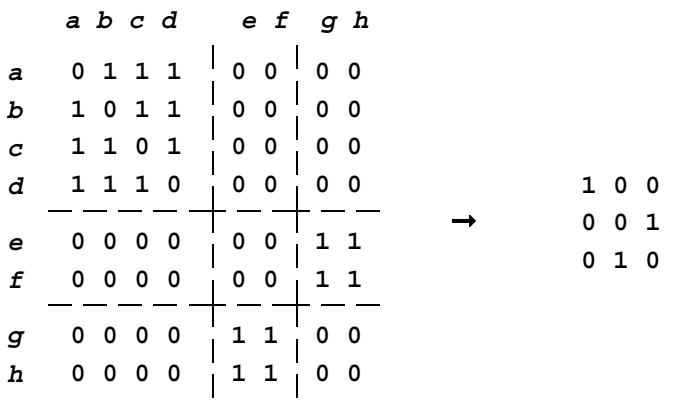
Introduction

Social network analysis (SNA) begins with data that describe the set of relationships among the members of a system. One goal of analysis is to obtain from the low-level relational data a higher-level description of the structure of the system which identifies various kinds of patterns in the set of relationships. For example, it may be of interest to find cohesive clusters of network members: those which have most of their connections with each other. It may also be of interest to find members with similar roles: those with few mutual connections but many connections to other similar sets of members. These two goals may be combined in the search for general patterns, which is the aim of block-modeling (Wasserman and Faust, 1994).

As an illustration of block-modeling, consider a network or graph G(V,E) as a set of nodes V (points, vertices) connected by a set of links E (lines, edges). For simplicity here, we will consider networks that are binary (edges have logical value 1 if a relationship/connection between the nodes exists, 0 if not), symmetric (an edge from node *I* to *j* implies an edge from node *j* to *I*), and without self-loops (no edges between I and I). We may represent such a network as the (square) adjacency matrix **A** = **A**(G) with:

A(*i*,*j*) = 1 if *i* is connected to *j*
A(*i*,*j*) = 0 otherwise

For example:



where the partitions of the network on the left map onto the blocks on the right. It is easy to see where the partitions (and hence blocks) should go in this example, since the rows and columns are ordered to make this obvious. In general, network data is not so conveniently ordered, nor is it so obvious where the blocks are. In this example, we see:

- the network is not connected; there are no links from the block in the upper left (*a*-*d*) to those in the lower right (*e*-*h*);
- the upper left block is on the diagonal; it is a clique (complete graph), with a link between every pair of nodes;
- the lower right blocks are off-diagonal and form a (complete) bipartite graph, with links from *e* and *f* to *g* and *h*, but no links between *e* and *f* or *g* and *h*.

There are a number of methods for finding an ordering and a blocking of network data. One approach is to choose a set of axes in the multidimensional space occupied by the network and rotate them so that the first axis points in the direction of the greatest variability in the data; the second axis, orthogonal to the first, points in the direction of greatest remaining variability, and so on. This set of axes is a coordinate system that can be used to describe the relative positions of the set of points in the data. Most of the variability in the locations of points will be accounted for by the first few dimensions of this coordinate system. The coordinates of the points along each axis will be an eigenvector, and the length of the projection will be an eigenvalue. The set of all eigenvalues is the spectrum of the network. Spectral methods (eigendecomposition) have been a part of graph theory for over a century. SNA researchers have used spectral methods either implicitly or explicitly since the late 1960's, when computers became generally accessible in most universities. Two of the earliest important programs were related to eigendecomposition: Negopy (Richards, 1971; Richards and Rice, 1981) was designed for finding cohesive clusters, and CONCOR (Breiger *et al*., 1975) aimed to solve the more general block-modeling problem. The eigenvalues of a network are intimately connected to important topological features such as maximum distance across the network (diameter), presence of cohesive clusters, long paths and bottlenecks, bipartite-ness, and how random the network is. The associated eigenvectors can be used as a natural coordinate system for graph visualization; they also provide methods for discovering clusters and other local features. For a more complete discussion of these matters, see Seary and Richards (2003).

As well as networks of people and relationships, SNA has long considered relationships between people and events (Davis *et al*., 1941), co-authorship networks (Crane, 1972), and other examples of so-called 2-mode networks (Wasserman and Faust, 1994) which involve relationships between two types of nodes. These networks are usually shown as rectangular **R** (with n_1 rows and n_2 columns), since in general there are not the same numbers of the two types of nodes. As Breiger (1974) shows, 2-mode matrices can be made square by matrix multiplication of **R** and its transpose $\mathbf{R}^T$. Another approach is to make a square **A** (with n_1+n_2 rows and columns) from **R** by appending $\mathbf{R}^T$ below and to the left of **R** along with necessary **0** matrices:

R =

	1	2	3	...	n2
1	1	1	0	...	1
2	0	1	1	...	1
:	:	:	:	...	:
n1	1	0	1	...	0

A =

0	**R**
$\mathbf{R}^T$	**0**

=

	$1\ 2\ \ldots\ n_1$	$1\ 2\ 3\ \ldots\ n_2$
1		1 1 0 ... 1
2	0	0 1 1 ... 1
:		: : : ... :
n_1		1 0 1 ... 0
1	1 0 ... 1	
2	1 1 ... 0	
3	0 1 ... 1	0
:	: : ... :	
n_2	1 1 ... 0	

This shows that 2-mode networks can be represented by the square adjacency matrices of symmetric bipartite graphs. (We will use this method for the data we describe later in this paper.) This representation is not generally used in SNA, probably because of the extra space taken up by the transpose and the **0** matrices. However, sparse matrix methods, which only store and

manipulate actual links, can allow rectangular **R** to be treated as square **A** very efficiently (Seary, 2005, p189).

The Normal Spectrum

The Normal Spectrum may be derived by considering the generalized quadratic placement problem (Hall, 1970; Seary and Richards, 1995) leading to the generalized eigenvalue equation:

$$\mathbf{Lx} = \lambda\, \mathbf{Dx} \text{ , where}^{1}$$

- **D** is a diagonal matrix of node degrees of G
- **L** = **D** - **A** is the Laplacian matrix of G (Cvetkovic *et al.*, 1995, Seary and Richards, 2003)
- λ is an eigenvalue
- **x** is a corresponding (column-)eigenvector

Assuming that **D** can be inverted (which it can be if every node has at least one link; i.e. no nodes are isolated)

$$\mathbf{D}^{-1}\,\mathbf{LX} = \mathbf{D}^{-1}\,(\mathbf{D} - \mathbf{A})\mathbf{X} = (\mathbf{I} - \mathbf{D}^{-1}\,\mathbf{A})\mathbf{X} = \mu\,\mathbf{X}$$

where **A** is the adjacency matrix of G, and **I** is an identity matrix of proper size. In fact, we usually take the defining equation to be

$$\mathbf{D}^{-1}\,\mathbf{An} = \mathbf{Nn} = \nu\mathbf{n} \quad \text{with} \quad \mathbf{D}^{-1}\,\mathbf{A} = \mathbf{N} \quad \text{and} \quad \nu = 1-\lambda, \text{ where}$$

- ν (the Greek letter nu) is an eigenvalue of the Normal matrix **N** and
- **n** is the corresponding eigenvector.

