue arcs link general directors and their corporations.

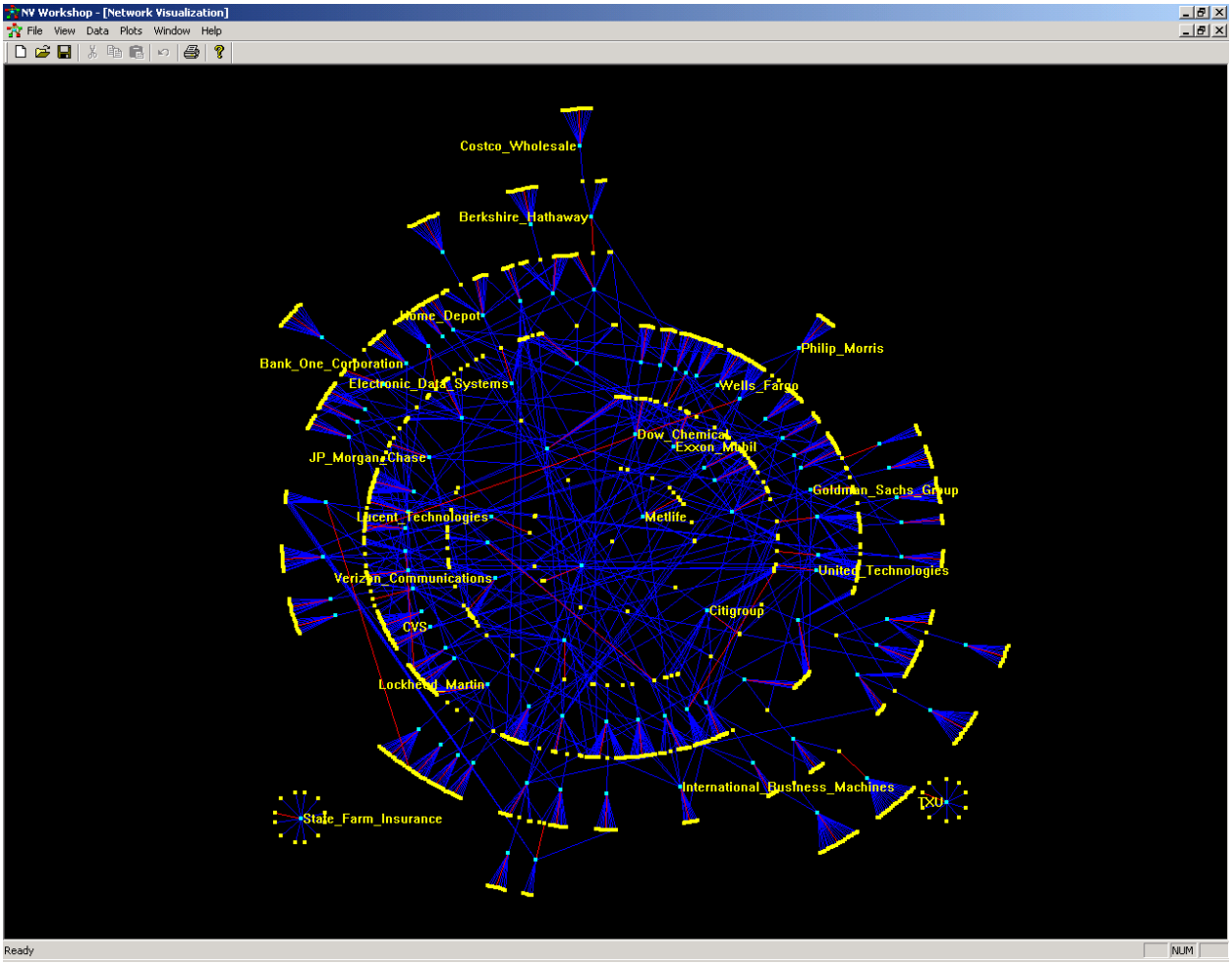


Figure 14: Board of Directors Network

For this example the node data set BDNODES contains a total of 1,178 individuals and corporations. The NAME and FULL_NAME variables list the names for the individuals and corporations. The corporations are classified by categories, as given by the CATEGORY variable. In addition, this variable indicates which individuals are CEOs. The TYPE variable indicates whether the node is an individual (I) or corporation (C). For an individual, the NUMBER_OF_BOARDS variable lists the number of boards of which that individual is a member. For a corporation, this variable lists the size of its board. The REVENUES of corporations are their real revenues, while the REVENUES of individuals do not have any actual meaning. A portion of the node data set, BDNODES, is shown in Figure 15.

