Asian Journal of Social Psychology (2008), **11**, 13–41 DOI: 10.1111/j.1467-839X.2007.00241.x

Social network analysis: A methodological introduction

Carter T. Butts
Department of Sociology and Institute for Mathematical Behavioral Sciences, University of California, Irvine, California, USA

Social network analysis is a large and growing body of research on the measurement and analysis of relational structure. Here, we review the fundamental concepts of network analysis, as well as a range of methods currently used in the field. Issues pertaining to data collection, analysis of single networks, network comparison, and analysis of individual-level covariates are discussed, and a number of suggestions are made for avoiding common pitfalls in the application of network methods to substantive questions.

Key words: relational data, social network analysis, social structure.

Introduction

The social network field is an interdisciplinary research programme which seeks to predict the structure of relationships among social entities, as well as the impact of said structure on other social phenomena. The substantive elements of this programme are built around a shared 'core' of concepts and methods for the measurement, representation, and analysis of social structure. These techniques (jointly referred to as the methods of social network analysis) are applicable to a wide range of substantive domains, ranging from the analysis of concepts within mental models (Wegner, 1995; Carley, 1997) to the study of war between nations (Wimmer & Min, 2006). For psychologists, social network analysis provides a powerful set of tools for describing and modelling the relational context in which behaviour takes place, as well as the relational dimensions of that behaviour. Network methods can also be applied to 'intrapersonal' networks such as the above-mentioned association among concepts, as well as developmental phenomena such as the structure of individual life histories (Butts & Pixley, 2004). While a number of introductory references to the field are available (which will be discussed below), the wide range of concepts and methods used can be daunting to the newcomer. Likewise, the rapid pace of change within the field means that many recent developments (particularly in the statistical analysis of network data) are unevenly covered in the standard references. The aim of the present paper is to rectify this situation to some extent, by supplying an overview of the fundamental concepts and methods of social network analysis. Attention is given to problems of network definition and data collection, as well as data analysis per se, as these issues are particularly relevant to those seeking to add a structural component to their own work. Although many classical methods are discussed, more emphasis is placed on recent, statistical approaches to network analysis, as these are somewhat less well covered by existing reviews. Finally, an effort has been made throughout to highlight common pitfalls which can await the unwary researcher, and to suggest how these may be avoided. The result, it is hoped, is a basic reference that offers a rigorous treatment of essential concepts and methods, without assuming prior background in this area.

The overall structure of this paper is as follows. After a brief comment on some things which are not discussed here (the field being too large to admit treatment in a single paper), an overview of core concepts and notation is presented. Following this is a discussion of network data, including basic issues involving representation, boundary definition, sampling schemes, instruments, and visualization. I then proceed to an overview of common approaches to the measurement and modelling of structural properties within single networks, followed by sections on methods for network comparison and modelling of individual attributes. Finally, I conclude with a discussion of some additional issues which affect the use of network analysis in practical settings.

Topics not discussed

The field of social network analysis is broad and growing, and new methods and approaches are constantly in development. As such, it is impossible to cover the entire network analysis literature in one article. Among the topics that are not discussed here are methods for the identification of cohesive subgroups, blockmodelling and equivalence analysis, signed graphs and structural balance, dynamic network analysis, methods for the analysis of two-mode (e.g. person by event) data, and a host of special-purpose methods. Likewise, for topics that are covered here, limitations of space require judicious selection from the set of available techniques. For readers desiring a more

Correspondence: Carter T. Butts, Department of Sociology and Institute for Mathematical Behavioral Sciences, University of California, Irvine, Irvine, CA 92697-5100, USA. Email: buttsc@uci.edu
Received 17 March 2007; accepted 17 April 2007.

extensive treatment, excellent book-length reviews of 'classic' network methods can be found in the volumes by Wasserman and Faust (1994) and Brandes and Erlebach (2005). Some more recent innovations can be found in Carrington, Scott, and Wasserman (2005) and Doreian, Batagelj, and Ferlioj (2005), while Scott (1991) and Degenne and Forsé (1999) serve as accessible introductions to the field. For those looking to keep abreast of the latest developments in network analysis, journals such as *Social Networks*, the *Journal of Mathematical Sociology*, the *Journal of Social Structure*, and *Sociological Methodology* frequently publish methodological work in this area. Due to the slowness of the academic publishing process, a growing (if not always welcomed) trend is the use of technical report and working paper series as an initial mode of information dissemination. While these sources are rarely peer reviewed, they frequently contain research which is 1–3 years ahead of that contained in the journals. Caution should be used when drawing upon such sources, but they can be a valuable resource for those seeking research on the cutting edge.