Adding an identity matrix shifts the eigenvalues by 1 without changing the eigenvectors. Note that for networks without isolated nodes **D** has an inverse and therefore an inverse square root $\mathbf{D}^{-1/2}$. In networks with isolated nodes, the network size is effectively reduced by the number of isolates because the analysis uses only the nodes with links. The number of eigenpairs (ν_i , $\mathbf{n}_i$) is equal to the number of nodes n. We generally label these with i=0,...,n-1 since i=0 corresponds to the trivial eigenpair ($\nu_0 = 1$, $\mathbf{n}_0 = \mathbf{1}$).

The Normal matrix **N**(G) is:

$$\mathbf{N}(i,j) = 1/\deg(i) \text{ if } i \text{ is connected to } j$$

$$\mathbf{N}(i,j) = 0 \text{ otherwise}$$

so that **N** is not symmetric. However, we can construct $\mathbf{M} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$, which is symmetric, and which is similar to **N** (it has the same eigenvalues).

---

[1] We have introduced some notation which will be followed throughout:

- matrices are represented by bold capitals: **D**
- (column-)vectors are represented by bold lower case: **e**
- eigenvalues are represented by greek letters, usually with some relationship to the latin letters representing a matrix and an eigenvector. E.g.(ν_i , $\mathbf{n}_i$) are the eigenpairs of Normal matrix **N**
- a boldface **1** refers to the vector (1,1, ... ,1)

Let (ν_i , $\mathbf{m}_i$) be the eigenpairs of **M**. Then the eigenpairs **N** are:

$$(\nu_i , \mathbf{n}_i) = (\nu_i , \mathbf{D}^{-1/2}\mathbf{m}_i)$$

The orthonormalization condition is:

$$\mathbf{n}_i \mathbf{D} \mathbf{n}_j = \delta_{ij} = 1 \text{ if } i=j, 0 \text{ otherwise}$$

That is, the vectors are orthonormal in the **D** (or χ^2) metric (Richards and Seary, 1997). The Normal spectrum is referred to as the Q-spectrum in (Cvetkovic, *et al.*, 1995). The multiplicity of 1 as an eigenvalue is equal to the number of connected components in G. If G is bipartite, then eigenvalues appear as pairs with opposite signs. Thus -1 is an eigenvalue if and only if G is bipartite.

The Normal matrix **N** has a number of interesting properties:

- It has a trivial constant eigenvector $\mathbf{n}_0 = \mathbf{1}$ with eigenvalue $\nu_0 = 1$
- The spectrum of **N** is bounded by $1 = \nu_0 \geq \nu_1 \ldots \geq \nu_{n-1} \geq -1$
- The rows of **N** sum to 1 (it is a stochastic matrix)
- The spectrum of **N** contains a 1 for every connected component
- The eigenvalue -1 occurs if and only if G is bipartite, in which case all eigenvalues occur in pairs with opposite signs
- **N** has been rediscovered a number of times: generalized or combinatorial Laplacian (Dodziuk and Kendall, 1985; Chung, 1995); Q-spectrum (Cvetkovic, *et al.*,1995).

Notice also that the similar matrix **M** satisfies the definition of Chi-squared. In practice, it is much simpler to solve the eigenproblem for **M**, since it is symmetric.

Four important properties of Normal eigenpairs

The following important properties of Normal eigenpairs will be useful in understanding the results obtained later.

1. Bipartite Representation of 2-mode networks

We can represent a 2-mode network by a square symmetric matrix with all the links in off-diagonal corners, so that the matrix is mostly 0's. The result is always a bipartite graph, so that all eigenvalues occur as positive and negative pairs (eg. 1, -1, 0.93, -0.93, ...). Generally we don't need most of the negative eigenpairs, but the eigenvector belonging to eigenvalue -1 can be very useful. We don't need to explicitly construct the full matrix, nor calculate all eigenpairs. Using sparse methods and automatic symmetrization, we only need store the links in one direction, and can calculate only a few eigenpairs with largest eigenvalues (Seary, 2005, p189).

Assume that there are n_1 items in one mode (the rows of the original matrix) and n_2 items in the other mode (the columns). Then the bipartite matrix will be square with (n_1+n_2) rows and columns. Thus each eigenvector also has (n_1+n_2) coordinates. By the bipartite construction, the

first n_1 coordinates correspond to the n_1 items in the first mode (the rows), and the remaining n_2 coordinates correspond to the n_2 items in the second mode (the columns).

For a pair of positive and negative eigenvalues of a normal spectrum, the only difference between the corresponding eigenvectors is that the first n_1 coordinates of one have opposite signs of the first n_1 coordinates of the other. In particular, the eigenvector belonging to eigenvalue -1 is the trivial constant eigenvector (**1**), except that the first n_1 coordinates are negative. The difference in signs can be used to identify the two modes.

2. Visualization

The eigenvectors of **N** can provide good visual representations of graphs which consist of blocks of nodes with similar connections. This follows from the relationship between the eigenvector coordinates for a node and those it is connected to (Seary and Richards, 1998). It is evident from the definition of eigendecomposition that:

$$n_i(s) = \frac{\sum_{s\sim t} n_i(t)}{(\nu_i \times \deg(s))} \qquad (1)$$

for the i^{th} eigenpair (ν_i, $\mathbf{n}_i$) of **N** (where $n_i(s)$ is the s^{th} component of the i^{th} eigenvector; "$s\sim t$" means "s is connected to t")

This equation shows that for eigenvalues ν_i near 1, each node is approximately at the centroid of those it is connected to. The exact difference from the centroid for node u for eigenvector $\mathbf{n_i}$ is:

$$n_i(s) - \sum_{s\sim t} n_i(t) \Big/ \deg(s) = (1-\nu_i) n_i$$

For "important" eigenvalues ν_i near 1, this produces very good visualization properties. Members of a block tend to be close to one another and not close to members of other blocks.

