Chapter 4
Visualization

Introduction

Much of the work conducted in the social sciences concerns a search for patterns and regularities in data (Greenacre 1984). In the course of such a search, a variety of formal mathematical and statistical models are generally employed in attempts to better describe, understand, and explain aspects of relatively complex social systems. Whereas these models may inform us as to, for example, the goodness-of-fit between any model and the original data, it is still incumbent on the researcher to interpret just what that fit means or entails from a theoretical standpoint and, more importantly, to communicate these interpretations to others. In yet other instances, we may be interested in exploring the relationships among a complex set of social entities in a more exploratory mode in the hopes of developing better theoretical insights that may be tested in future analysis or research (Glaser and Strauss 1967). In either case, our need to better interpret our explanatory models or our need to produce better explanatory models through an exploration of the data can be made simpler and more effective with the use of visualization techniques. There is an abundance of work that clearly shows the importance of visual representations in communicating information. But in our attempts to communicate and explore we must be concerned about the most effective way to represent properties of the network data so that we ensure the most valid interpretation possible.

One of the most important constraints on the valid graphical representation of data and models concerns human limitations of perception. There has been a great deal of work on this topic and an in-depth discussion is beyond the scope of this chapter. For a good review of issues concerning human color perception and the communication of network graphical information see Krempel (2002) or for a more general review see Munzner (2000). For our purposes we seek to explore the ways in which information can be communicated in network graphs that provide adequate representations of the various properties of nodes and arcs that may aid in a better understanding of network structure. In addition we seek means for reducing the complexity of network structures by discussing methods for reducing the graphical complexity of network representations through such mechanisms as turning on and off nodal or arc clusters or categories in order to reveal potentially hidden structural properties. There has been a long history of the use of graphs in the study of social networks. Freeman (2000) has provided a historical overview of network graphs.

One of the first things most people want to do with network data is construct a visual representation of it – in short, draw a picture. Seeing the network can provide a qualitative understanding that is hard to obtain quantitatively. A network diagram consists of a set of points representing nodes and a set of lines representing ties. Various characteristics of the points and lines, such as color, size, and shape, can be used communicate information about the nodes and the relationships among them.

This chapter discusses the ins and outs of visualizing social networks. In the discussion to follow, please note that we distinguish carefully between network elements and their graphical representation – i.e., between nodes and the points that represent them, and between ties and the lines that represent them.

Layout

The layout of a network diagram refers to the position of the points in the diagram. It is the most important aspect of network visualization. A badly laid out network diagram communicates very little information. As an example, consider the diagram in Figure xxx.

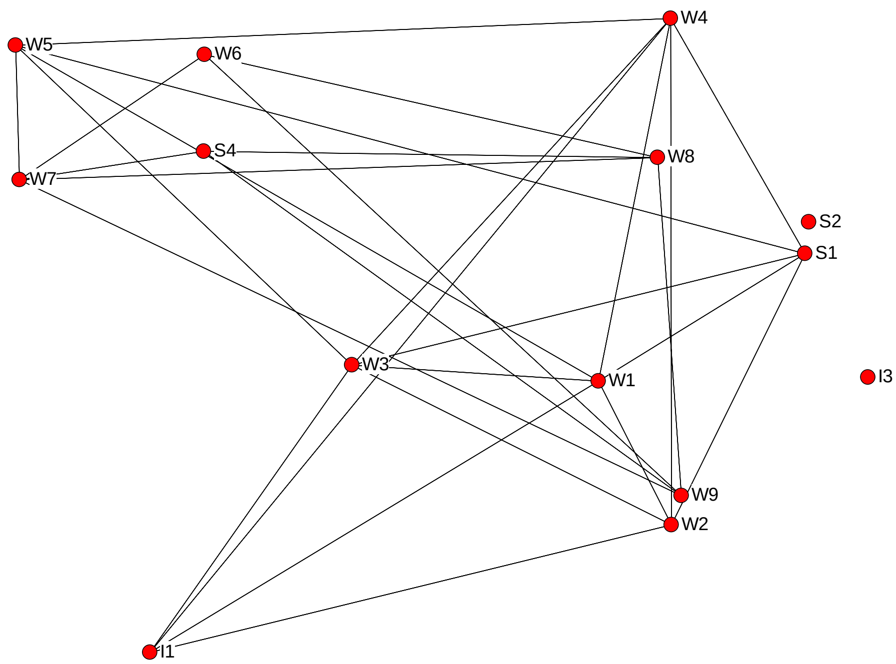


Figure xx. Random layout of the Games relation in the Bank Wiring room dataset.

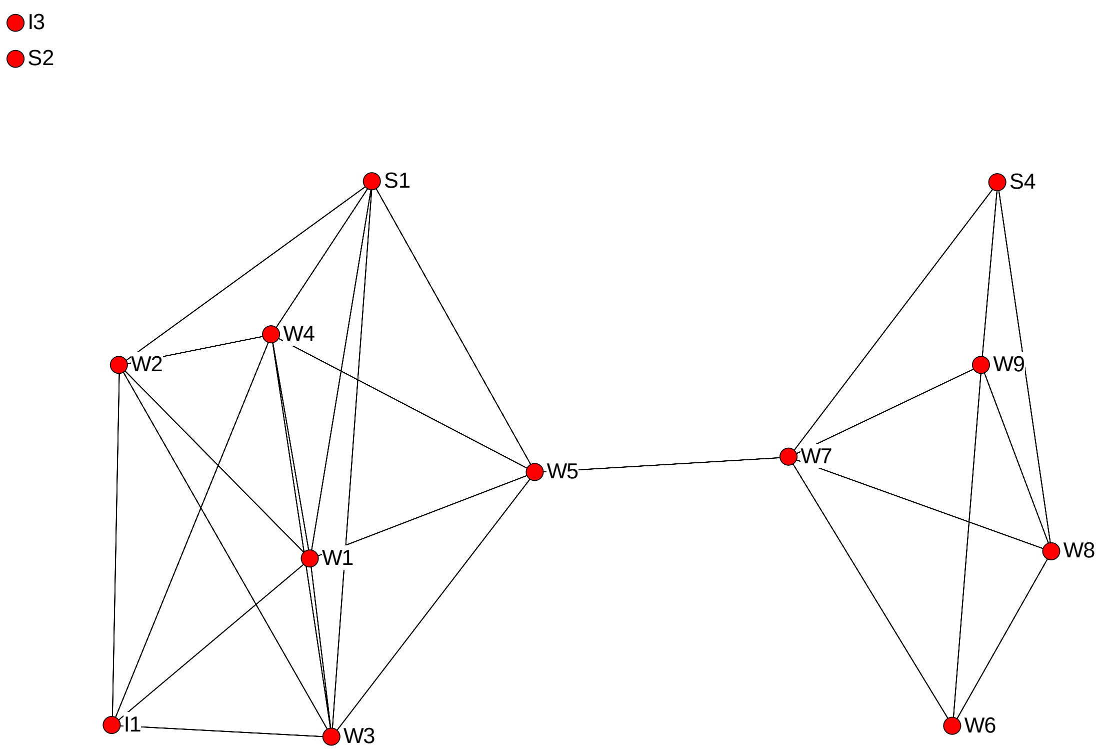


Figure xxx. Games relation rearranged to reveal structure.

There are three basic kinds of methods used to position nodes: attribute-based scatter plotting, scaling or ordination, and graph-theoretic layout algorithms. We call the diagrams produced by each of these methods scatters, ordinations, and graph-layouts respectively.

The scatters are simply scatter plots based on attributes of the nodes. For example, we can plot the points based on the age and income of the corresponding nodes. We then draw lines between the points to represent ties. These kinds of displays are useful when we are interested in visualizing how attributes of the nodes affect who is connected to whom. Scatters are most successful when both attributes are continuous (i.e., not categorical like gender or department), and when the attributes in fact do affect who is tied with whom.

The ordinations are diagrams based on multivariate statistics techniques such as principal components and metric multidimensional scaling. In these layouts, the distances between points are meaningful in the sense that there is a known mathematical relationship between the distances and the social proximities of the nodes. For example, in metric multidimensional scaling, if the data contain information on the strength of ties between nodes, the resulting layout positions the points so that the points that near each other are the ones that are strongly connected to each other, and the nodes that are far apart are the ones that are only weakly connected. When no strengths of tie data are available, the standard thing to do is compute geodesic distances between nodes. By geodesic distance we mean the number of links in the shortest path between a pair of nodes, what is called “degrees of separation” in the popular press. The ordination would then lay out the points corresponding to nodes with high geodesic distance between them would be far apart in the diagram, and the points corresponding to nodes with short geodesic distance would be close together.