bdnodes.sas7bdat

| | | CATEGORY | NAME | FULL_NAME | TYPE | NUMBER_OF_BOARDS | REVENUES |
|---|---|---|---|---|---|---|---|
| | | Nom | Nom | Nom | om | Int | Int |
| 1070 | χ² | Individual | Carter_MH | Mollie_H_Carter | I | 1 | 15000 |
| 1071 | χ² | Individual | Coan | Gaylord_Coan | I | 1 | 15000 |
| 1072 | χ² | Individual | De Boon | Herman_De Boon | I | 1 | 15000 |
| 1073 | χ² | Individual | Mimran | David_Mimran | I | 1 | 15000 |
| 1074 | χ² | Individual | Mulroney_M | M_Mulroney | I | 1 | 15000 |
| 1075 | χ² | Individual | Strauss | Robert_Strauss | I | 1 | 15000 |
| 1076 | χ² | Individual | Vanier | John_Vanier | I | 1 | 15000 |
| 1077 | χ² | Individual | Webb_O | O_Webb | I | 1 | 15000 |
| 1078 | χ² | Individual | Young_A | Andrew_Young | I | 1 | 15000 |
| 1079 | χ² | Retail Trade | Walmart | Wal-Mart_Stores | C | 16 | 219812 |
| 1080 | χ² | Energy | ExxonMobil | Exxon_Mobil | C | 13 | 191581 |
| 1081 | χ² | Manufacturing | GM | General_Motors | C | 14 | 177260 |
| 1082 | χ² | Manufacturing | Ford | Ford_Motor | C | 14 | 162412 |
| 1083 | χ² | Energy | Enron | Enron | C | 17 | 138718 |
| 1084 | χ² | Manufacturing | GE | General_Electric | C | 16 | 125913 |
| 1085 | χ² | Finance And Insurance | Citigroup | Citigroup | C | 17 | 112022 |
| 1086 | χ² | Energy | ChevronTexaco | Chevron_Texaco | C | 15 | 99699 |
| 1087 | χ² | Technology | IBM | International_Business_Machines | C | 14 | 85866 |
| 1088 | χ² | Others | PhilipMorris | Philip_Morris | C | 14 | 72944 |
| 1089 | χ² | Telecommunication | Verizon | Verizon_Communications | C | 16 | 67190 |
| 1090 | χ² | Finance And Insurance | AIG | American_International_Group | C | 20 | 62402 |

Figure 15: BDNODES Data Set Excerpt

The link data set BDLINKS contains a total of 1,332 links between directors and corporations. The BOARD_CHAIRMAN variable indicates whether the individual is the board chairman of the corresponding corporation (1) or not (0). Similarly, CEO indicates whether the individual is the CEO of the corresponding corporation (1) or not (0). NUMBER_OF_BOARDS lists the number of boards on which the individual sits, and BOARD_SIZE lists the size of the board. A portion of the link data set, BDLINKS, is shown in Figure 16.

bdlinks.sas7bdat

| | | INDIVIDUAL | CORPORATION | BOARD_CHAIRMAN | CEO | NUMBER_OF_BOARDS | BOARD_SIZE |
|---|---|---|---|---|---|---|---|
| | | Nom | Nom | Int | Int | Int | Int |
| ▶ | χ² | Breyer | Walmart | 0 | 0 | 1 | 16 |
| 2 | χ² | Chambers | Walmart | 0 | 0 | 2 | 16 |
| 3 | χ² | Coughlin | Walmart | 0 | 0 | 1 | 16 |
| 4 | χ² | Friedman_S | Walmart | 0 | 0 | 2 | 16 |
| 5 | χ² | Gault | Walmart | 0 | 0 | 1 | 16 |
| 6 | χ² | Glass | Walmart | 0 | 0 | 1 | 16 |
| 7 | χ² | Hernandez | Walmart | 0 | 0 | 1 | 16 |
| 8 | χ² | Lepore | Walmart | 0 | 0 | 1 | 16 |
| 9 | χ² | Reason | Walmart | 0 | 0 | 1 | 16 |
| 10 | χ² | Sanders | Walmart | 0 | 0 | 1 | 16 |
| 11 | χ² | Scott_H | Walmart | 0 | 1 | 1 | 16 |
| 12 | χ² | Shewmaker | Walmart | 0 | 0 | 1 | 16 |
| 13 | χ² | Soderquist | Walmart | 0 | 0 | 1 | 16 |
| 14 | χ² | Villarreal | Walmart | 0 | 0 | 1 | 16 |
| 15 | χ² | Walton_J | Walmart | 0 | 0 | 1 | 16 |
| 16 | χ² | Walton_S | Walmart | 1 | 0 | 1 | 16 |
| 17 | χ² | Boskin | ExxonMobil | 0 | 0 | 1 | 13 |
| 18 | χ² | Dahan | ExxonMobil | 0 | 0 | 1 | 13 |
| 19 | χ² | Esrey | ExxonMobil | 0 | 0 | 3 | 13 |
| 20 | χ² | Fites | ExxonMobil | 0 | 0 | 2 | 13 |