3. Relation to χ^2:

The χ^2 matrix is defined in terms of the row and column *marginals* (sums). A typical element is $(\text{Observed}_{ij} - \text{Expected}_{ij})^2 / \text{Expected}_{ij}$ where

$$\text{Expected}_{ij} = \frac{\deg(i)\deg(j)}{\sum \deg(i)}$$

We can write χ as $O/\sqrt{E} - \sqrt{E}$, where the second term corresponds to the trivial eigenvector which can be dealt with separately. In matrix notation $\chi = \mathbf{D}^{-1/2}\mathbf{A}\,\mathbf{D}^{-1/2}$ which has eigenpairs (ν_i, $\mathbf{D}^{-1/2}\mathbf{n}_i$). Thus (omitting the expected term corresponding to trivial $\nu_0 = 1$, $\mathbf{n}_0 = \mathbf{1}$) we have[2] the following:

$$\chi^2 = \sum_{j=1}^{n-1} \nu_i^2 \sum_{i,j=1}^{n} a_{ij}$$

[2] For the bipartite representation, only the positive eigenvalues contribute to χ^2 of the 2-mode network.

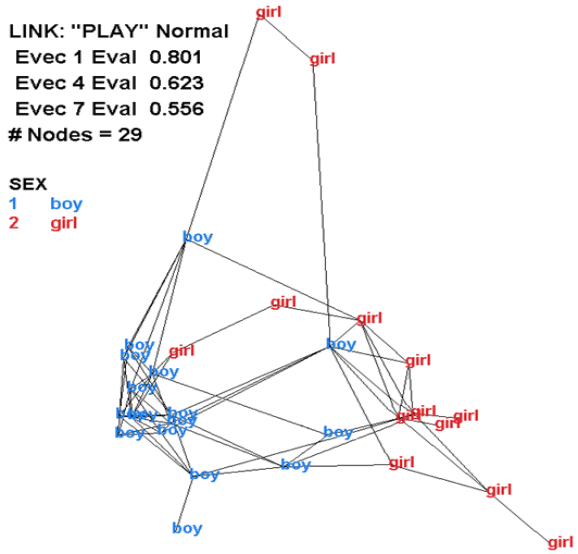
This equation shows how much each dimension contributes to χ^2 which is a measure of *dependence* between rows and columns (or of deviation from what would be expected if the node degrees by themselves would give a complete description of the network's structure). In this interpretation, if $|\nu_1|$ is small ($|\nu_1| << \nu_0 = 1$), then χ^2 is also small: there is no structure or pattern to explain in the network beyond the node degrees, and so there is no "signal" above the expected "background." On the other hand, if $|\nu_1|$ is close to 1, then χ^2 will be large and there is a relation between rows and columns of **A**, with the first eigenvector pointing in the direction of the maximum variability in χ^2. If $|\nu_2, \nu_3, ...\nu_k|$ are also large, we need k+1 eigenvectors to describe the patterns in the χ^2 matrix. Thus we can tell from the eigenvalues how many eigenvectors we need to explain most of the χ^2 of the network, and which are the most "important" ones, since they contribute most to χ^2 (Greenacre, 1984).

4. Partitions

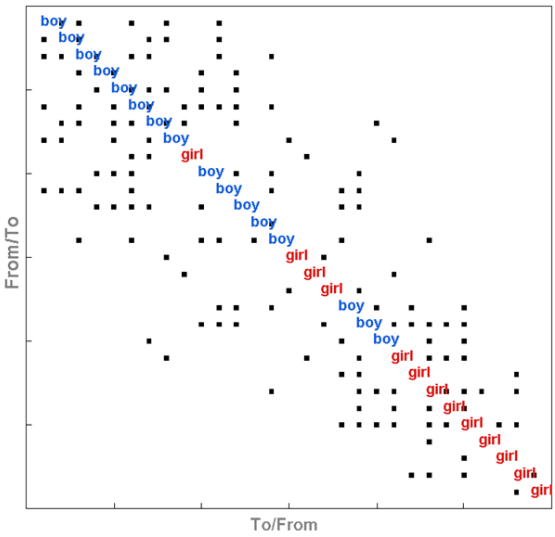
There is a large body of literature on the use of eigenvector coordinates to provide partitions of graphs. Most of these methods use eigenpairs of the adjacency matrix (Powers, 1988) or the Laplacian (Pothen *et al.*, 1990). Fiedler (1975) was the first to show that Laplacian eigenpairs could provide good approximate solutions to the min-cut problem: partition a graph into parts with approximately equal numbers of nodes and few links between them. We can add an additional constraint that the *number of links* in each part also be roughly equal by weighting the node sets by their total degrees (Dhillon, 2001). This is exactly what a partition based on $\mathbf{n}_1$ from **N** gives us, since $\mathbf{n}_1$ points in the direction of maximum variability in χ^2. Similarly, further partitions based on $\mathbf{n}_2$, $\mathbf{n}_3$, ... will also produce sets of nodes with a large number of edges in common (as long as ν_2, ν_3, ... make significant contributions to χ^2). Partitions based on positive eigenvalues will produce blocks of edges on the diagonal of **A**, while those based on negative eigenvalues produce nearly bipartite off-diagonal blocks, which occur in pairs if the network is symmetric (Seary and Richards, 1995). In both cases, the concentration of links to specific parts of the network leads to a large value of χ^2 for the partition.

As an example of these properties, figure 1 shows visualizations of children at a day-care centre[3]. The network is defined by observing which children "Play" with each other (all links are therefore symmetric). Figure 1a is a two-dimensional visualization, labelled by the sex of the children. It is clear that the x-dimension (eigenvector 1) is important (eigenvalue = 0.801) and that the clusters on the left and right are related to sex. Figure 1b is a one-dimensional visualization, showing the adjacency matrix as permuted by the coordinates on eigenvector 1. It is clear that this permutation based on the maximum variability in χ^2 has moved most of the links close to the diagonal. Figure 1c shows the same adjacency matrix as permuted by the sex of the children (boys in upper left, girls in lower right). Some clustering is evident. Figure 1d shows the adjacency matrix permuted by the *signs* of eigenvector 1 ("n" for negative, "p" for positive). The partition, which is now based on the network itself, is better than that for sex in a sense that will be described in the next section.