Ordinations based on geodesic distance typically work very well in the sense that the resulting diagrams are relatively uncluttered, cohesive subgroups are clearly visible, and points corresponding to the more central nodes tend to be found in the center of the diagram. In addition, they have the advantage of interpretability – we know exactly why some nodes are further apart than others.

The graph-layouts are diagrams based on either heuristic or combinatorial optimization algorithms that try to locate the points in such a way as to optimize a variety of criteria simultaneously. One of these criteria – the correspondence between point distance and path distance between nodes – is the same as in ordinations. But the graph-layouts include other criteria as well, such as preventing nodes from getting too close to each other, minimizing the number of crossed-lines, and a preference for equal-length lines. As a result, the distances between points in the diagram no longer correspond in a 1-to-1 way to path distances between nodes. Thus, we give up a measure of interpretability in order to get cleaner diagrams that are easier to read.

It is important to realize that the information in graph-layouts is contained in the pattern of which nodes are connected to which others. The locations of the points do not necessarily reflect any mathematical or sociological properties – they are chosen based on essentially aesthetic criteria. As such, one must not attach too much meaning to the exact location of a node since the algorithm is not explicitly trying identify cliques or place central nodes in the center.

In addition any arrangement of nodes in space is equally valid as long as no ties are added or dropped. In other words, if we drag a node out of the center and put it on the periphery (dragging all of its ties along with it), the resulting diagram is no less valid than the original. This is not true of scatters and ordinations, in which the physical distances between the points have meanings which would be violated if the points were moved arbitrarily.

Embedding Node Attributes

There are number of obvious means for representing the qualitative or quantitative properties of nodes in a network. In the sections to follow we discuss a number issues surrounding the representation of node and arc properties in NETDRAW. Nodal color, shape, label, or nodal geometric distortions (e.g., ratios) are all examples of possible ways for conveying information of either a nominal or numerical kind. Both qualitative and quantitative information can be simultaneously displayed at a given node as, for example, when a node's color represents group affiliation while size reflects the magnitude of some nodal property. Whereas there are a whole range of possibilities, one must be aware of the interpretability and complexity of graphs as more and more colors and shapes are added. Some work has suggested that no more than six colors should be used in any computer graph (Derefeldt and Marmolin 1981) although others have suggested the maximum may be more like nine (Smallman and Boynton 1990). Much of this will depend on the size and complexity (e.g., density) of the network and the number of attributes to be explored. However, as the number of nodal attributes increase, the ability to conduct dyadic assessments of connections among clustered attributes (e.g., the linkages between republicans and democrats) can become cumbersome due to the number of permutations. Thus, care should be exercised in the selection and number of attributes to be explored.

NETDRAW has the ability to switch among lists of node attributes and aspects of node attributes. In the former the graph author can predefine multiple color, size, or shape lists for nodes to represent different attribute sets to be represented or explored (e.g., a color list based on sets of structurally equivalent nodes and one based on regularly equivalent nodes). The viewer can then interactively switch among lists thereby facilitating comparisons among nodal attribute sets. The program allows for the comparisons of the degree to which nodes have specified attributes from a list of such attributes. Thus, for example, nodal color gradients can be used to represent the degree to which nodes have a specific attribute (e.g., degree of depression, anxiety , tension, etc. among actors in a network).

As stated above, points have a number of properties that can be put to use to communicate information about the nodes they represent. For example, nodes can be different sizes and shapes. The shapes can have rims of varying widths and colors. Nodes can have textual labels attached, which can have varying sizes and fonts. Each of these properties can be mapped to a node attribute, although it is our experience that using more than two properties at a time can be more distracting than informative.

In general, it is best to reserve size differences for representing continuous attributes such as age or rank, while using color and shape differences to represent categorical attributes like gender or department.

An example is given in Figure xx, which depicts the CAMPNET dataset. The size of the nodes is used to represent the betweenness centrality of each node (see the chapter on centrality for an explanation), while the shape of nodes is used to represent gender.

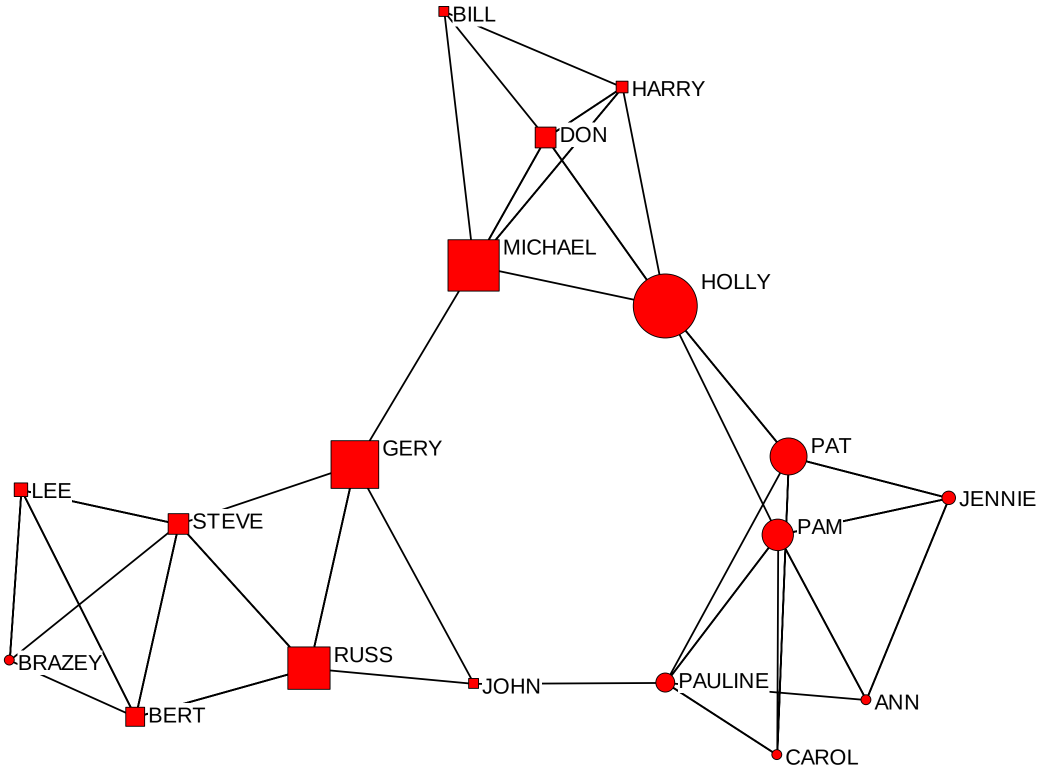


Figure xxx. Circle shapes indicate women and square shapes indicate men. Size corresponds to each node's betweenness centrality.

Color can also be used to represent the nominal properties of nodes in terms of such things as gender, membership in groups, political party affiliation, tribal membership, etc. Any qualitative property of the node can be represented by color. There are a number of considerations for the selection of colors that may either facilitate or hinder the communication of information. Aside from the number of colors, there may be colors that are more intuitive or culturally appropriate for the communication of specific sorts of information. Some might see red as an action color or blue as a color of status or gender. We all know in US culture that the appropriate color for boys is blue and girls pink. Although possibly not politically correct in today's world such use of colors do have cultural foundations and do often aid in conveying complex information. For example, in trying to show gradients (e.g., ratings scales) cool colors can show "comfortable" atomic contacts while red signaled trouble spots. Selective use of bright, contrasting colors draws the eye very effectively to points of emphasis, but overuse of them can easily overwhelm the transfer of real information. Therefore relatively more pale, dim, or similar colors work very well for large amounts of ordinary or framework data.

Color background is also an important issue. A white background works well for most applications but a black background can provide the illusion of empty space giving the illusion of higher dimensionality. However, for some applications in which the use of certain colors is important, or non-network information is being mapped on to the structure, a white background may be most effective.

Node Selection

It is often useful in analyzing networks to see what the network looks like when certain classes of nodes are removed. Sometimes this is done to remove nodes that are peripheral to a given research interest. Other times, it is to gauge the importance of the group in connecting others. Programs like UCINET's NetDraw procedure make it easy to click groups of nodes on and off (as well as individuals).