Notation and core concepts

Because structural concepts are not well described using natural language, scientists in the social network field use specialized jargon and notation. Much of this is borrowed from graph theory, the branch of mathematics which is concerned with discrete relational structures (for an overview, see West, 1996 or Bollobás, 1998). Indeed, the close relationship between graph theory and the study of social networks is much like the relationship between the theory of differential equations and the study of classical mechanics:[1] in both cases, the mathematical literature provides a formal substrate for the associated scientific work, and much of the theoretical leverage in both scientific fields comes from judicious application of results from their associated mathematical subdisciplines. While the graph theoretical formalisms used within the social network field can seem daunting to the newcomer, the core concepts and notation are easily mastered. We begin, therefore, by reviewing some of these elements before advancing to a discussion of network data and methods.

A social network, as we shall here use the term, consists of a set of 'entities', together with a 'relation' on those entities. For the moment, we are unconcerned with the specific nature of the entities in question; persons, groups, or organizations may be objects of study, as may more exotic entities such as texts, artifacts, or even concepts. We do assume, however, that the entities which form our network are distinct from one another, can be uniquely identified, and are finite in number. (Extensions to incorporate more general cases are possible, but will not be treated here.) Likewise, we constrain the set of potential relations to be studied not by content, but by their formal properties. Specifically, we require that relations be defined on pairs of entities, and that they admit a dichotomous qualitative distinction between relationships which are present and those which are absent. A wide range of relations can be cast in this form, including attributions of trust or friendship, interpersonal communication, agonistic acts, and even binary entailments (e.g. within mental models). Relations which do not satisfy these constraints include those which necessarily involve three or more entities at once (e.g. the respective A-B-O or P-O-X triads of Newcomb (1953) and Heider (1946)), or those for which the presence/absence of a relation is not a useful distinction (e.g. spatial proximity). Formalisms which can accommodate these more general cases exist; see Wasserman and Faust (1994) for some examples.

Within the above constraints, we may represent social relations as graphs. A graph is a relational structure consisting of two elements: a set of entities (called vertices or nodes), and a set of entity pairs indicating ties (called edges). Formally, we represent such an object as $G = (V, E)$, where V is the vertex set and E is the edge set. Where multiple graphs are involved, it can sometimes be useful to treat V and E as operators: thus, $V(G)$ is the vertex set of G, and $E(G)$ is the edge set of G. When used alone (as V and E) these elements are tacitly assumed to pertain to the graph under study. We represent the number of elements in a given set by the cardinality operator, $|\cdot|$, and hence $|V|$ and $|E|$ are the numbers of vertices and edges in G, respectively. The number of vertices in a given graph is known as its order or size, and will be denoted here by $n = |V|$ where there is no danger of confusion. We will also use simple set theoretical notation to describe various collections of objects throughout this paper (as is standard in the network literature). In particular, $\{a, b, c, \ldots\}$ refers to the set containing the elements a, b, c etc., and $(a, b, c \ldots)$ refers to an ordered set (or tuple) of the same objects. Note that the order of elements matters only in the latter case; thus $\{a, b\} = \{b, a\}$, but $(a, b) \neq (b, a)$. Intersections and unions of sets are designated via $\cap$ and $\cup$, respectively, so that, for example, $A \cup B$ is the union of sets A and B. Setwise subtraction is denoted via the backslash operator, so that $A \backslash B$ is the set formed by removing the elements of B from A. Subsets are denoted by $\subset$ (for proper subsets) and $\subseteq$ (for general subsets), such that $A \subset B$ means that A is a proper subset of B. Set membership is similarly denoted by $\in$, with $a \in A$ indicating that object a belongs to set A. Finally, we use the existential ($\exists$, reading as 'there exists') and universal ($\forall$, reading as 'for all') quantifiers in making statements about objects and sets. While this notation may be unfamiliar to some readers, it provides a precise and compact language for describing structure which cannot be obtained using natural language. This notation is frequently encountered within the network literature, particularly in more technical papers.