Figure 16: BDLINKS Data Set Excerpt

The BDNODES and BDLINKS data sets are used to produce the network shown in Figure 14 above. After loading the data sets into NV Workshop, the data attributes are assigned by selecting **Data**: **Edit Attributes** from the pull-down menu. The attributes and associated variables from BDNODES are *Id:* NAME, *Label:* FULL_NAME, and *Color:* TYPE. For BDLINKS they are *From:* INDIVIDUAL, *To:* CORPORATION, and *Color:* NUMBER_OF_BOARDS. The network plot is created by selecting **Plots**: **Visualize Network**: **Hierarchical** from the pull-down menu.

IDENTIFYING LINKS BETWEEN CORPORATIONS

The density of the network shown in Figure 14 above makes it difficult to detect patterns involving individuals and corporations of interest, such as the individuals sitting on multiple boards. By using a combination of data tables, statistical graphs, and network plots, subnetworks can be displayed clearly. The subnetwork involving the individuals sitting on four or more different boards is shown in Figure 17. From this you can see that there are nine individuals who sit on four or more different boards and that the most "influential" individual, William Gray, sits on six different boards. There are also two directors who sit on five boards. These patterns suggest possible relationships between companies that might not be readily apparent at first glance.

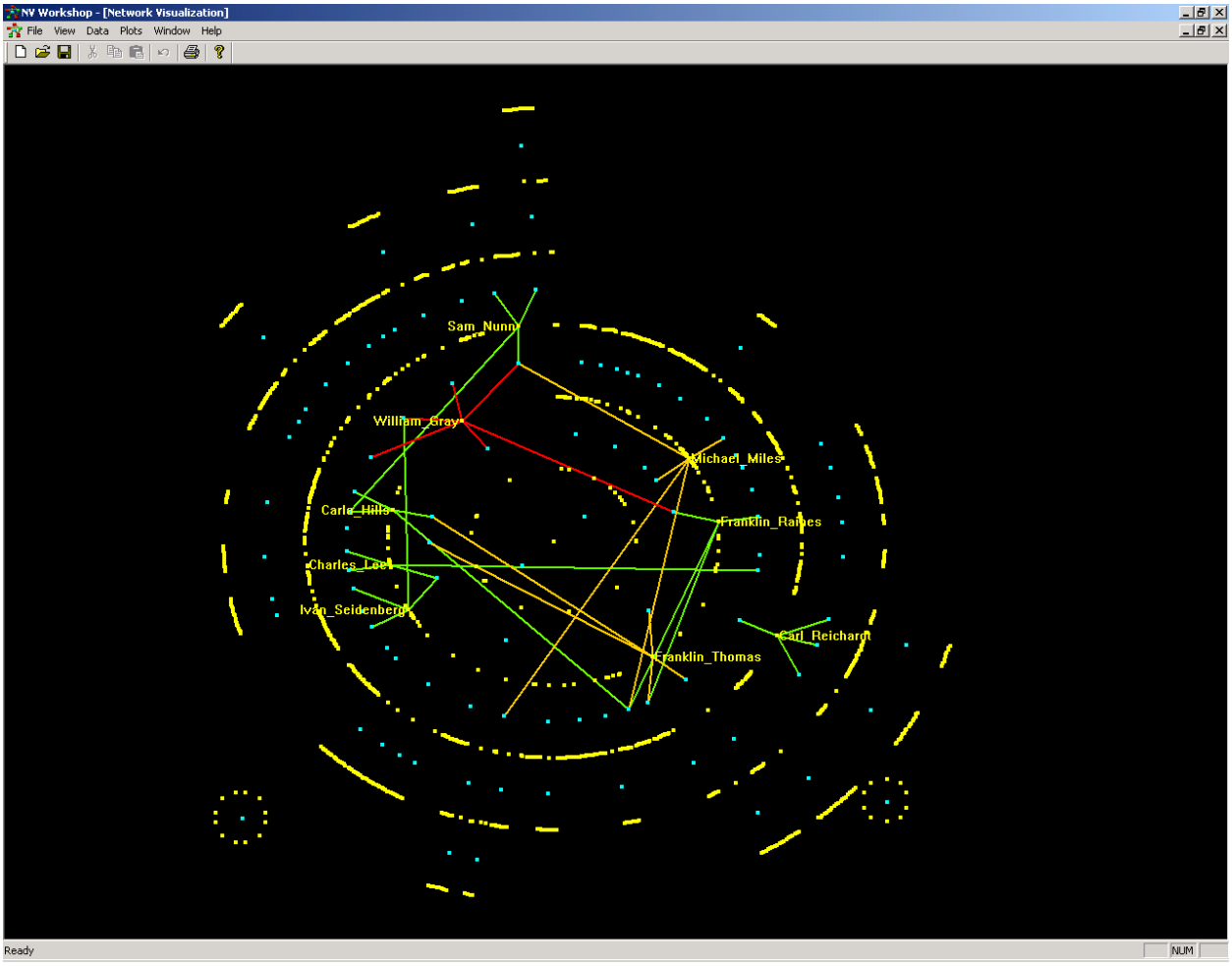


Figure 17: Subnetwork of Corporate Links

The following steps describe how to create the display shown in Figure 17:

1. Load data using **File**: **Open**:
   - Node data set: bdnodes.sas7bdat
   - Link data set: bdlinks.sas7bdat (select **Link** for **Data Type** in the **Open** dialog box)
2. Assign data attributes using **Data**: **Edit Attributes:**
   - Node id: NAME
   - Node label: FULL_NAME
   - Node color: TYPE
   - Link from: INDIVIDUAL
   - Link to: CORPORATION
   - Link color: NUMBER_OF_BOARDS
3. Create a network plot with **Plots**: **Visualize Network**: **Hierarchical.**
4. Activate the link data set and create a histogram by selecting **Plot**: **Create Histogram** from the pull-down menu. Select NUMBER_OF_BOARDS as the X Variable.

5. In the histogram, right-click on the horizontal axis to open the **Axis Properties** window. Here you can change the axis scale, as shown in Figure 18.
6. In the histogram, select the bars for the numbers greater than or equal to four, as shown in Figure 19.
7. Right-click on the graph area of the network plot to open the **Graph Properties** window. Here you can increase the line width, as shown in Figure 20.