Ego Networks

Another useful exercise is to examine the ego networks of particular nodes. By "ego networks" we mean the set of ties among the nodes connected to a focal individual (ego). This is particular useful when used to compare the structures around two different egos. For example, Figures xxx and xxx show the acquaintanceship networks two drug injectors in the city of Hartford, CT (Weeks et al).

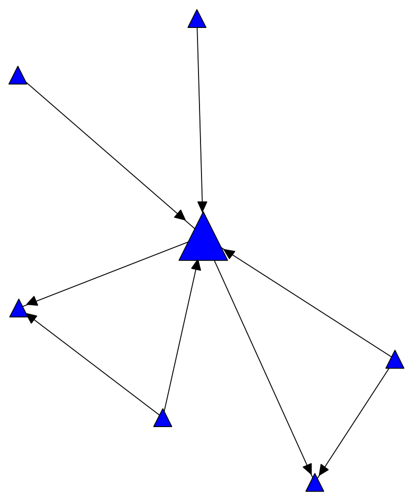


Figure xxx. Relatively open ego network of Puerto-Rican drug-injector. Large node is ego.

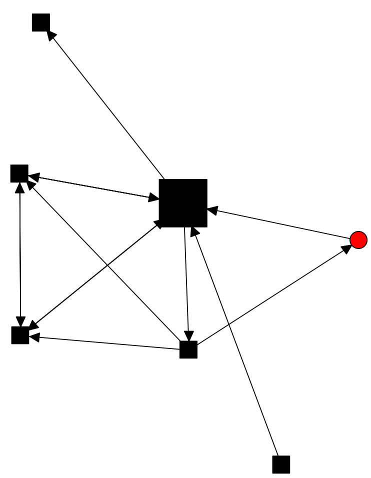
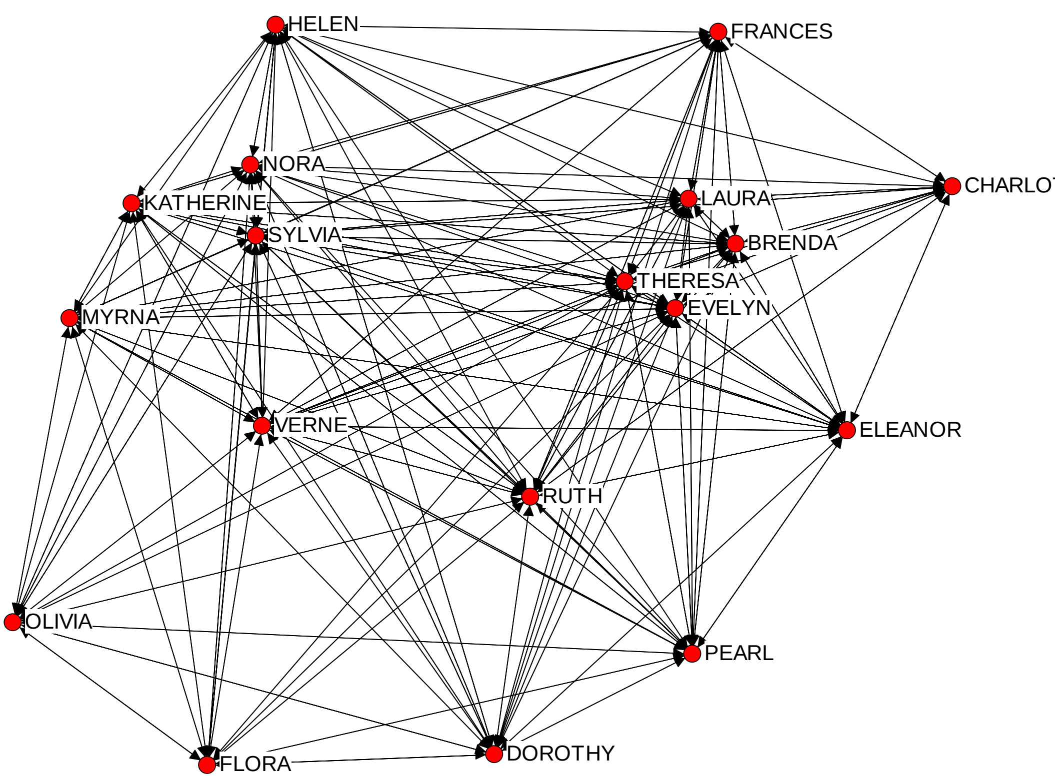


Figure xxx. Relatively closed ego network of African-American drug-injector. Large node is ego.

Strength of Ties

There are several ways to communicate information about strength of ties. One way is to use the distance between points to communicate the strength of tie between the corresponding nodes. This is the ordination approach discussed above. As an example, we use a valued, 1-mode dataset derived from the Davis, Gardner and Gardner () women-by-events data.[1] The data were transformed using the Affiliations procedure in UCINET to create a woman-by-woman matrix in which the cells indicate the number of events that each pair of women attended in common. Figure xx shows an ordination of these data. A line is shown between two points if the corresponding women attended at least one social event in common.



[1] For other ways of visualizing 2-mode data, consult the 2-mode chapter in this book.

Figure xxx. Ordination plot. Distance is inversely proportional to strength of tie.

We can make this diagram easier to read by suppressing lines representing weaker ties. For example, we might show a line between points only if the women attended at least 3 events in common. As shown in Figure xx, this approach makes the 2-group structure of the data very evident.

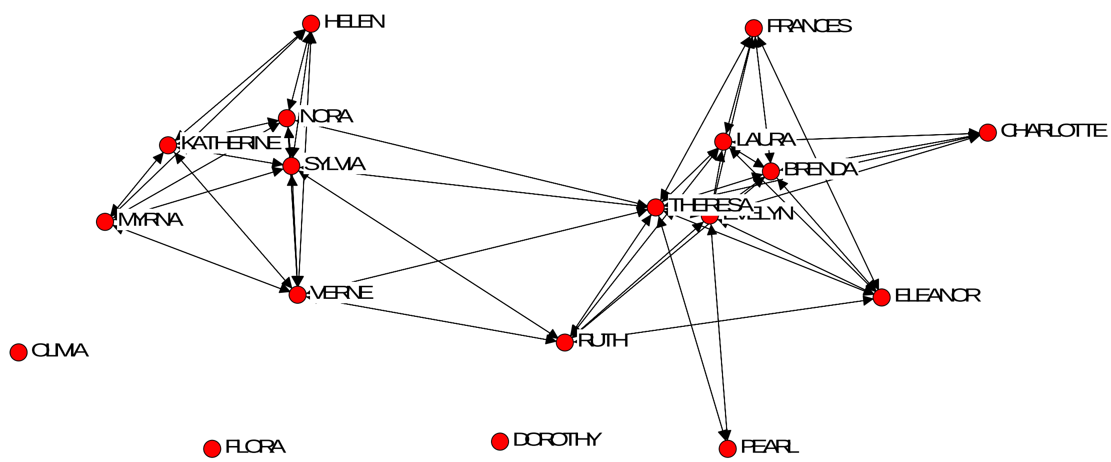


Figure xxx. Ordination plot with lines corresponding to weak ties suppressed.

Another approach to displaying strength of tie information is to make the thickness of lines proportional to the strength of tie. An example is shown in Figure xxx, in which the position of the nodes is determined by ordination. Essentially this diagram uses both physical distance and line thickness to communicate social proximity. One can see that the thicker lines tend to be within the left group and within the right group, but not between the groups.