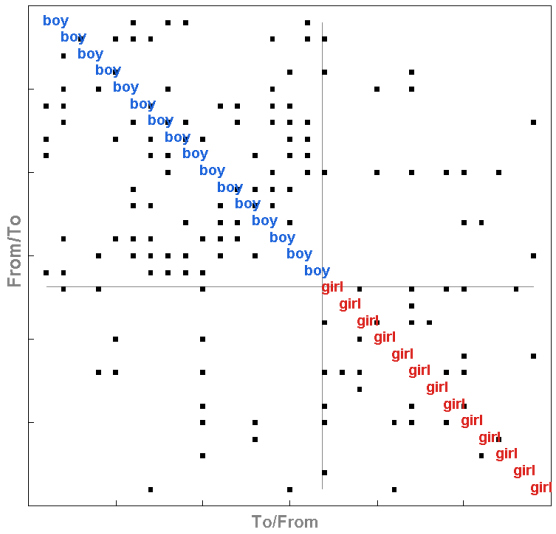
[3] These data were collected by students in a course Richards taught taught in 1988. The students watched the children in a daycare centre (ages: 6 to 10) and, over the course of a day, noted the children they saw playing together and, later in the day, asked each who they had played with.



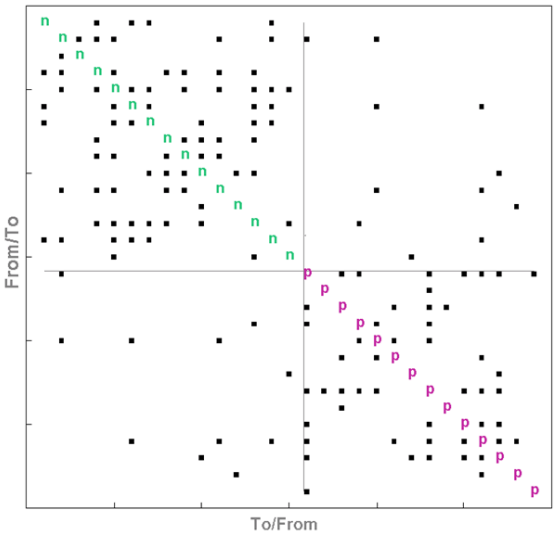
a) Two-dimensional visualization of Play network based on positive Normal eigenvectors. Nodes labelled by Sex.



b) Play adjacency matrix permuted by Normal eigenvector 1 coordinates. Nodes labelled by Sex.



c) Play adjacency matrix permuted by Sex. Nodes labelled by Sex.



d) Play adjacency matrix permuted by signs of Normal eigenvector 1. Nodes labelled by signs.

Figure 1. Four visualizations of the Play network.

Contingency tables and panigrams

Once a partition has been found, it is a simple matter to form a contingency table by counting the number of links within and between each block. The quality of partitions can also be compared by calculating the χ^2 for each contingency table. These tables may be visualized by using *panigrams*[4]. A panigram contains most of the information in a contingency table. The

[4] Seary suggested the name "panigram" for the two-dimensional analog of histograms. "Histos" (ιστοσ) is Greek for the mast of a ship, whereas a 2-dimensional sail is "pani" (πανι) in Greek.

percentages in the left column are the row marginals. The percentages under the columns are the column marginals. The numbers in the cells are the column percents you would see in the corresponding cells of the contingency table. The height of the segments in the "Totals" column are proportional to the row marginals. The width of the other columns are proportional to the column marginals. Thus the areas of the segments are proportional to the percent in the corresponding cell in the contingency table. If the row variable is independent of the column variable, the segment heights in all columns are the same as the ones in the "Totals" column. This is not the case if the row variable is not independent of the column variable.

In table 1a and figure 2a (compare to figure 1c) we see the counts within and between a block-model based on sex. In table 1b and figure 2b (compare to figure 1d) we see the counts within and between a block-model based on the component signs of the first normal eigenvector. Clearly the latter is superior based both on a larger χ^2 and more within-block and fewer between-block counts.

Table 1. Contingency tables for partitions based on: a) node attribute and b) Normal eigenvector

a) Crosstabulation of sex.
Chi-squared = 44.613

COUNT
ROW %
COL %

ROWS = FROM sex
COLS = TO sex

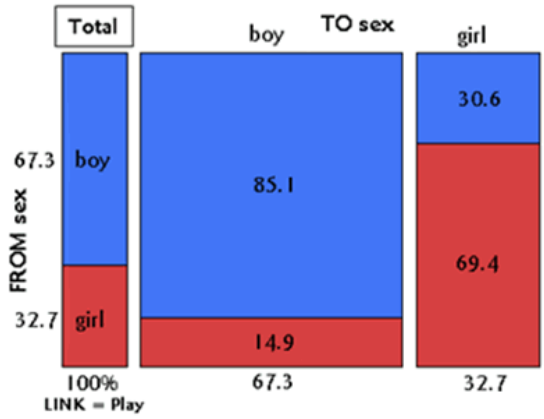
	boy	**girl**	**TOTAL**
boy	86 85.15% 85.15%	15 14.85% 30.61%	101 67.33%
girl	15 30.61% 14.85%	34 69.39% 69.39%	49 32.67%
TOTAL	101 67.33%	49 32.67%	150

b) Crosstabulation of np(Play)
Chi-squared = 73.282

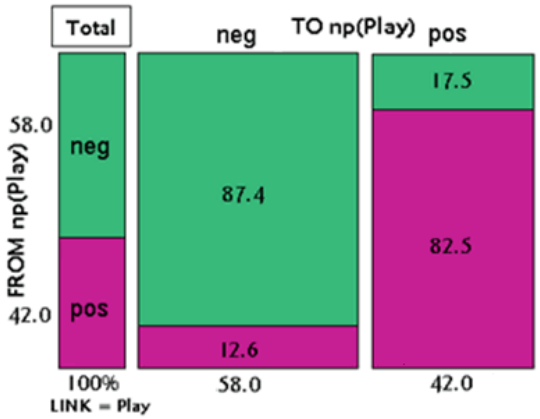
COUNT
ROW %
COL %

ROWS = FROM np(Play)
COLS = TO np(Play)

	neg	**pos**	**TOTAL**
neg	76 87.36% 87.36%	11 12.64% 17.46%	87 58.0%
pos	11 17.46% 12.64%	52 82.54% 82.54%	63 42.0%
TOTAL	87 58.0%	63 42.0%	150



a) Panigram for partition of Play network based on Sex. χ^2 = 44.613



b) Panigram for partition of Play network based on signs of Normal eigenvector 1. χ^2 = 73.282

Figure 2. Panigrams for Play network with partitions based on a) node attribute and b) Normal eigenvector 1 interactive Fig2

The MultiNet Network analysis program

MultiNet is a Windows-based computer program designed for interactive exploratory data analysis of social and other large, sparse, multivariate networks[5]. It was designed for exploratory analysis and visualization of large, complex networks, and to provide details of the values of the link and node variables that make up the networks. Three aspects of the program are relevant to this discussion:

- **Eigenspaces**: Visualize networks and create variables and partitions from graph spectra.
- **Variables**: Univariate statistics and transform, combine, create and delete link or node variables. We will make use of the **Recode** function which allows a variable to be created by combining existing variables, then transformed into a categorical variable by quantiling, for use in a contingency table.
- **Analyse**: Perform statistical analyses on two or three link and/or node variables. We will create contingency tables visualized as Panigrams.