Returning to the matter of graphs, we note that they appear in several varieties. These varieties are defined by the type of relationships they represent, as reflected in the content of their edge sets. Graphs which represent dyadic (i.e. pairwise) relations which are intrinsically symmetric (i.e. no distinction can be drawn between the 'sender' and the 'receiver' of the relation) are said to be undirected (or non-directed), and have edge sets which consist of unordered pairs of vertices. For these relations, we express this principle formally via the statement that $\{v, v'\} \in E$ if and only if ('iff') vertex v is tied (or adjacent) to vertex v' (where $v, v' \in V$). By contrast, other graphs represent relations which are not inherently symmetric, in the sense that each relationship involves distinct 'sender' and 'receiver' roles. These graphs (which are called directed graphs or digraphs) have edge sets which are composed of ordered pairs of vertices. Formally, we require that $(v, v') \in E$ iff v sends a tie to v'. Note that, as shorthand, it is sometimes useful to use arrow notation to denote ties, such that $v \rightarrow v'$ should be read as 'v sends a tie to v'' (or, equivalently, v is adjacent to v'). An edge from a vertex to itself is a special type of edge known as a loop, and may or may not be meaningful for a particular relation. Relations which are irreflexive (i.e. have no loops) and which are not multiplex (i.e. do not allow duplicate edges) are said to be simple. Graphs used here will be presumed to be simple unless otherwise indicated.

When working with graphs, it is often useful to be able to speak of smaller elements within a larger whole. In this vein, we define a subgraph to be a graph whose elements are subsets of a larger graph; formally, H is a subgraph of G (denoted $H \subseteq G$) iff $V(H) \subseteq V(G)$ and $E(H) \subseteq E(G)$. One important type of subgraph is formed by taking a set of vertices, together with all edges between those vertices. For vertex set $S \subseteq V$, we refer to this as the subgraph induced by S, or $G[S]$. Another important type of substructure is the neighbourhood, which consists of all vertices which are adjacent to a particular vertex. For simple graph G, $N(v) \equiv \{v' \in V: \{v, v'\} \in E\}$ denotes the neighbourhood of vertex v (where $\equiv$ should be read as 'is defined as'). The directed case obviously forces the distinction between neighbours to whom ties are directed (out-neighbours) and neighbours from whom ties are received (in-neighbours). These are denoted, respectively, as $N^+(v) \equiv \{v' \in V: (v, v') \in E\}$ and $N^-(v) \equiv \{v' \in V: (v', v) \in E\}$, with the joint neighbourhood $N(v) \equiv N^+(v) \cup N^-(v)$ being the union of the two. When discussing neighbourhoods, we often refer to the focal vertex (v) as ego with neighbouring vertices ($v' \in N(v)$) referred to as alters; indeed, this language may be used whenever we consider a particular individual and those who relate to him or her. Two vertices with identical neighbourhoods are said to be copies of each other, or (as it is better known in the social sciences) are said to be structurally equivalent (Lorrain & White, 1971).[2] Combining ideas, we also note that $G[v' \cup N(v)]$ is a succinct way of referring to the subgraph of G formed by selecting v and its neighbours along with all edges among them; this structure (called an egocentric network) will surface frequently throughout the present paper.

While graphs derived from empirical data are frequently complex, there are a number of useful graph theoretical terms for simple structures which are encountered (if only as subgraphs) in various settings. The simplest of these is the empty graph (or null graph), which consists of a vertex set with no edges. The null graph on n vertices is traditionally denoted N_n, and has the trivial structure $N_n = (V, \varnothing)$ where $\varnothing$ denotes the null set. A vertex whose neighbourhood is empty is referred to as an isolate and, hence, the null graph can be thought of as a graph that contains nothing but isolates. The corresponding opposite of the null graph is the complete graph or clique on n vertices, denoted K_n. K_n consists of n vertices, together with all possible ties among them (discounting loops, if the relation in question is simple). N_n and K_n are said to be complements of each other, in that an edge exists in one graph iff that edge does not exist in the other. More generally, the complement of G (denoted $\overline{G}$) is defined as the graph on $V(G)$ such that $v \rightarrow v'$ in $\overline{G}$ iff $v \not\rightarrow v'$ in G. Finally, another 'special' graph of which it is useful to be aware is the star, which consists of one vertex with ties to all others, and no other edges. The star on n vertices is denoted $K_{1,n-1}$, reflecting the fact that the star is a complete bipartite graph. A graph is said to be bipartite if its vertices can be divided into two non-empty disjoint sets, A and B, such that $G[A]$ and $G[B]$ are both null graphs. A complete bipartite graph is one in which all possible between-set edges exist but (from the definition of a bipartite graph) no within-set edges exist, and is denoted $K_{a,b}$ (where a and b are the cardinalities of A and B, respectively). It follows therefore that a graph with one vertex which is adjacent to all others (none of which are adjacent to each other) can be thought of as a complete bipartite graph in which one of the two vertex sets has only one member (and hence a $K_{1,n-1}$).