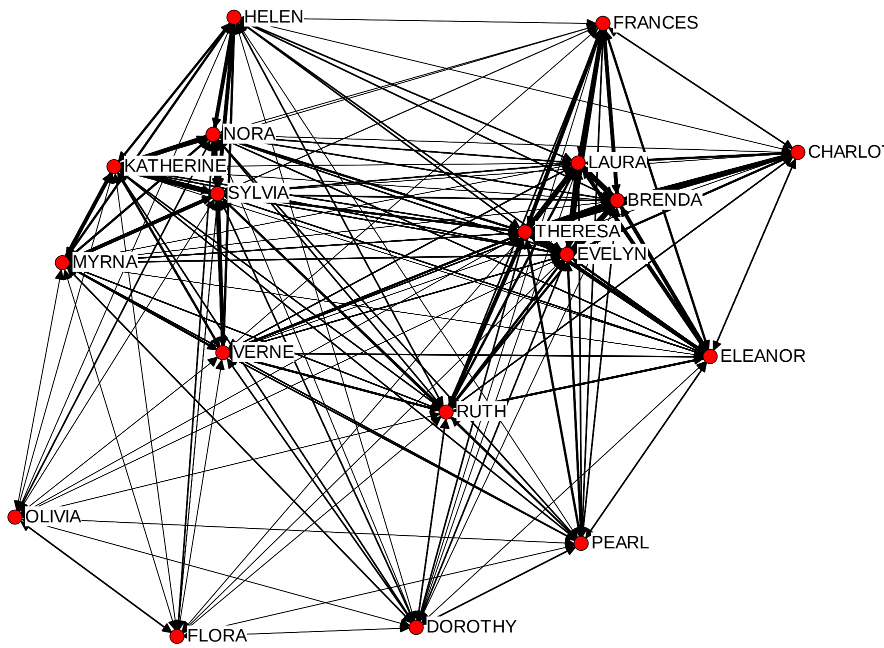


Figure xxx. Ordination+thickness. Both physical distance and thickness of line are used to represent strength of ties among women.

A different approach is to abandon ordination in favor of graph-theoretic layouts, in combination with dichotomizing the data so that only strong ties are considered. By systematically increasing the cutoff value for dichotomization, one can create a series of diagrams that portray increasingly strong ties (see Figures xx-yy).[2]

[2] To reproduce these diagrams using UCINET, open the affiliations matrix in NetDraw, then use the Ties window to raise the cutoff level by 1 unit. Then press the layout button (a lightning bolt with an equals sign). Repeat several times until no more ties are visible.

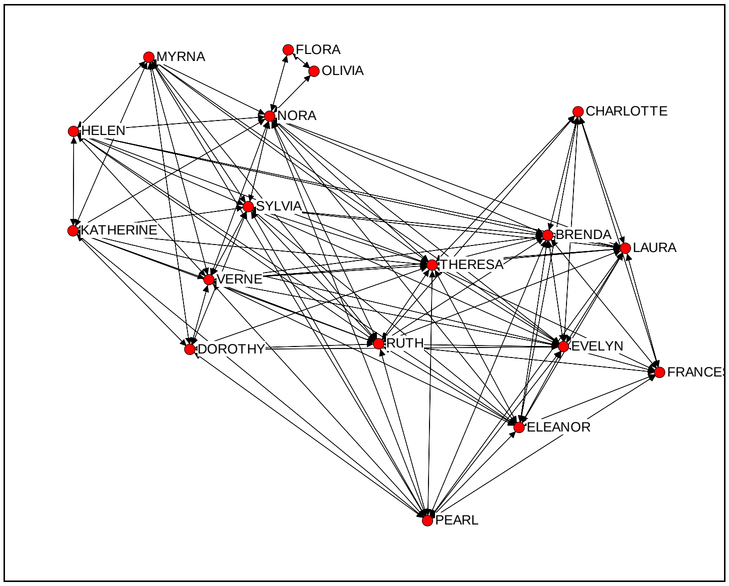


Figure xxxa. Ties of strength 1 or greater. Graph-theoretic layout.

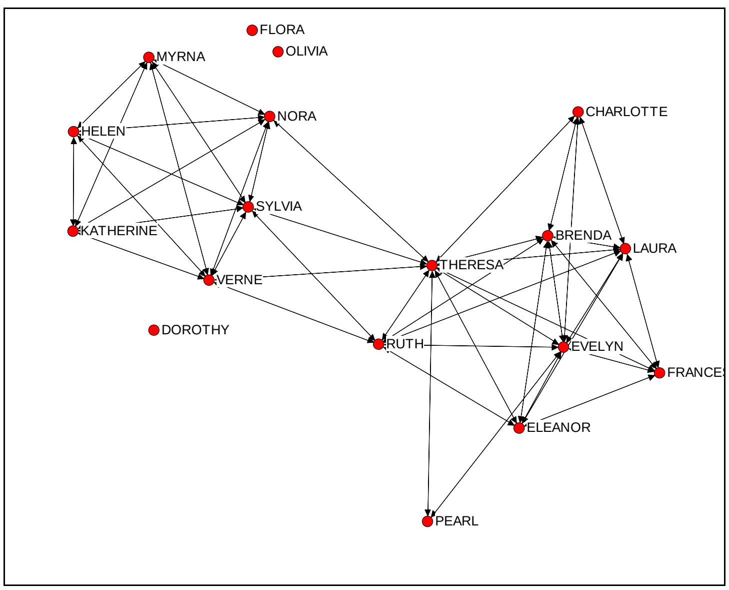


Figure xxxb. Ties of strength 2 or greater. Graph-theoretic layout.

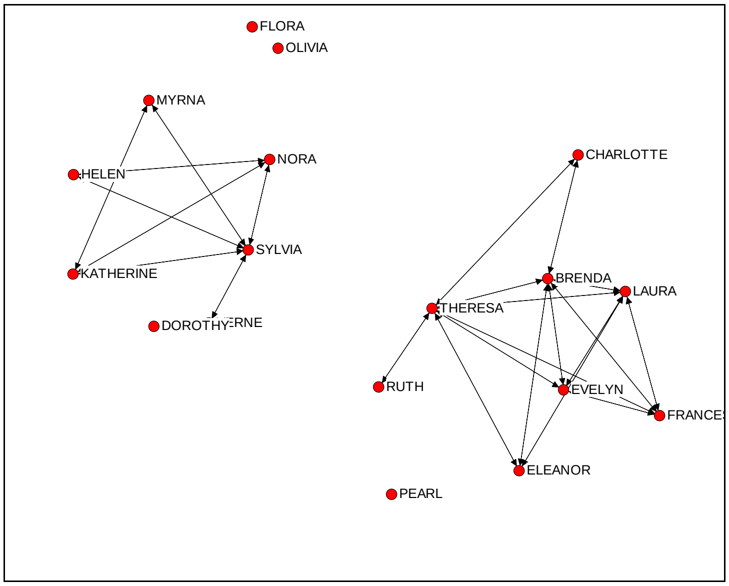


Figure xxxc. Ties of strength 3 or greater. Graph-theoretic layout.

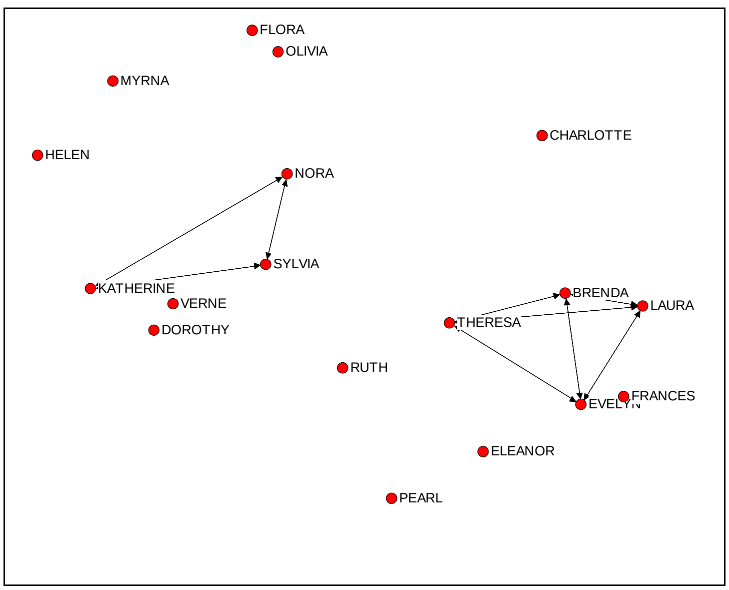


Figure xxxd. Ties of strength 4 or greater. Graph-theoretic layout.

Multiple Relations

In most studies, we have measured several different social relations on the same set of nodes. Programs like UCINET's Netdraw make it easy to switch between relations while maintaining nodes in the same positions.

For example, one of the standard UCINET datasets is the Bank Wiring room dataset. To collect these data, a researcher observed interactions among a set of employees in one room over a period of months, recording a number of social interactions such as playing games during breaks or having conflicts over such things as whether the room's windows should be open or closed. Figure xxx shows game playing ties among the men, while Figure xxx shows conflict ties. Since the nodes remain fixed in the same positions, it is easy to see that game playing interactions occur within each of the two subgroups, but rarely between, and conflict interactions occur mostly between the two subgroups but also within the right-hand subgroup. Thus, it appears that while two groups exist, the left-hand group is more cohesive.