MultiNet always produces both graphical displays and textual reports; all the figures and tables in this paper were prepared using the program.

Figures 1d and 2b and table 1b use categorical partition variable np(Play) with two unique values ("n" and "p") based on the signs of Normal eigenvector 1 for the Play network. MultiNet makes it easy to define a real-valued variable based on actual eigenvector coordinates; this variable can then be used to perform further operations on the eigenvector coordinates, such as selection of subsets of nodes and binning or quantiling into categories. Relationships between categorical variables can be examined with the resulting contingency tables visualized as Panigrams. These definitions, transformations and analyses are all that will be used in this paper. Further and more detailed information on the program's capabilities can be found in Seary (2005).

A 3-mode medical network of people, symptoms, and exposures

Bipartite representation can also be used for three-mode networks, which have three types of objects and one relationship which is meaningful only between but not within object types. An example is a) people, b) reported symptoms, and c) exposures that were believed to produce the symptoms[6]. Using the method described above in the discussion of bipartite representation of 2-mode networks, we can represent the data as shown in Table 2.

---

[5] There is currently no limit (apart from memory) on the number of nodes and links that can be handled by the Analyse and Variables modules. There is similarly no limit on the number of node and link variables that may belong to a MultiNet data file, and new node and link variables can easily be created when desired.

[6] This dataset came from a University of Toronto study conducted by co-authors and medical researchers Cornelia Baines and Gail McKeown-Eyssen (McKeown- Eyssen, G., Baines, C., *et al*. 2001).

Questionnaires were filled out by patients in general practices. They listed symptoms they had experienced in the last year and any substances (exposures) that they thought might have caused symptoms. Respondents were not asked to link specific symptoms to specific exposures. The medical researchers initially categorized 68 selected symptoms into 14 types (Table 3a) and the 85 exposures into 8 types (Table 3b). The analysis described here did not link individual symptoms to individual exposures because of the large numbers of each reported by some patients (in one case 63 symptoms and 61 exposures). The analysis found a relationship between types of symptoms and types of exposures reported by respondents which was unexpected to the medical researchers. The analysis also suggested a further grouping of types of symptoms based on obvious monotonic trends in the Food (Group A) and Standard Allergens (Group C) exposures.

Table 2. Bipartite representation of 2-mode networks

		people	**symptoms**	**exposures**
		1 2 ... n1	1 2 ... n2	1 2 ... n3
people	1 2 : n1	**0**	**reports of symptoms***	**reports of exposures***
symptoms	1 2 3 : n2	**symptoms reported****	**0**	**0**
exposures	1 2 : n3	**exposures reported****	**0**	**0**

** transposed data matrix * original data matrix

To define the network we begin with three types of nodes: 1340 people, 68 symptoms, and 85 exposures (the network thus has a total of 1493 nodes). A link is defined between a person and a symptom if the person reported that symptom; a link is defined between a person and an exposure if the person reported that exposure. The resulting link variable is called “sym-exp” in figures 3 and 4. For each person, this variable has a value for each symptom and each exposure. The value is “1” if a person reports a particular symptom or a particular exposure; otherwise it is “0.”

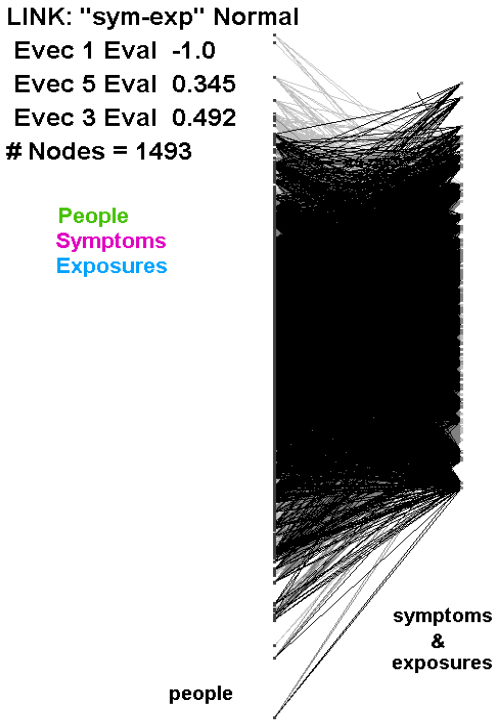
Table 3a. Categories of Symptoms (n = number in category, N=total number reported)

Category	n	N	Group	Examples
Neurocognitive	9	3147	A	forgetfulness, trouble finding words
Affect/Mood	6	3051	A	feeling tense, depressed
Vegetative	3	743	A	sleeping more, compulsive sleepiness
Energy	2	751	A	tiredness, general weakness
Musculoskeletal	5	1283	A	muscle pain, muscle weakness
Endocrine	1	253	A	fast heartbeat
Headache	2	1250	B	other headache, migraine
Gastro-intestinal	6	3129	B	excess gas, bloating
Connective	1	292	B	burning eye
Cardiovascular	1	277	B	irregular heartbeat
Sensory	6	1236	B	light sensitive, bad taste
Infection	4	1841	C	sore throat, hoarse voice
Allergy	12	4323	C	itchy eye, watery eye
Miscellaneous	10	3641	C	sinus fullness, sinus headache
TOTAL	68	25217		

Table 3b. Categories of Exposures (n=number in category, N=total number reported)

Category	n	N	Group	Examples
Food	29	2002	A	coffee, dairy products
Environmental	11	947	B	tobacco smoke, auto exhaust
Home/Work	15	1005	B	household cleaners, disinfectants
Furnishings	10	620	B	TV screen, carpet
Grooming	4	587	B	perfume, cosmetics
Renovation	3	232	B	paint, sawdust
Pharmaceuticals	5	240	B	prescription, non-prescription medicine
Standard Allergens	6	1480	C	pollen, house dust
TOTAL	85	7113		

Figure 3 shows nodes placed according to the coordinates of eigenvectors 1, 3, and 5. In Figure 3a and 3b the green dots correspond to people, the magenta dots to symptoms and the cyan dots to exposures. Since the first eigenvector with eigenvalue -1 perfectly captures bipartite-ness, the two parts of the bipartite network (people and symptoms or exposures) each lie along straight lines in the direction of the Y-axis. People report so many symptoms and exposures (high degrees) that the lines representing links obscure the display, so they are turned off in



a) Bipartite (people on left, symptoms and exposures on right) nature captured perfectly by eigenvector 1, but the lines hide the relationships.

b) With lines off and display rotated slightly, clustering on the right becomes clearly visible.

interactive Fig3

c) Close-up of symptoms (1-15) and exposures (16-23) labelled by type. Upper and lower extremes show a relationship between symptom and exposure types. Coloured by symptom and exposure groups.