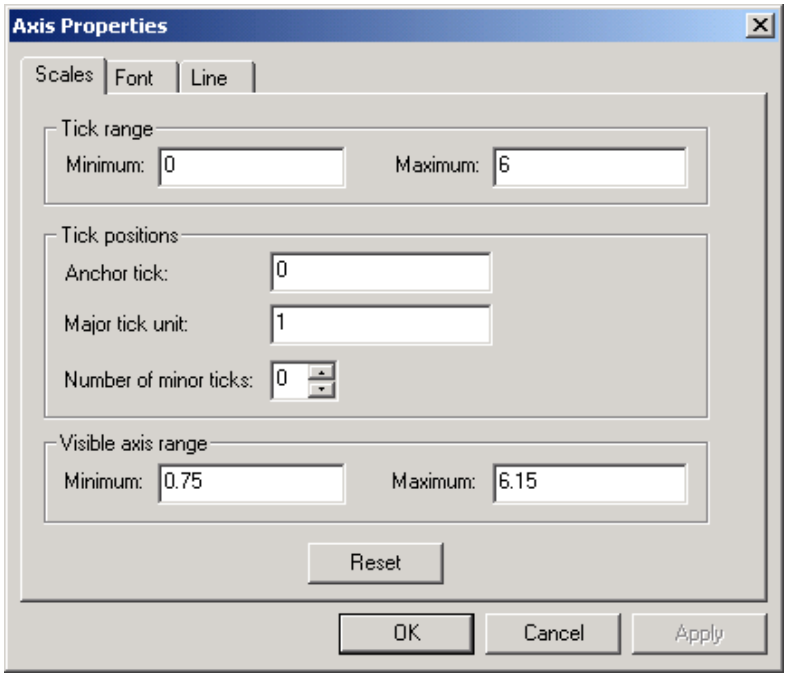


Figure 18: Axis Properties Assignments

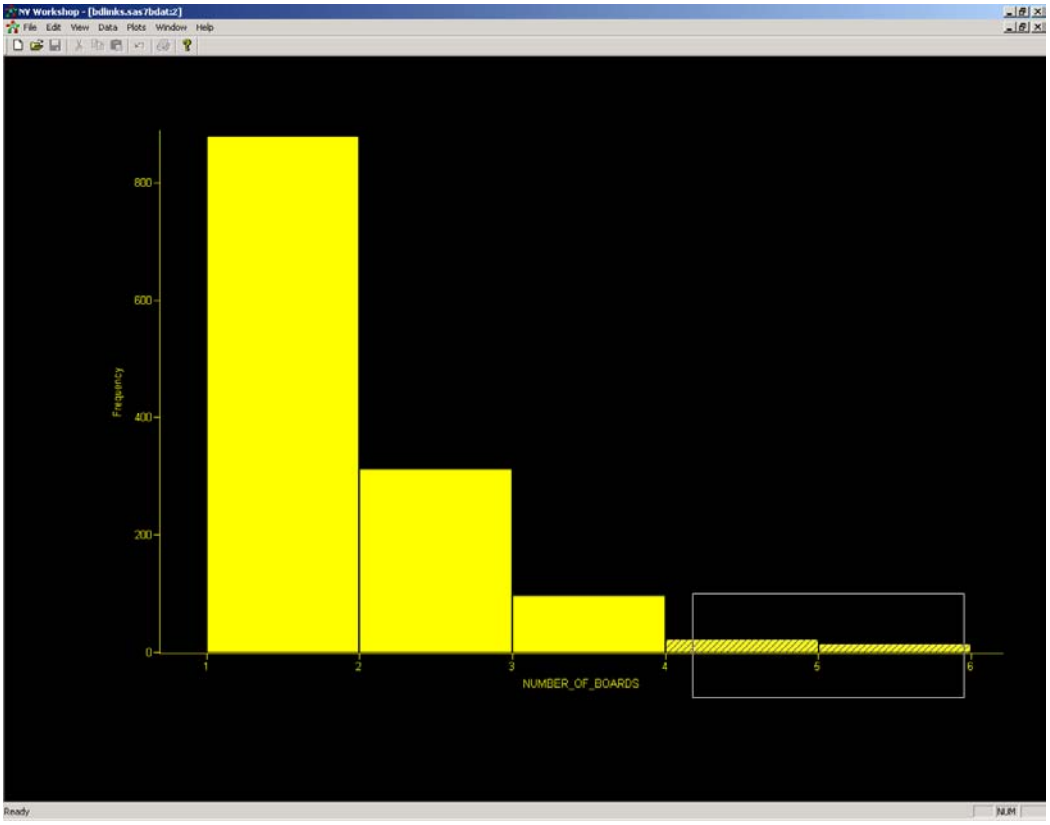


Figure 19: Histogram

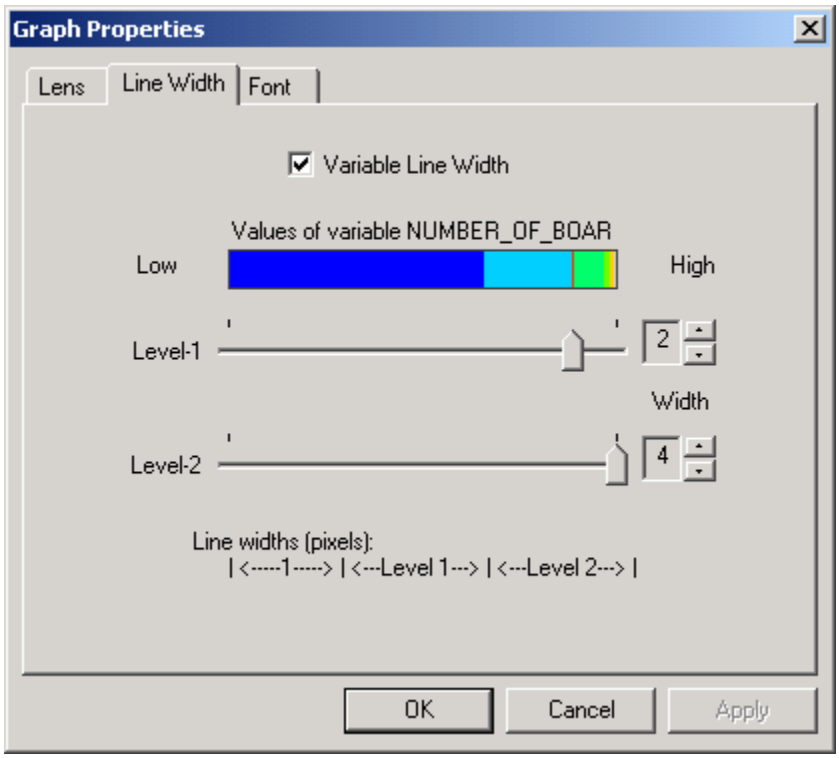


Figure 20: Line Width Assignment

DETECTING INTERLOCKS

According to the Investor Responsibility Research Center's (IRRC) definition, there is an *interlock* when two CEOs sit on the boards of each other's companies. Interlocks are often viewed with a degree of suspicion; therefore, identifying interlocks in the network is crucial. The subnetwork of interlocks is shown in Figure 21. Here you can see that there are two interlocks: Sanford Weill, the CEO of Citigroup, sits on the boards of AT&T and United Technologies; Michael Armstrong and George David, the CEOs of AT&T and United Technologies, respectively, sit on the board of Citigroup.