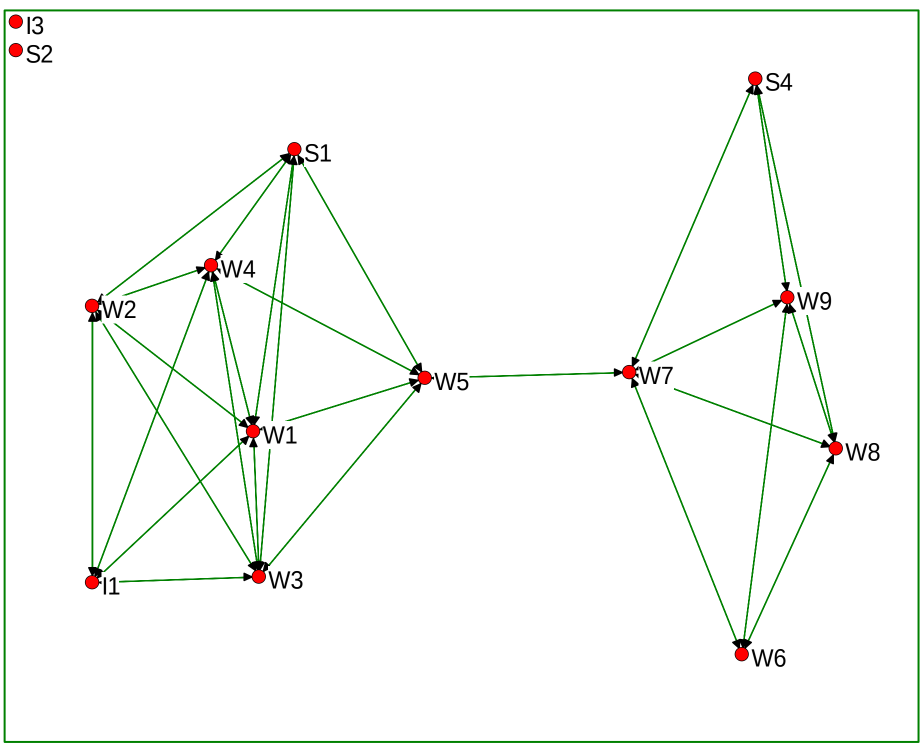


Figure xxx. Game-playing relation among Bank Wiring room employees.

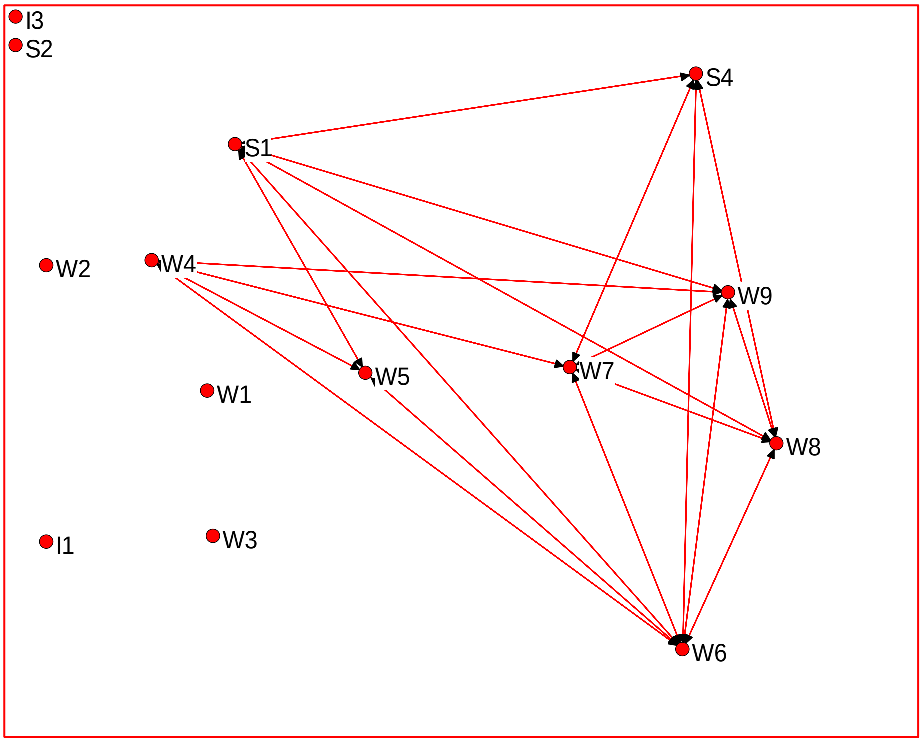


Figure xxx. Conflict relation among Bank Wiring room employees.

Changes over Time

If network data are collected at multiple points in time, we can just treat each time point as a different relation and use all of the techniques described above. We can also create a kind of meta-display by showing relationships between the time points rather than actors. For example, Burkhardt collected advice giving relations among employees of a government agency at five points in time. Using the techniques described in Chapter xx, we can correlate the adjacency matrices corresponding to each time period. Table xx shows the correlation matrix obtained for the Burkhardt data.

	T1	T2	T3	T4	T5
T1	1.000	0.684	0.483	0.440	0.300
T2	0.684	1.000	0.582	0.543	0.335
T3	0.483	0.582	1.000	0.613	0.341
T4	0.440	0.543	0.613	1.000	0.371

T5	0.300	0.335	0.341	0.371	1.000

As you might expect, the correlations with time 1 (first row) decrease from left to right, indicating that the social structure is increasingly different with each passing time period. In addition, the largest correlations for any time period are usually with the two periods on either side of it, indicating a kind of orderly change from period to period. However, the change is not linear. A metric multidimensional scaling of this correlation matrix (Figure xx) shows a gap between time 2 and time 3, and another gap between time 4 and time 5, suggesting periods of incremental change punctuated by instances of more radical change.

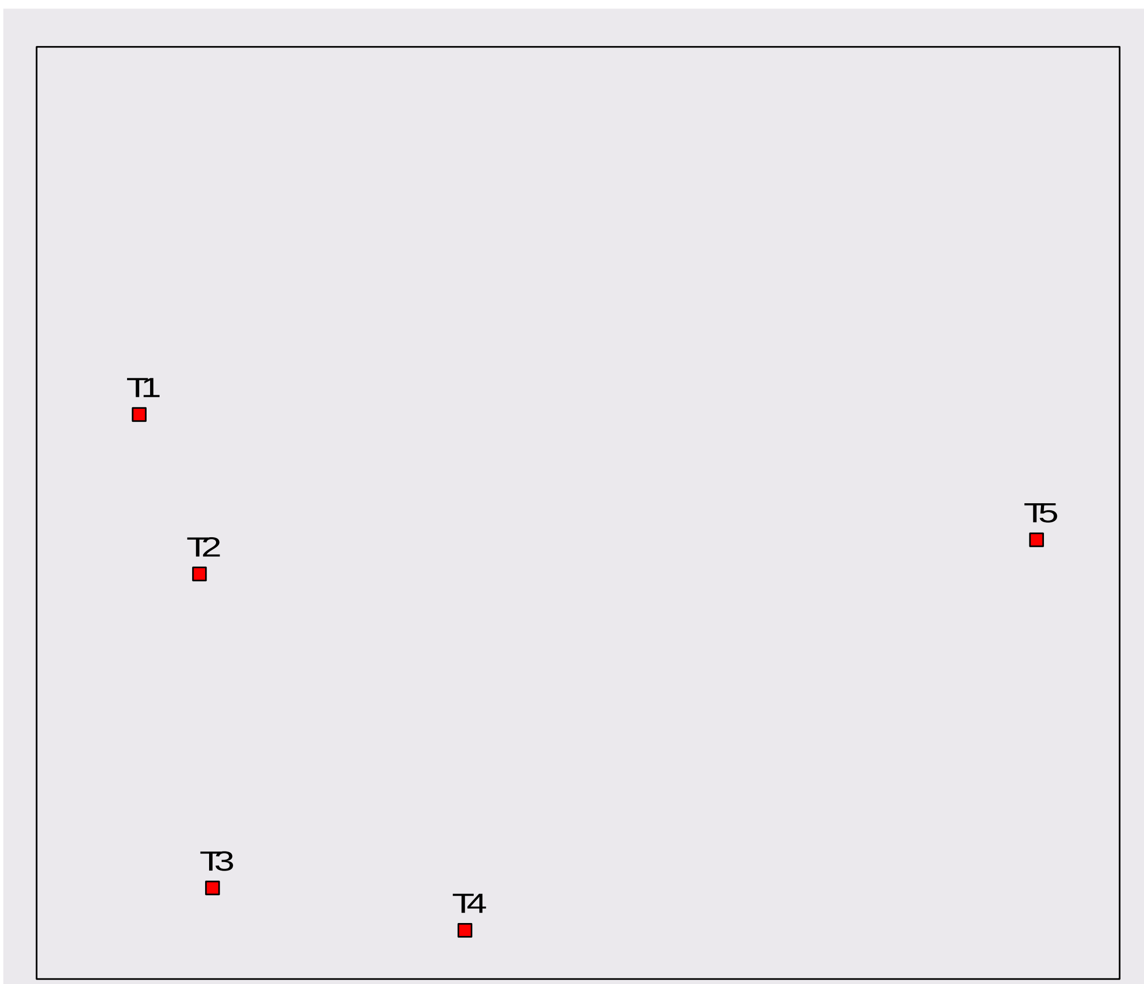


Figure xx. Correlations among time periods for the Burkhardt data, represented via metric multidimensional scaling.