Although idealized structures such as the above are helpful when describing graphs, there are also other properties for which special terminology is useful. In many cases, we will be interested in determining whether one vertex could reach another by traversing a series of edges within the network. A sequence of distinct, serially adjacent vertices $v, \ldots, v'$ together with their included edges is called a path (or a directed path, if G is directed), and the existence of a path from v to v' implies that the two vertices are in some way connected. In an undirected graph, there is only one form of connectedness: v and v' are connected iff there exists some v, v' path in G. In directed graphs, by contrast, several distinct notions of connectedness are possible. At the lowest level, we may consider v and v' to be connected iff there exists a sequence of vertices from v to v'

such that, for any adjacent pair (v'', v''') in the sequence, $v'' \rightarrow v'''$ and/or $v''' \rightarrow v''$. Such a structure is called a semipath, and two vertices joined by a semipath are said to be weakly (or semipath) connected. A slightly more stringent condition is for there to exist either a directed path from v to v' or such a path from v' to v (but possibly not both). This does require a sequence of vertices which can be traversed in order to get from one end of the path to the other, but this condition is not required to hold in both directions. A vertex pair satisfying this condition is said to be unilaterally connected. A criterion which is more stringent yet is to require that there exists a directed path from v to v' and that there exists a directed path from v' to v; vertex pairs for which this condition is met are said to be strongly connected. Finally (and most stringently of all), we may require not only the existence of directed v, v' and v', v paths, but also that these paths traverse the same intermediate vertices. Vertex pairs satisfying this reciprocal condition are said to be recursively connected. This same terminology can be extended to describe larger sets of vertices as well. In particular, a vertex set is said to be connected if all pairs of vertices within it are connected (with the type of connectivity being specified in the directed case). Likewise, a graph G is said to be connected if all pairs of vertices in V are connected. Specific types of connectivity (weak, unilateral etc.) are again relevant in the directed case, with strong connectivity being the conventional 'default' assumption if no qualifier is given. A maximal set of connected vertices in G is said to form a component of G, with G as a whole being connected iff it has only one component. Components and connectedness play an important part in the study of phenomena such as information transmission, and will be invoked here on multiple occasions.

Several additional path-related concepts also bear mentioning. A geodesic from v to v' is a v, v' path of minimal length; the length of such a path is called the geodesic distance (or simply distance) from v to v'. The path concept may also be generalized in various ways, some of which are important for our present purposes. A sequence of distinct, serially adjacent vertices which both begins and ends with vertex v (together with its included edges) is called a cycle; this is directly analogous to a path, save in that the start and end-points are the same. Both the path and the cycle are special cases of the 'walk', which is simply a sequence of serially adjacent vertices together with their included edges. Unlike a path, a walk may visit a given edge or vertex multiple times and, hence, can be of any length. A path, by contrast, must have a length of, at most, $n - 1$, as vertices within a path may not be repeated. A path of length $n - 1$ must touch all vertices, and is known as a spanning (or Hamiltonian) path. More generally, any subgraph of G which contains all elements of V is known as a spanning subgraph, with spanning paths, walks, cycles etc. being special cases. Interestingly, for many classes of graphs, the average geodesic distance among connected vertices (or mean geodesic distance) can be very small compared to the length of a spanning path – this result lies behind the 'small world' phenomenon famously studied by Travers and Milgram (1969), Pool and Kochen (1979), Watts and Strogatz (1998), and others.