Figure 3. Eigenspace displays of 3-mode symptom-exposure-people medical data.

Figure 3b, which is also rotated slightly around the Y-axis for clarity. Eigenvector 3[7] captures the difference in frequency of symptoms and exposures, separating the higher frequency symptoms from lower frequency exposures. As the totals in tables 3a and 3b show, the fre-quencies of symptoms and exposures are quite different with 25,217 symptoms reported (mean of 18.82 symptoms per person) and only 7,113 exposures reported (mean of 5.31 exposures per person). Eigenvector 5 captures the simultaneous clustering of symptoms and exposures. Figure 3c shows more detail of the symptoms and exposures nodes labelled by the types assigned by the medical researchers. Both symptoms and exposures cluster by type, with extremes belonging to Neurocognitive symptoms and Food exposures at the upper left and right, and Allergic Symptoms and Standard Allergen Exposures at the lower left and right.

The analytic strategy

In order to quantify the clustering that appears visually in Figure 3, we start with a Normal eigendecomposition of the network using for "links" the variable that describes reported symptoms and exposures. The first and fifth eigenvectors were used to create new variables. The values of the fifth eigenvector for symptoms and exposures were converted to missing, resulting in a variable that contained only values for people. This variable was recoded into deciles so the lowest ten percent of people were "1"; the next ten percent were "2", etc.

We then performed a crosstabulation of symptom reports, using the symptom's type for rows and the person's Decile for columns. We did the same with exposure reports, using the exposure's type for rows and the person's Decile for columns. The set of steps used to do this analysis in MultiNet are explained in an appendix.

The results are shown graphically in panigrams in Figures 4a and 4b. In both cases, each column of the table is represented as a bar with width proportional to the column's marginal percentage. In columns, segments correspond to rows of the table. The heights of these segments are proportional to the column percentages in the corresponding cells of the table.

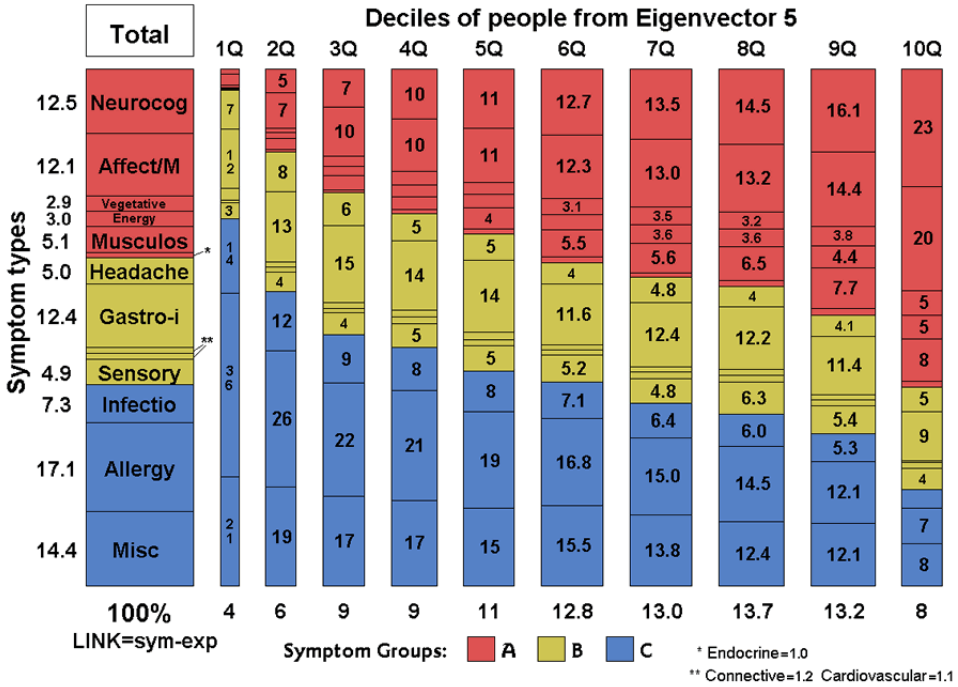
Discussion of results

Figure 3c suggests, and the tables[8] visualized in Figure 4 confirm, that there is a relationship between types of symptoms and types of exposures people report. People who report symptoms in certain categories tend to report exposures in certain categories. For example, more than 50% of the exposures reported by the people in decile 1 are Standard Allergens; almost 75% of the symptoms they report were in Group C (and more than 35% in "Allergy"). That people who report sensitivity to allergens tend to also report allergies is not a surprise, but at the other extreme (of both Figure 3 and 4) is the result that more than 50% of the exposures reported by people in decile 10 were "Food" and more than 60% of the symptoms they report were in