To see how the network changed from Time 1 to Time 5, we can simply draw the two networks. Figures xx and yy show these two time points using a graph-theoretic layout and showing only strong ties. As we can see, at Time 1 the network shows evidence of three groups (left, bottom right and top right). At Time 2, the left and bottom right groups are still separate from each other, but the top left group seems to be in the process of being adopted by the other two groups. In addition we can see individual changes in position. For example, node R53 is a central figure in the bottom right group at Time 1, but becomes an isolate by Time 5.

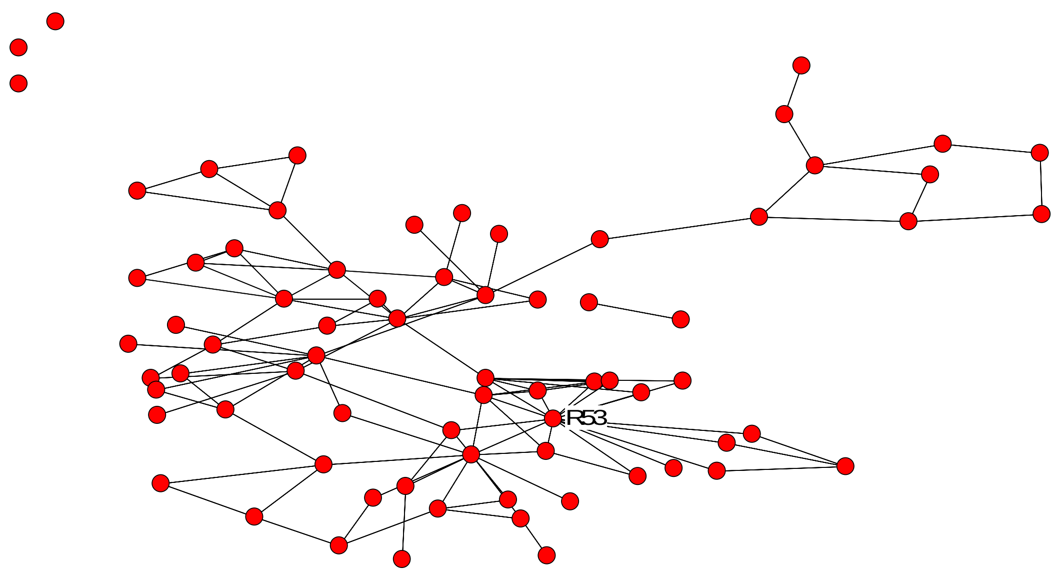


Figure xx. Friendship ties at Time 1 for the Burkhardt data.

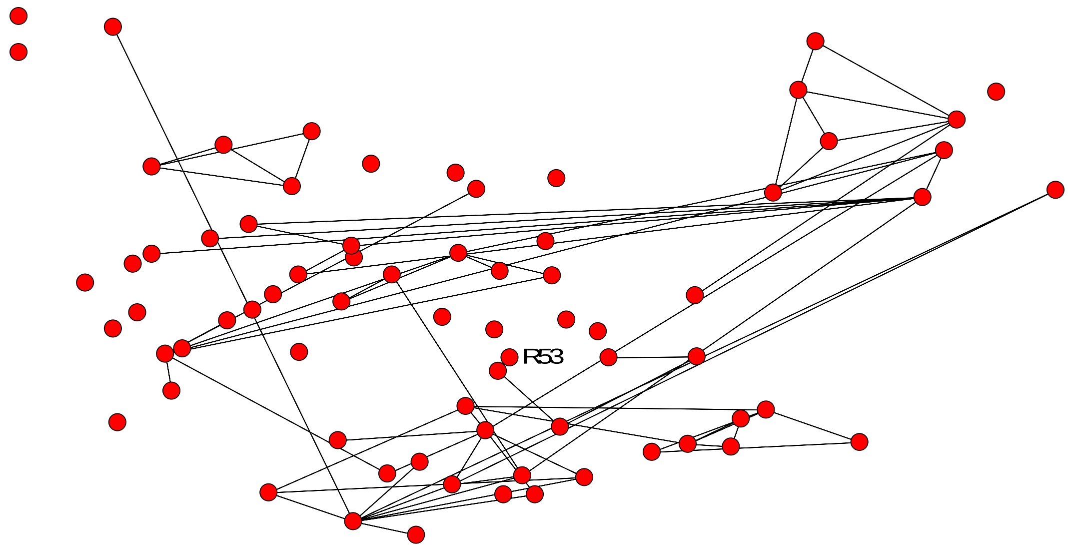


Figure yy. Friendship ties at Time 5 for the Burkhardt data.

Another way to visualize change over time is to focus on the change in an actor's position in the network over some period instead of concentrating on changes in the overall structure over time. This can be readily accomplished by stacking the network matrices on top of one another and then subject this stacked matrix to correspondence analysis. Figure zz shows how the matrices for 2 time periods can be stacked on top of one another. The data consists of ratings of reported interactions (on a scale from 0 to 10) at the beginning and end for people attending a workshop.

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
		--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
1	SUE	0	5	6	6	5	8	8	10	7	5	6	7	9	6	8	9	7
2	MARY	6	0	8	5	10	9	7	7	6	5	9	10	8	8	8	8	6
3	JOHN	8	3	0	8	8	8	5	9	9	5	9	6	9	5	9	6	9
4	BRYN	8	3	8	0	9	10	5	10	10	2	6	8	10	3	9	4	6
5	JEFF	3	9	7	5	0	9	4	4	4	3	8	5	4	2	4	2	3
6	STEVE	10	10	7	10	10	0	8	9	10	5	8	8	10	7	9	8	9
7	CHRISTINE	7	6	3	4	4	3	0	6	6	2	7	2	3	4	7	3	5
8	MARTIN	9	5	5	8	6	8	7	0	8	2	6	6	9	4	8	5	3
9	RICHARD	6	1	6	9	4	9	3	9	0	1	3	5	9	1	7	1	5
10	JEAN	2	1	2	2	2	1	2	2	2	0	2	2	2	1	2	2	1
11	ABBI	5	7	6	9	8	6	9	6	7	4	0	6	6	8	7	7	6
12	LISA	7	8	2	5	5	3	3	4	3	1	3	0	3	3	7	7	7
13	DAVID	9	10	10	10	10	10	8	10	8	1	8	8	0	6	9	5	6
14	LYNN	6	8	3	3	2	2	7	2	2	2	2	4	2	0	9	8	10
15	PETE	7	5	9	4	3	5	6	7	7	2	4	6	6	9	0	6	4
16	DELORES	7	5	7	5	5	6	4	5	6	2	6	8	6	5	8	0	4
17	SASHA	5	3	7	5	4	7	8	5	8	2	6	7	6	9	9	8	0
18	SUE2	0	6	8	8	8	10	8	10	9	5	7	9	10	9	10	9	9
19	MARY2	7	0	7	7	10	9	8	7	6	5	9	10	9	9	9	8	6
20	JOHN2	8	5	0	8	9	9	7	8	8	5	8	7	9	5	9	5	8
21	BRYN2	7	4	8	0	8	10	4	10	10	2	6	8	10	6	8	6	4
22	JEFF2	5	10	10	8	0	10	4	8	7	3	9	6	10	3	6	4	5
23	STEVE2	9	8	8	10	9	0	7	10	9	0	6	8	10	6	7	7	7
24	CHRISTINE2	1	6	1	4	4	3	0	4	4	0	6	2	3	6	2	2	3
25	MARTIN2	10	6	6	9	8	9	8	0	9	2	6	8	10	4	8	6	4
26	RICHARD2	10	4	10	9	7	10	4	10	0	2	7	10	10	5	9	7	9
27	JEAN2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	ABBI2	5	7	6	7	7	5	7	5	6	2	0	6	5	6	5	5	5
29	LISA2	8	8	8	5	8	8	5	8	8	1	5	0	7	1	8	5	5
30	DAVID2	7	8	8	9	8	10	6	9	9	1	6	7	0	2	8	5	5
31	LYNN2	3	10	5	4	3	4	8	5	3	2	9	5	3	0	9	10	10
32	PETE2	8	3	8	2	4	4	4	8	8	2	2	8	5	5	0	6	4
33	DELORES2	6	6	6	6	6	6	4	6	6	4	5	6	6	5	6	0	5
34	SASHA2	7	5	7	6	7	7	6	6	8	0	7	7	6	9	8	9	0