Before concluding this section, I note some additional concepts which are subtle but important for what follows. A one-to-one function ℓ which takes V onto itself is said to be a permutation or labelling function for V. A relabelling or graph permutation of G is then a transformation of G which relabels its vertex set by ℓ, i.e. (in a slight abuse of notation) $\ell(G) = (\ell(V), E)$. A permutation which preserves the adjacency structure of G is said to be an automorphism of G. ℓ is hence an automorphism iff $\ell(G) = G$. Relatedly, two distinct graphs G and G' on vertex set V are said to be isomorphic iff there exists a permutation ℓ such that $\ell(G) = G'$. This is denoted $G \simeq G'$, with $\simeq$ read as 'is isomorphic to'. Isomorphic graphs are structurally identical, differing only in the identity of their respective vertices. A maximal set of mutually isomorphic graphs is referred to as an isomorphism class, and each graph within the set can be converted into any other by means of a graph permutation. Another transformation-related concept is the graph minor, which is a graph formed by merging (or condensing) adjacent vertices of G. In particular, let v, v' be adjacent vertices in G, and form the graph $G' = (V', E')$ by letting $V' = V \backslash v$ and setting E' such that $N(v') = (N(v') \cup N(v)) \backslash v$. Then, G' is a graph minor of G. Furthermore, if G'' is a graph minor of G' and G' is a graph minor of G, then G'' is said to be a graph minor of G as well. Thus, a graph formed by condensing any sequence of vertices of G is a graph minor of G. As we shall see, graph minors are useful for defining the number of 'levels' in a hierarchical structure, a substantively important property of directed graphs. For further reading on graph minors, isomorphism, or the other concepts discussed here, West (1996) provides an accessible introduction.

Finally, I note that the above concepts may be expanded in various ways to accommodate more general relational structures. Of particular importance are valued edges (i.e. edges which are associated with the value of a variable such as frequency, tie strength, etc.) and vertex attributes (sometimes called 'colours' in the graph-theoretical literature). Edge values and vertex attributes are frequently encountered in empirical network data, as I shall discuss below.

Network data

Before considering how networks may be analyzed, I first begin with a general discussion of network data. As network data are represented in a different form from the

matrix/vector format familiar to most social scientists, I begin with a brief discussion of how such data may be numerically represented. This is useful both notationally (for the discussion which follows) and also pragmatically, as most available network analysis tools assume some basic familiarity with the representation of network data. From this, I turn to a discussion of network boundary definition, the most fundamental issue to be determined when creating or assessing a network study. I also say a few words about the collection of network data (designs and instruments), with particular emphasis on the collection of data on the connections between individuals. Finally, I provide some background on the visualization of network data, a problem which has been foundational to the development of modern network analysis (Freeman, 2004).

Representation

Network data can be represented in a number of ways, depending upon what is most convenient for the application at hand. We have already seen that networks can be represented using graph theoretical notation, and I shall use this representation extensively in more conceptual discussions. For practical purposes, however, network data are more often represented in other ways. The most common data representation in empirical contexts is the adjacency matrix, an $n \times n$ matrix whose ijth cell is equal to 1 if vertex i sends an edge to vertex j, and 0 otherwise. For an undirected graph G with adjacency matrix $\mathbf{A}$, it is clear that $A_{ij} = A_{ji}$ (i.e. the adjacency matrix must be symmetric). This is not generally true if G is a digraph. If G is simple (i.e. G has no loops), then all elements of the diagonal of $\mathbf{A}$ will be identically 0. Otherwise, $A_{ii} = 1$ iff vertex i has a loop (this being identical for directed and undirected graphs).

Several other data representation issues also bear mention. In the special case of networks with valued edges, we use the above representation with the minor modification that A_{ij} is the value of the (i, j) edge (conventionally 0 if no edge is present). When representing multiple relations on the same vertex set, it is also useful to extend the notion of the adjacency matrix to encompass the adjacency array. For a set of graphs $G_1, \ldots, G_m$ on a common vertex set V having order n, we use the $m \times n \times n$ adjacency array $\mathbf{A}$ such that $A_{ijk} = 1$ if j sends an edge to k in G_i, and 0 otherwise. As usual, we replace cell values with edge values in the non-dichotomous case.