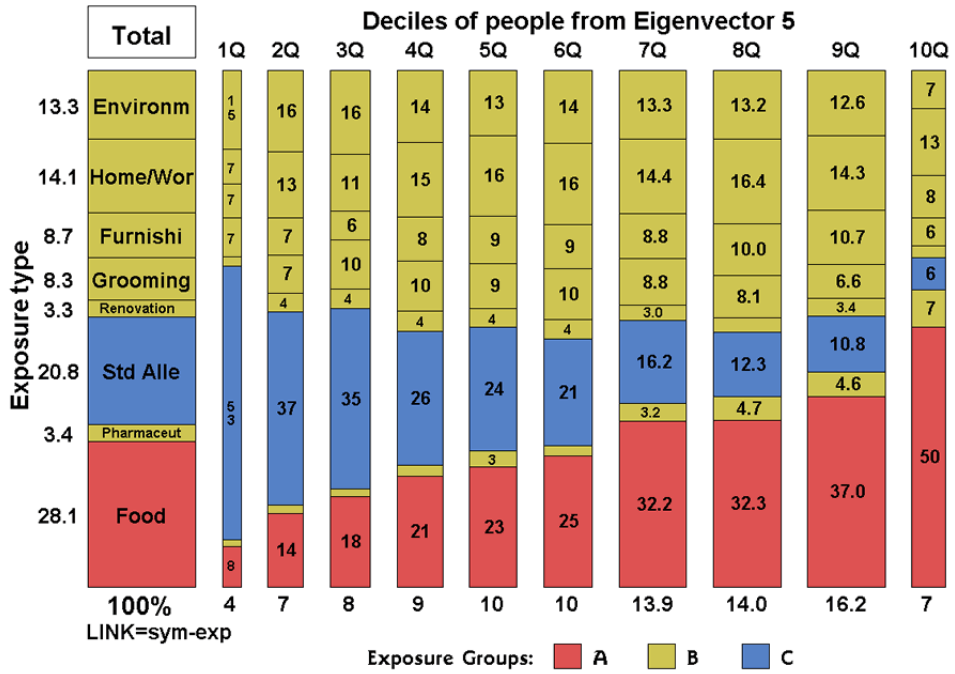
[7] In bipartite graphs, eigenvalues come in pairs with opposite signs. The eigenvectors associated with each pair contain the same values, but the component values for one part have signs reversed, repeating the bipartite-ness captured by eigenvalue -1. For this reason, we did not use eigenvectors 2 and 4 because the eigenvalues they are associated with are the negative copies of 3 and 5.

[8] The cross-tab tables visualized by these panigrams are large — the first one has 15 rows and 10 columns (and marginal rows and columns) resulting in about 500 numbers (including row, column percentages, and counts). Please email the authors if you wish to see these tables.

Group A, with most of these either Neurocognitive (23%) or Affect/Mood (20%). To our knowledge, the relationship between Food exposures and Neurocognitive and Affect/Mood symptoms has not been previously reported.



a) Panigram of counts of symptom types from each of the deciles of people. Coloured by symptom groups.



b) Panigram of counts of exposure types from each of the deciles of people. Coloured by exposure groups.

Figure 4. Panigrams based on symptom and exposure types, with people ordered by eigenvector 5. In both panigrams, unlabelled cells contain less than 3% of their column totals. interactive Fig4a

It is clear that eigenvector 5 captures a difference between people who report symptoms related to allergens and those who report symptoms related to food. On the basis of the trends in eigenvector 5, we collected the categories of both symptoms and exposures into the following groups (table 2):

- Group A has column percents which *increase* steadily (almost monotonically) from decile 1 to decile 10. For Exposures, this group consists of Food. For symptoms, this group includes Neurocognitive, Affect/Mood, Vegetative, Energy, Musculoskeletal, and Endocrine, with the first two contributing more than 50% to the counts. In figures 3c and 4, group A is coloured red.
- Group B does not change monotonically from decile 1 to decile 10. This group is coloured yellow.
- Group C has column percents which *decrease* steadily (monotonically except for one data point) from decile 1 to decile 10. For exposures, the group consists of Standard Allergens. For symptoms this includes categories Infection, Allergy, and Miscellaneous, with the last two contributing about 80% of the counts. In Figures 3c and 4 group C is coloured blue.

These groupings and colourings are used in Figures 3c and 4 to show the smooth relationship between categories (the deciles) of people and the symptoms and exposures they report.

The clusters shown in these figures arise from the relationship between the coordinates of any node in an eigenvector and the coordinates of the nodes it is connected to. This relationship is expressed by equation (1):

$$n_i(s) = \sum_{s \sim t} n_i(t) \Big/ (v_i \times \deg(s))$$

showing that coordinate $n_i(s)$, the s^{th} component of the i^{th} eigenvector, is approximately at the centroid of the coordinates of the nodes s is connected to. The approximation is exact for the constant trivial eigenvector with eigenvalue 1 (where every node has exactly the same coordinate). For eigenvalues far from 1 (which is the case for eigenvector 5 with eigenvalue 0.345), the coordinates can be quite far from the centroid, so that any clusters can be quite smeared out, as we see in Figure 3c. Nevertheless, the analytic strategy outlined here can detect small signals and suggest directions for further analysis.

One reviewer suggested that comparable results could be found by using methods such as factor analysis. However, factor analysis would necessarily require reduction over the "cases" of the data (the people), while fitting to the "variables" (the symptoms and exposures). For example, the default SPSS "Factor" routine would apply Principal Components Analysis (Joliffe, 1986) to the symmetric (and therefore square) matrix produced by correlating the columns of variables, which loses all details about the people. Our method is similar to Correspondence Analysis (Greenacre, 1984) which uses Singular Value Decomposition (Press et. al., 1986) to find the related eigenspaces of the symmetric matrices of cases and variables formed by pre- and post-multiplying the data matrix by its transpose. Our method forms a symmetric matrix by constructing a bipartite graph which retains all details about the cases and variables; the cases and variables are in a single eigenspace and are given their own sets of coordinates as the two parts of the bipartite graph. This allows easy calculations and visualizations such as those shown in the panigrams of Figure 4.

Another reviewer suggested relating Panigrams to Mosaic displays (proposed by Hartigan & Kleine, 1981) to represent contingency tables. Though there are superficial similarities, the two methods were developed independently and have evolved in different directions. Richards developed panigrams as a way to make the information in large crosstabulation tables easily comprehensible (Richards, 1987). Subsequent developments (Richards, 1988, 1993; Seary, 2005) include transposes, three-way contingency tables, three-mode ANOVA, and interactive exploration and interpretation (e.g., Figures 2 and 4). Panigrams have always included row and column marginals, which have never been part of Mosaic displays. In the early 1990's, Michael Friendly extended mosaic displays so they would incorporate residuals into the tiles, allowing the analyst to know whether the observed data deviates from an expected model. While panigrams are used to illustrate the column (or row) percentages and marginals in a two-dimensional crosstab table, with colours representing the categories of row or column variables, Friendly's method uses colour and shading to represent the sign and magnitude of standardized residuals from a specified loglinear model (Friendly, 1991, 1992).

Conclusions

The results presented here show that the combination of spectral methods, for visualizing and partitioning, and contingency tables with panigrams can lead to the extraction of unsuspected relationships, even with high network density and low signal. In this case the categories given by the medical researchers were a good match to the patterns in the data. Without such pre-existing categorizations these methods can also suggest alternative ways of categorizing the data by block models which maximize χ^2 derived from spectral results.

References

Breiger, RL. 1974. The duality of persons and groups. *Social Forces*, **53**(2): 181-190

Breiger, RL, Boorman, S & Arabie, P. 1975. An Algorithm for Clustering Relational Data with Applications to Social Network Analysis and Comparison with Multidimensional Scaling. *Journal of Math. Psych.*, **12**(3): 328-382.