Figure zz. Matrices (n x n) for 2 time periods for a workshop group stacked on top of one another to form an n x m matrix. The ratings of interaction for time one are not shaded while the ratings of interaction fort time 2 are shaded.

The stacked matrix can be analyzed using correspondence analysis in order to visualize changes in the structural position of each of the actors in the network across multiple time periods. Figure zztop shows the changes in position for 2 time periods for reports of interaction for the workshop group. The figure reveals a tendency for members of the group to move closer to one another over time. The group appears to be coming more cohesive and in fact density does increase between Time 1 (5.90) to Time 2 (6.13).   Lisa, in particular, makes a movement from the groups periphery to its' core followed to a smaller extent by Lynn.

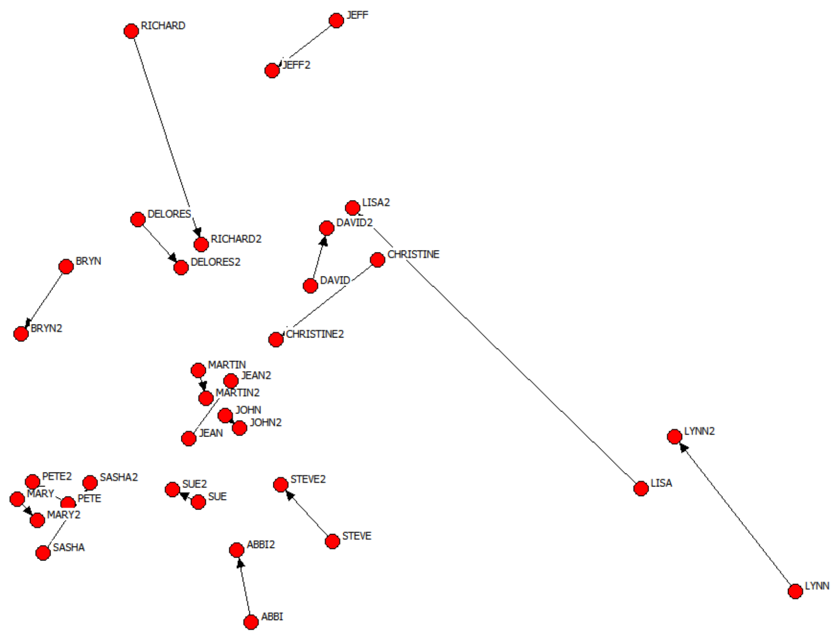


Figure zztop. A graph of the changes in network position between 2 time periods using correspondence analysis.

Multiple time points can also be visualized in that multiple matrices representing time points can be stacked and then visualized using correspondence analysis. Figure zztop2 is a graph showing vote co-occurrences for Supreme Court Justices by year over a 10-year period. The graph shows the spatial location of each of the justices in each of the years. Figurezztop3 shows the spatial movement for Rehnquist over the course of the 10 years. What this graph clearly reveals is that Rehnquist himself has often entered what one might think of as swing vote spatial territory in the course of his voting behavior. In Figure zztop3 voting blocks (i.e., conservative, swing, liberal), as identified by media sources, are encompassed by convex hulls. Here the the extreme edges of each of the blocks can be easily determined. For example, Scalia over the 10-year period consistently defines the extremes of the conservatives while Stevens consistently defines the extremes of the liberals. Although Kennedy and O’Connor were considered swing votes there is a definite bias towards the conservative side of the graph.

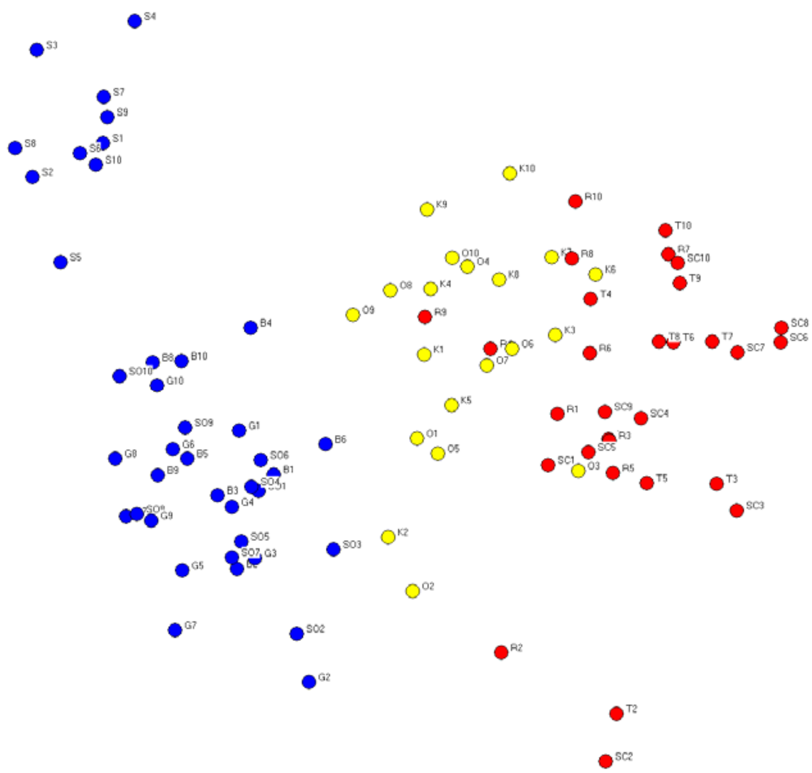


Figure zztop2. Stacked correspondence analysis of the co-occurrence matrices for the 10 years of voting behavior among the Supreme Court Justices.