Although adjacency arrays are simple to work with, they can be unwieldy where n is very large (especially if G is very sparse). In such cases, it is common to store networks via edge lists, or pairs of vertices which are tied to one another. Another representation which is sometimes useful is the incidence matrix, a $n \times |E|$ matrix $\mathbf{I}$ such that $I_{ij} = 1$ if i is an end-point of edge j and 0 otherwise. Direction within incidence matrices is denoted via signs, such that $I_{ij} = -1$ if i is the source of the jth edge of G, and $I_{ij} = 1$ if i is, instead, the destination of the jth edge. Incidence matrices are relatively unwieldy, and are defined only up to a column permutation; as such, they are not often used in conventional network research. However, incidence matrices are very useful for representing hypergraphs (i.e. networks whose edges involve more than two end-points) and for two-mode data (i.e. networks consisting of connections between two disjoint types of entities). I do not treat these applications here, although the interested reader may turn to Wasserman and Faust (1994) for an introductory account.

Network boundary definition

As noted above, a social network is defined by a set of entities, together with a social relation on those entities. As such, a network is bounded by the set of entities on which it is defined. While the same principle applies to any social grouping, network boundaries are of particular importance due to the intrinsically interactive nature of relational systems. Specifically, a misspecified network boundary may include or exclude not only some set of relevant or irrelevant entities, but also all relationships between those entities and others in the population (not to mention all relationships internal to the included/excluded entities). Furthermore, many structural properties of interest (e.g. connectivity) can be affected by the presence or absence of small numbers of relationships in key locations (e.g. bridging between two cohesive subgroups). Thus, the inappropriate inclusion or exclusion of a small number of entities can have ramifications which extend well beyond those entities themselves, and which are of far greater importance than the types of misspecification which occur in most non-relational settings. As such, it is vital to define the network boundary in a substantively appropriate manner, and to ensure that subsequent analyses reflect that choice of boundary (and not, for example, a boundary which simply happens to be methodologically convenient). In practice, of course, network boundaries are set in a number of ways, and it is useful to review those most frequently encountered in the network literature.

Exogenously defined boundary. In the ideal case, one has a clearly specified substantive theory which indicates the entities that are relevant for some phenomenon of interest, and whose ties are, hence, relevant for subsequent analysis. The network boundary is then exogenously defined by one's substantive knowledge, and one's research task then shifts to measuring ties among the indicated entities. Exogenously defined boundaries are common in small group and intra-organizational studies, wherein membership is well defined and one is frequently concerned only with interactions among group members (e.g. Krackhardt & Stern, 1988; Lazega, 2001). Studies of relationships within spa-

tially defined units (e.g. residential studies like those of Festinger, Schachter, and Back (1950) and Yancey (1971)) serve as another example, although it is important to ensure that the theoretically relevant relations are truly restricted to the spatial boundary. Indeed, the same problem may surface in organizational settings, when researchers suddenly shift focus from a locally defined question (e.g. who has the most within-group friendships?) to one which has non-local elements (e.g. who has the most friendships overall?). The extent to which a given sample may be regarded as exogenously bounded thus depends on the research question being pursued, rather than the data in hand.

Relationally defined boundary. A less common means of defining a network boundary is endogenously (i.e. by specifying the relevant entities as those who satisfy some condition of social closure). Intuitively, the presumption in this case is that entities and relations within the 'closed' set do not depend on those beyond that set and, hence, may be studied separately. Definition of the network boundary is thus determined by the closure condition, and usually by a set of 'seed' entities who are defined as being of intrinsic interest. For instance, in a study of interaction among community organizations, a researcher might define the relevant network as consisting of some small set of 'core' organizations (e.g. the Mayor's Office or Chamber of Commerce) together with all the organizations that can be reached by the core organizations through some path in the relevant network. As organizations not in this set do not (by construction) have any contact with those in the set, the resulting network may be presumed to be sufficiently decoupled from its surroundings to permit independent analysis. (See Freeman, Fararo, Bloomberg, and Sunshine (1963) for a related discussion.) As with exogenous boundary definitions, the plausibility of this assumption must rest on substantive knowledge regarding the phenomenon under study, and should not be naïvely assumed. For instance, if a lack of ties to external organizations (e.g. major employers) were critical to the phenomenon of interest, then the network boundary definition in the above example would be inappropriate. The use of relationally defined boundaries does not, therefore, exempt one from verifying that one's inclusion criterion is theoretically appropriate.