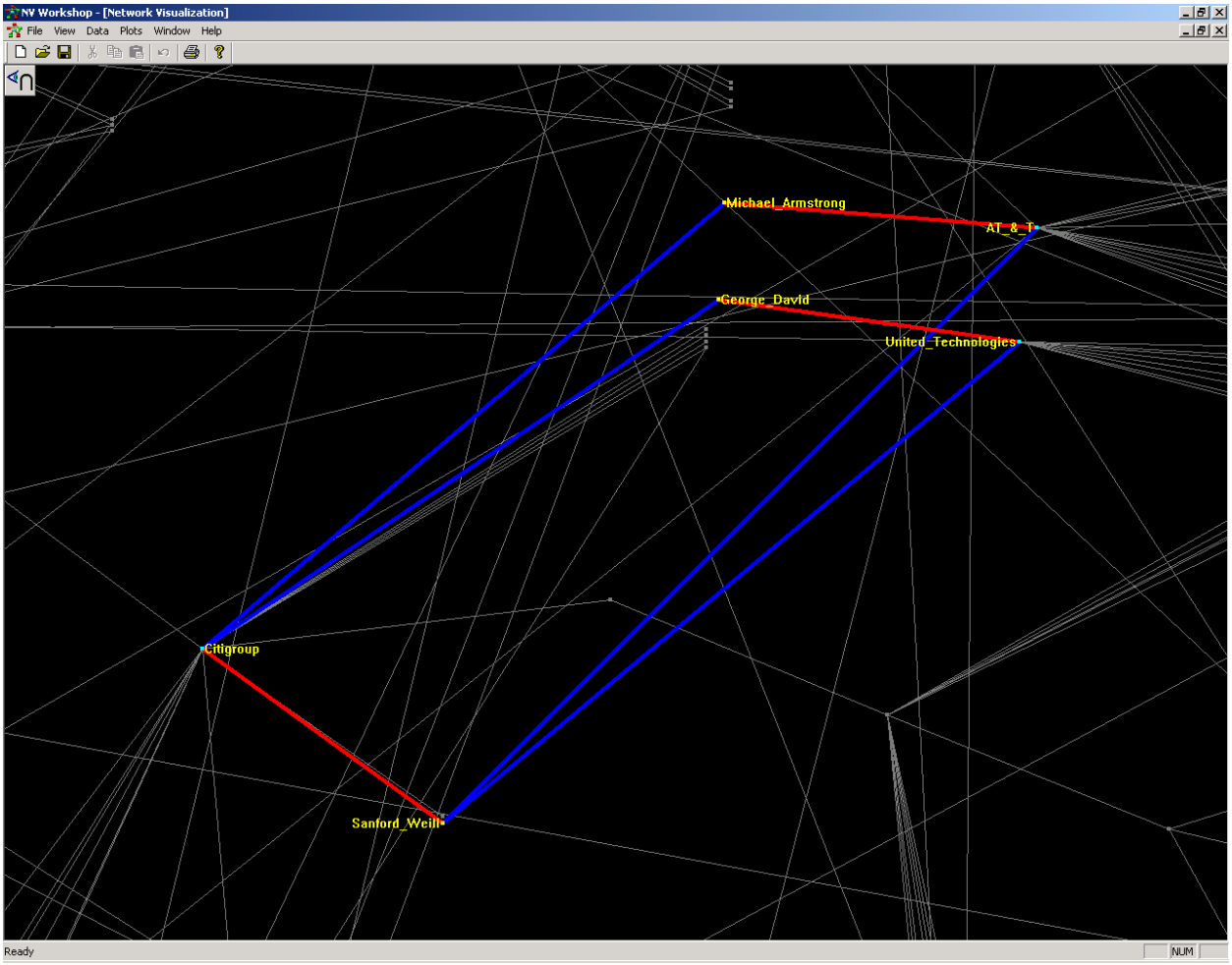


Figure 21: Subnetwork of Interlocks

The interlocks shown in Figure 21 were detected by examining the data in Figure 22 showing the subnetwork of CEOs who sit on two or more boards. A quadrilateral consisting of two CEOs and two corporations represents an interlock; the interlocks can be detected by finding the quadrilaterals in this much sparser subnetwork. Figure 21 is simply a zoomed version of Figure 22, with all three executives labeled.