Chow, TY. 1997. The Q-spectrum and spanning trees of tensor products of bipartite graphs", *Proc. Am. Math. Soc.*, **125**(11): 3155-3161.

Chung, FKR. 1995. *Spectral Graph Theory*, CBMS Lecture Notes, AMS Publication.

Crane, D. 1972. *Invisible Colleges: Discussion of Knowledge in Scientific Communities*. Chicago: University of Chicago Press.

Cvetkovic , D, Doob, M, and Sachs, H. 1995. *Spectra of Graphs*. Academic Press.

Davis, A, Gardner, B & Gardner, MR. 1941. *Deep South*. Chicago University Press.

Dhillon, IS. 2001. Co-clustering documents and words using bipartite spectral graph partitioning, UT CS Technical Report #TR 2001-05.

Dodziuk, J & Kendall, WS. 1985. Combinatorial Laplacians and Isoperimetric Inequality. In K.D. Ellworthy (ed.). *From Local Times to Global Geometry*, Pitman Research Notes in Mathematics Series, **150**: 68-74

Fiedler, M. 1973. Algebraic Connectivity of Graphs, *Czech. Math. J.* **23**: 298-305.

Friendly, M. 1991. Mosaic displays for multi-way contingency tables. York Univ.: *Dept. of Psychology Reports*, 1991, No. 195.

Friendly, M. 1992. Mosaic displays for loglinear models. *American Statistical Association, Proceedings of the Statistical Graphics Section*, 1992: 61-68.

Greenacre, M. 1984. *Theory and Application of Correspondence Analysis*. Academic Press.

Hall, K. 1970. An r-Dimensional Quadratic Placement Algorithm. *Management Science*. **17**(3): 219-229.

Hartigan, JA, & Kleiner, B. 1981. Mosaics for contingency tables. In W. F. Eddy (Ed.), *Proceedings of the 13th symposium on the interface between computer science and statistics*, 268-273. New York: Springer-Verlag.

Joliffe, IT. (1986). *Principal Components Analysis*, Springer-Verlag, New York.

McKeown-Eyssen GE, Baines C, Marshall LM, Jazmaji V, & Sokoloff E. 2001. Multiple Chemical Sensitivity: Discriminant Validity of Case. Archives of Environmental Health, **6**: 406-412.

Pothen, A, Simon, H, & Liou, K-P, 1990. Partitioning Sparse Matrices with Eigenvalues of Graphs. *SIAM J. Matrix Anal. App.*, **11**(3): 430-452.

Powers, D. 1988. Graph partitioning by eigenvectors. *Linear Algebra Appl.*, **101**: 121-133.

Press, W, Flannery, B, Teukolsky, S, & Vetterling, W. (1986). *Numerical Recipes*, New York: Cambridge University Press

Richards, WD & Seary, AJ. 1997. Convergence analysis of communication networks, in Barnett, G.A. (ed.), *Advances in Communication Sciences, Vol 15*, Ablex: Norwood NJ, 141-189.

Richards, WD. 1993. FATCAT v4.1VP. A computer program for categorical analysis of multivariate multiplex communication network data. http://www.sfu.ca/~richards/Pdf-ZipFiles/fat41.zip

Richards, WD. 1988. FATCAT ... for Thick Data, *Connections* (the Bulletin of the International Network for Social Network Analysis), Vol. XI, No. 3.

Richards, WD & Rice, R. 1981. The NEGOPY Network Analysis Program, *Social Networks*, **3**(3): 215-224.

Richards, WD. 1971. An Improved Conceptually-Based Method for the Analysis of Communication Networks in Large Complex Organizations. Presented to International Communication Association, Phoenix, Arizona.

Seary, AJ 2005. MultiNet Manual. http://www.sfu.ca/~richards/Multinet/Pages/manual.pdf

Seary, AJ & Richards, WD. 2003. Spectral methods for analysing and visualizing networks: an introduction. In National Research Council, *Dynamic Social Network Modelling and Analysis: Workshop Summary and Papers*, (209-228). Eds. Ronald Breiger, Kathleen Carley and Phillipa Pattison, Division of Behavioral and Social Sciences and Education. Washington DC. The National Academics Press.

Seary, AJ & Richards, WD. 1998. Some Spectral Facts. Presented at INSNA XIX, Charleston. http://www.sfu.ca/~richards/Pages/specfact.pdf

Seary, AJ & Richards, WD. 1995. Partitioning Networks by Eigenvectors. Presented to European Network Conference, London. Published in Everett, M.G. and Rennolls, K. (eds). 1996. *Proceedings of the International Conference on Social Networks, Volume 1: Methodology*. 47-58. http://www.sfu.ca/~richards/Multinet/Pages/variables.pdf

Wasserman, S and Faust, K. 1994. *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.

Appendix

Perform a Normal eigendecomposition of the network using for "links" the variable that describes reported symptoms and exposures:

- Use **Eigenspaces →Normal** with "sym-exp" – the link variable that describes reported symptoms and exposures
- Use **Define →Variables** to create two new variables ("1N-sym-exp" and "5N-sym-exp") from eigenvectors 1 and 5
- Use **Recode →Equation** to select coordinates on the 5th eigenvector for people (and to exclude symptoms and exposures). This is a two-step process. First, take advantage of the fact that people have negative coordinates on eigenvector 1 (Figure 3): multiply the variable that contains eigenvector 5 by "1N-sym-exp<0" which evaluates to "1" if true and "0" if false. The equation (1N-sym-exp<0)*5N-sym will make the coordinates for Symptoms and Exposures equal to 0. This uses properties 1 (bipartiteness) and 2 (visualization) to isolate the people.
- Use **Recode →Zero ->Missing** to turn these 0 coordinates into missing data, excluding them from the next steps of the analysis. The distribution now includes only coordinates for people.
- Use **Recode →Discrete** option **Quantiles** to categorize the people into deciles. Create a new variable to specify which decile each of the 1340 people is in ("Deciles of people from Eigenvector 5" in figure 4). This uses properties 3 and 4 to produce a partition that should result in large χ^2.

Next, perform a network crosstabulation of symptom reports where symptom type is used for rows and Deciles for columns, then another with exposure types for rows and Deciles for columns:

- Use **Network →Xtabs** to form contingency tables counting the types of symptoms reported by people in each of the 10 deciles (Figure 4a).
- Use **Network →Xtabs** to form contingency tables counting the types of exposures reported by people in each of the same 10 deciles (Figure 4b).