Methodologically defined boundary. Finally, the network boundaries for many studies are determined by the methodology that is used to obtain the network in question. For instance, sampling interaction via a given communication medium (e.g. email, radio communication etc.) may implicitly limit the measured network to those using the medium in question; more explicit boundary effects may result from measurement designs such as those described below. While sometimes problematic for the reasons described above, there are some circumstances in which methodologically defined boundaries may be appropriate. In particular, if it can be shown that inference for some quantity of substantive interest requires only the observation of particular ties (e.g. ego's alters and all ties among them), then it may be both reasonable and efficient to restrict one's data collection to the particular relationships that are required for the intended purpose. This is, in fact, a form of theory-based boundary definition, save that it is the relevant theory of inference, rather than a theory of process or structure, which guides the process. While this is a legitimate approach where applicable, one must still ensure that the inferential theory being used is substantively appropriate, and that the information being gathered is, in fact, adequate to draw inferences which are of substantive interest. One cannot justify choosing a network boundary on methodological grounds if the methodology in question is not itself appropriate for the problem at hand.

Common measurement designs

A question apart from (but related to) the network boundary definition is the question of network measurement. Broadly speaking, the designs used in network measurement attempt to permit inference at one of three levels. Personal or egocentric inference centres on the properties of individuals' local networks. These may be limited to the number of alters to whom ego is tied, but may also include individual attributes of those alters and/or the existence of ties among them. Strict egocentric inference does not seek to generalize beyond ego's local structure and, hence, does not involve the 'linking' of personal networks among multiple individuals (even where this is possible); while it is limited in its ability to yield insights regarding global structure, egocentric inference has modest data requirements, and is easily adapted to large-scale survey research. For this reason, most population-level network studies (e.g. the network modules of the General Social Survey (Davis & Smith, 1988) and International Social Survey Program) are of this type. A more ambitious goal than egocentric inference is general network inference, in which the goal is detailed reconstruction of the entire social network on a given population. Studies of this kind (sometimes called 'complete network' or 'network census' studies) allow for the determination of both global and local social properties, and are hence the 'gold standard' of network analysis. Most organizational and small group studies are designed with the goal of complete network inference, but the strict data requirements make this goal difficult to obtain for networks on large populations. Finally, a third level of inference involves the attempt to estimate cognitive social structures (Krackhardt, 1987a) (i.e. the view of the complete social structure as understood by each member of the network). Although distinct from complete network inference in the above sense, knowledge of cognitive social structures can

serve as a basis for accomplishing the former via appropriate data aggregation models (Romney, Weller, & Batchelder, 1986; Batchelder & Romney, 1988; Butts, 2003). Cognitive social structures are nevertheless important targets of inference in their own right, and should not be assumed to be exact replications of behavioural networks (Bernard, Killworth, Kronenfeld, & Sailer, 1984; Krackhardt, 1987a).

Given that we may seek to infer structure at the personal network, complete network, or cognitive level, there are a number of designs which can be used to meet this objective. Here, I briefly outline some of the major varieties that are currently used in the study of interpersonal networks. Each grouping listed here has many subvariants, which will not be treated in detail. Further descriptions of many related issues can be found in Marsden (1990, 2005) and Morris (2004).

Own-tie reports. The most common designs in interpersonal network measurement consist of variants on the own-tie report scheme: selected informants are asked to report on the ties to which they are an end-point. For directed relations, some own-tie reporting schemes are one-way; that is, ego is asked to provide either incoming or outgoing ties, but not both. In other cases, ego may be asked to provide both incoming and outgoing ties of which he or she is an end-point. The egos sampled for own-tie reporting schemes are generally the entire set of network members (where inference is sought regarding all ties in the network), or a probability sample thereof (when only average properties of alters are required). When implemented in the former case (with all egos reporting), own-tie designs supply either one (for one-way) or two (for two-way) reports per potential edge. As such, they tend to be vulnerable to both non-response and measurement error, although the former is much less problematic in personal network studies (wherein no attempt is made to infer the entire network).