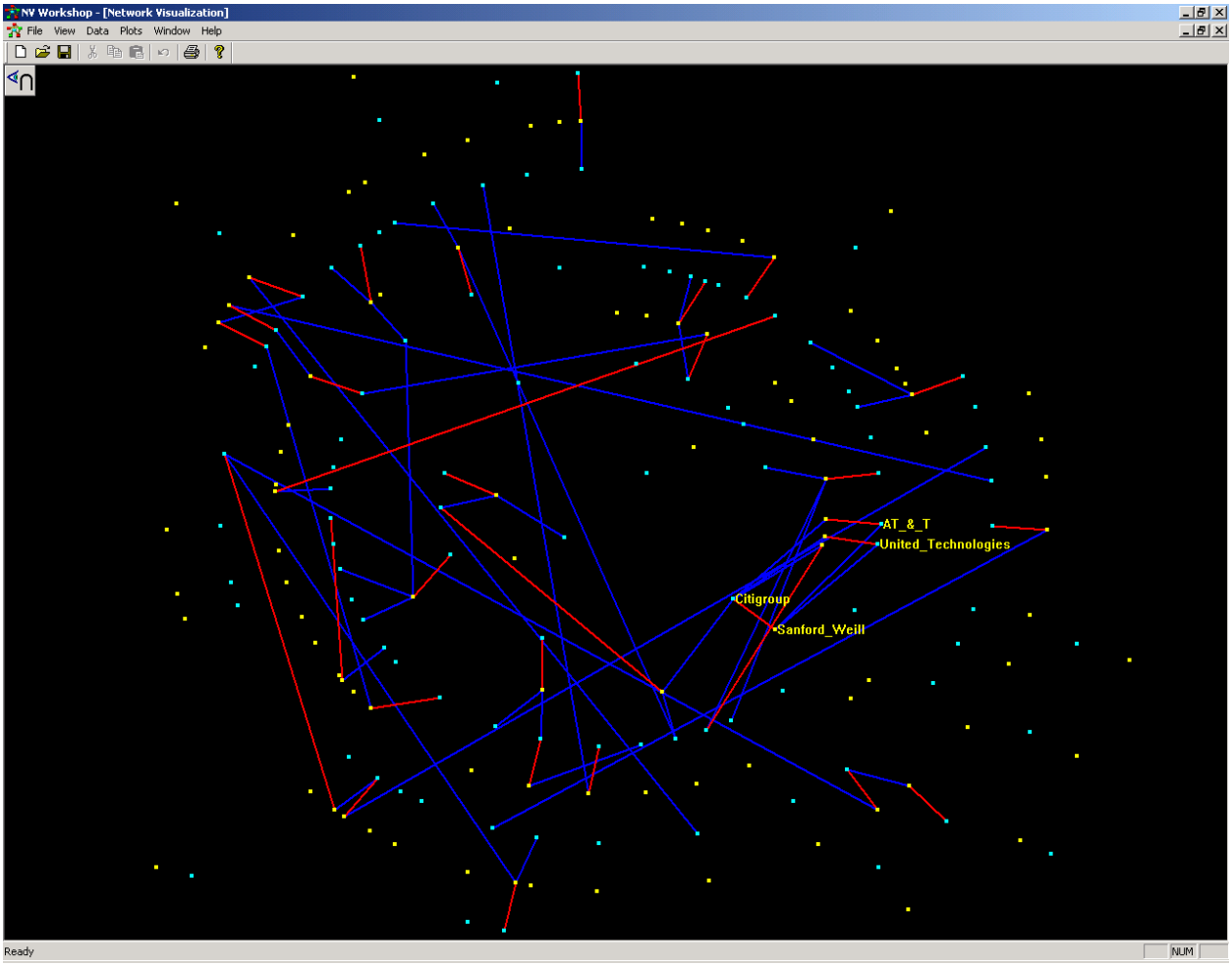


Figure 22: Subnetwork of CEOs Who Sit on Two or More Boards

Suppose a hierarchical network plot has been created by following steps 1 to 3 for creating Figure 17. The following steps describe how to generate the display in Figure 22:

1. Select the node data set and create a histogram by selecting **Plot**: **Create Histogram** from the pull-down menu. Select CATEGORY as the X Variable.
2. In the histogram, select the bar for CEO.
3. On the network plot, right-click and choose **Selection Mode**. Select **Local selection mode**: **Observer View**: **Intersection**.
4. Select all observations in the node data set.
5. In the histogram, select the bar for Other with the **Shift** key depressed.
6. Activate the link data set and create a histogram by selecting **Plot**: **Create Histogram**. Select NUMBER_OF_BOARDS as the X Variable.
7. In the histogram, right-click on the horizontal axis to open the **Axis Properties** window. Here you can change the axis scale as shown in \fref{axis}.
8. In the histogram, select the bars for the numbers greater than or equal to two.
9. Right-click on the graph area of the network plot to open the **Graph Properties** window. Here you can increase the line width.
10. Right-click on the graph area of the network plot and choose the **Zoom** Tool to magnify the plot.
11. Right-click on the graph area of the network plot and choose the **Label** Tool to add labels to the network plot.

In summary, this example shows how to use NV Workshop to investigate relationships among Fortune 100 boards of directors. By using the visualization features and observation filtering capabilities of NV Workshop, you can identify the influential individuals and detect the interlocks.