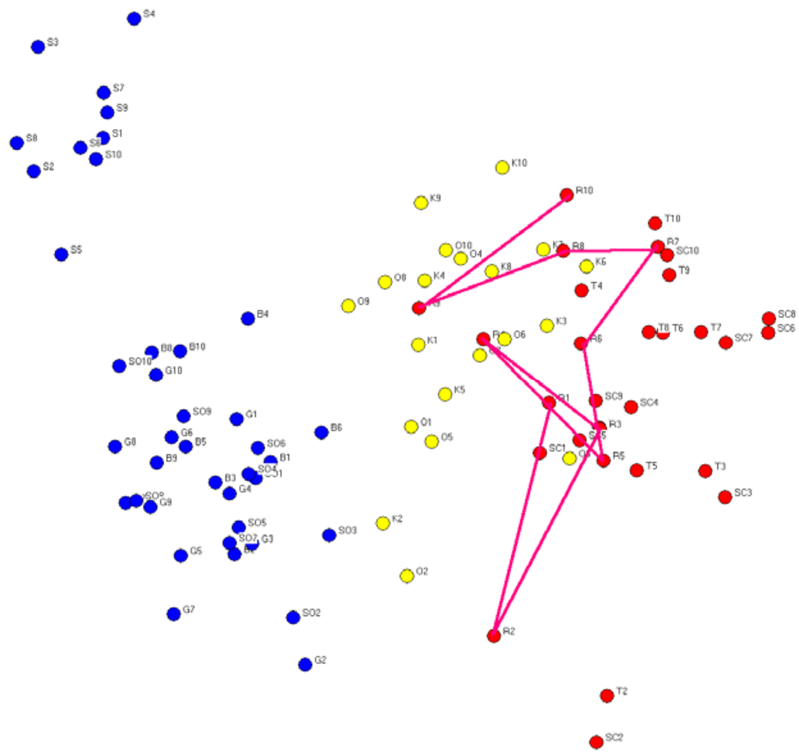
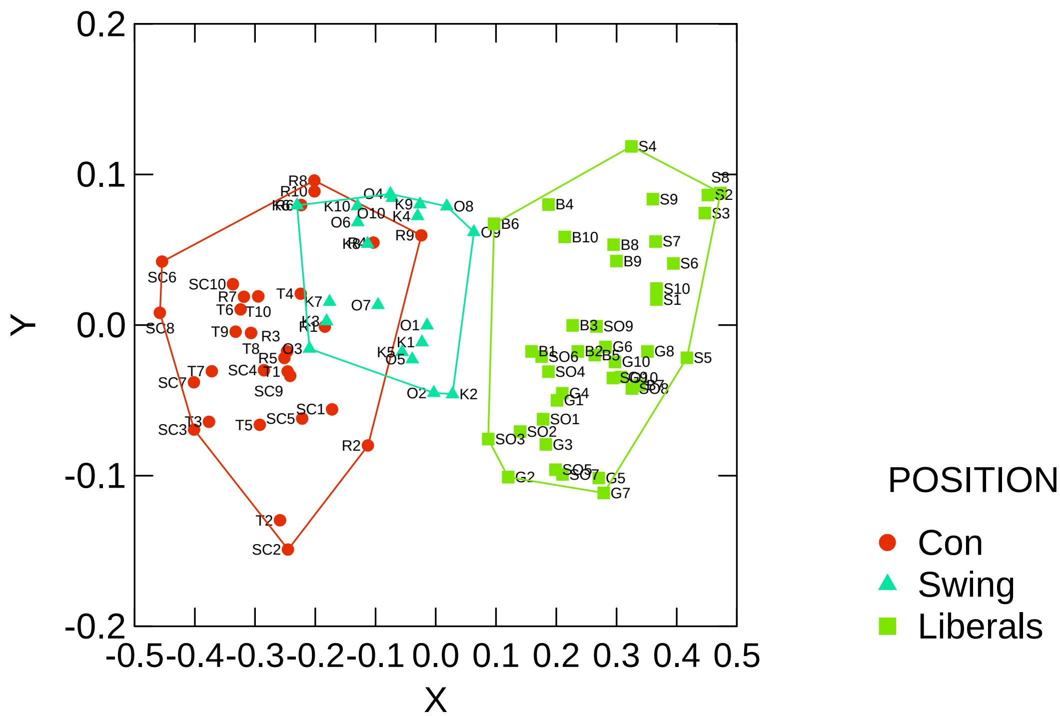


Figure zztop3. Stacked correspondence analysis of the co-occurrence matrices for the 10 years of voting behavior among the Supreme Court Justices with Renquists' spatial movements connected over time.



Figurezztop4. Stacked correspondence analysis of all time periods with conservative, swing and liberal judges identified as determined from media sources with members encompassed by convex hulls.

Visualizing Large Networks

Visualizing large networks is a skill in its own right. Clearly one approach is to apply the reduction techniques discussed above and then visualize either fragments of the network or aggregated versions of the network. The nodes that make up the aggregated network can then visualized separately. The software program Pajek has been specifically designed to visualize large networks and has a number of features which support this kind of approach.

If the network is well structured it is still possible to produce revealing visualization which help in the analysis of large networks. In this case we are not so concerned with individual actors but with the overall pattern. This is analogous to looking at a galaxy where we are not concerned about the position of individual stars but by the overall spiral pattern that they exhibit together.

The general approach is to try and peel away actors (and if valued edges) which detract from the underlying structure. In the first instance since we are not concerned with the individual actors any display should not include labels. We now demonstrate the general approach using the software package UCINET. Figure 12.2 is a network of just under

1000 edges are valued between 1 and 28. The labels have been removed and the nodes are randomly arranged.[that's cheating]

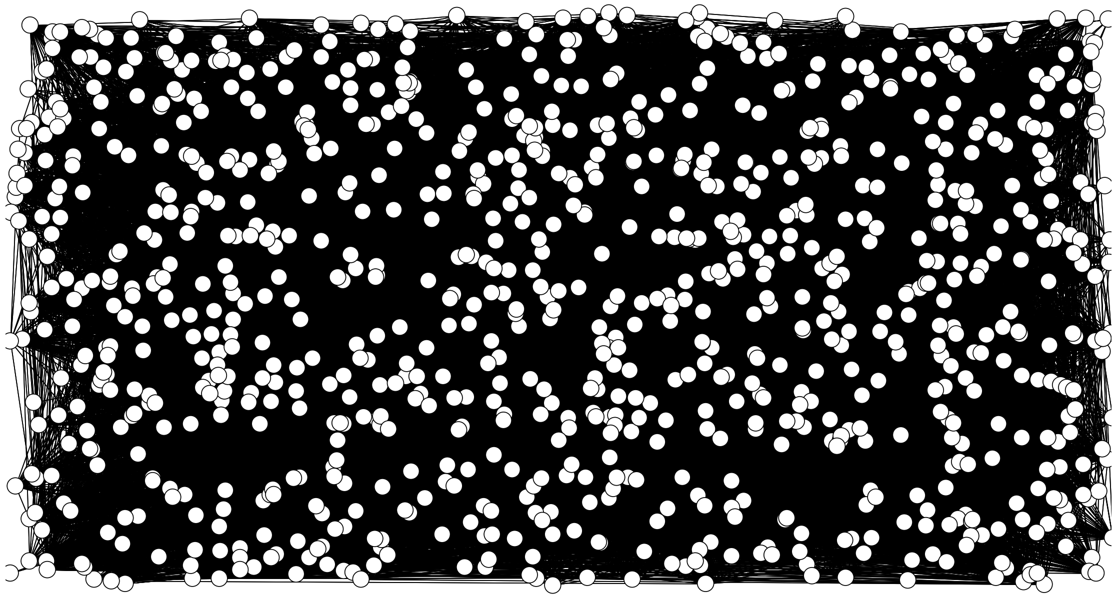


Figure 12.2

This is clearly unsatisfactory and of no use in helping understand the data. As a first step we reduced the number of edges by dichotomizing at a value of 6. The components of this network were then computed. There were now a large number of isolates and small components. The largest component had 203 actors, while the next largest had just 11. The original network consisted of one component so all the smaller components are linked into the large component at some level. We can therefore conclude that the large component is highly representative of the general underlying structure. This component can be seen in Figure 12.3.

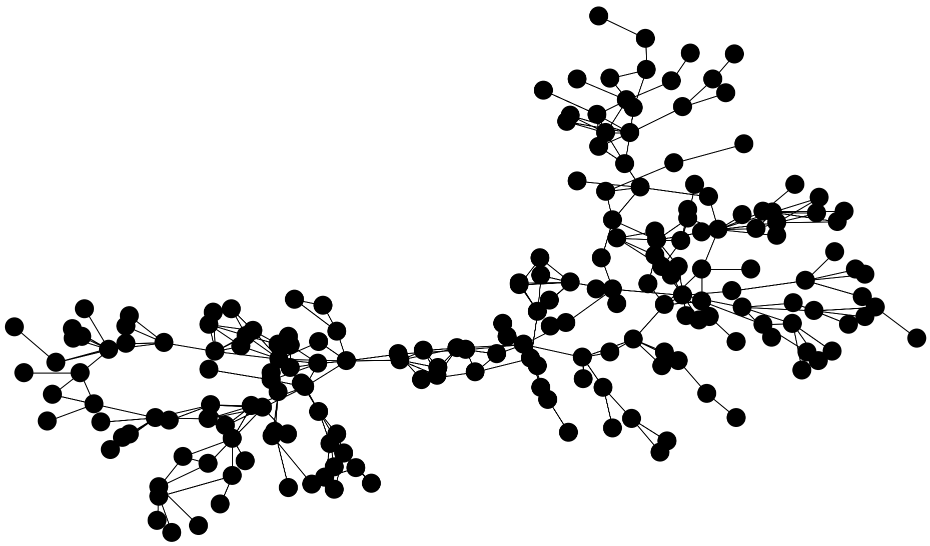


Figure 12.3

We are now able to see some of the underlying structure of the network. There are two main groups and a third smaller group (at the top) that is forming at a small distance from the larger groups.