## CONCLUSION

There is no question that networks and network data play an increasingly prominent role for more and more businesses, organizations, and individuals. Just as surely as the importance of network data is growing, the scope and detail of network data is growing as well. Unfortunately, larger network data—and especially very large network data—is difficult to interpret and explore. This creates the paradox of network data becoming more and more difficult to use just as it becomes more and more important that we use it.

NV Workshop, an interactive graphics-oriented application, enables you to work effectively with large network data by providing statistical and network graphics that cut through the volume of the data to help you highlight the important relationships concealed by the sheer scale and detail of the network. Its flexibility enables you to customize your graphical data investigation methods in order to utilize the individual and linked data views and graphical plots that most readily provide you with useful insights.

## REFERENCES

"Introduction to the SAS/OR Network Visualization Workshop," SAS Institute Inc., 2005.

## RECOMMENDED READING

For further information on the science of networks, the following works are recommended:

Barabási, Albert-László, *Linked: The New Science of Networks,* Perseus Publishing, 2002.
Buchanan, Mark, *Nexus—Small Worlds and the Groundbreaking Science of Networks,* Norton & Company, 2002.
Watts, Duncan J., *Six Degrees—The Science of a Connected Age,* Norton & Company, 2003.

## CONTACT INFORMATION

Ed Hughes
SAS Institute Inc.
SAS Campus Drive, R5322
Cary, NC 27513
Work Phone: 919-531-6916
Fax: 919-531-9445
Email: Ed.Hughes@sas.com

Phil Meanor
SAS Institute Inc.
SAS Campus Drive, R5416
Cary, NC 27513
Work Phone: 919-531-6043
Fax: 919-677-4444
Email: Phillip.Meanor@sas.com

For more information, you may also contact:

Ravi Devarajan
SAS Institute Inc.
SAS Campus Drive, R2242
Cary, NC 27513
Work Phone: 919-531-7793
Fax: 919-677-4444
Email: Ravi.Devarajan@sas.com