Complete egocentric designs. Another common set of designs comprises the complete egocentric family. In a complete egocentric design, selected informants are first asked to nominate those with whom they are tied (as in an own-tie report design). This is then followed by a second phase, in which ego is asked to identify which pairs of alters are tied to one another. As with own-tie designs, these identifications may be one way or two way in the directed case, and egos may be chosen in a number of ways. Most commonly, complete egocentric designs are used in personal network research, where egos are sampled from a larger population (and no attempt is made to link alters across egos). In this case, the complete egocentric designs have the advantage of providing information regarding ego's local structural context, while still being simple enough to be administered via standard survey instruments. Although uncommon, complete egocentric designs can also be used when attempting a network census, in which case they provide some redundant information regarding particular edges. (Specifically, each potential edge will receive one report per informant who reports being tied to both end-points, or who is an end-point and who reports being tied to the other end-point.) Unfortunately, such third-party reports are non-ignorably dependent upon informant error rates and, hence, the use of network inference models like those of Butts (2003) is non-trivial for such data. More generally, it should be noted that reporting errors on the part of ego regarding his or her personal ties will affect ego's reports of alters' ties under a complete egocentric design, as reports are elicited only for edges among those to whom ego claims to be tied. The consequences of this potential for complete egocentric network designs to amplify measurement error are not well studied at this time.

Link-trace designs. To provide valid inferences, the above designs require ignorable methods of drawing egos from the population of network members (to infer personal network structure) or taking a census of egos (for complete network inference). In some cases, however, we may lack a sampling frame for network membership (e.g. when studying a hidden population) or may need to estimate global network property without measuring all members of a large population. In such settings, link-trace designs serve as a potential option. Broadly speaking, link-trace designs are adaptive sampling methods (Thompson, 1997) which operate by iteratively eliciting alters from a current set of egos (as in own-tie report), and then using these alters as egos in further waves of data collection. In this way, link-trace designs 'walk' through the network, following chains of ties from current respondents to future respondents. Variants of link-trace designs include snowball sampling (Goodman, 1961), random-walk sampling (Klovdahl, 1989), and respondent-driven sampling (Heckathorn, 1997, 2002), all of which use somewhat different procedures for selecting an initial 'seed' sample, contacting egos within each wave, determining which alters to trace in additional waves, and deciding how many waves to use. While complex to implement and analyze, link-trace methods have the desirable feature that they can generate reasonable estimates without representative seed samples; somewhat counterintuitively, the Markovian properties of the sampling mechanism tend to reduce the impact of the seed sample on subsequent waves (see Heckathorn, 2002 for a discussion, and Tierney, 1996 for related commentary on convergence in Markov chains). Furthermore, link-trace designs can allow for some types of global network inference, despite the fact that not all edges are measured (see Thompson & Frank, 2000 for details). However, link-trace designs generally provide, at most, one to two measure-

ments per potential edge (depending on the elicitation scheme used), and share with complete egocentric designs the problem that sampling is potentially contaminated by reporting error. How robust these designs are to such errors is currently unknown, as are many other aspects of their performance in realistic settings. As such, link-trace designs have a great deal of promise, but should be used with caution.

Arc sampling designs. A final category of designs are those based on arc sampling ('arc' being another term for directed edge). Arc sampling designs differ from the others discussed here in that they begin by selecting particular edges to measure, and then seek information on those edges. Importantly, this information need not come from the individuals who are end-points to the edges in question: observer or third party informant reports, archival materials, or even sensor data (Choudhury & Pentland, 2003) can serve to produce observations. The observational data famously reported by Killworth and Bernard (1976); Bernard and Killworth (1977); Killworth and Bernard (1979); Bernard, Killworth, and Sailer (1979) can be understood as arising from an arc sampling design, as is the cognitive social structure (CSS) design used by Krackhardt (1987a) (in which every network member is asked to report on the ties between all other network members). Frank (2005) describes arc sampling designs which arise from contexts in which one samples on realized interactions, rather than potential interactions; some archival data are of this form (e.g. news accounts of partnerships among firms). Another family of arc sampling designs is described by Butts (2003), in which multiple sources are queried about the state of various potential edges, such that each potential edge is measured a fixed number of times (with measurements being balanced across sources). This family of designs is intended for use with data from informants or observers, and provides a way to reduce the considerable respondent burden imposed by the CSS design.

Because they allow for multiple measurements on each potential edge, arc sampling designs can be used to provide complete network estimates which are highly robust to reporting error and missing data (Butts, 2003). However, the number of observations required can prove burdensome to respondents, and the more complex designs can be difficult to execute. Most such designs also require that the target population be known in advance, although they do not necessarily require that network members be willing or available to supply information on their own ties; observers, sensors, or informants may be used to provide information on persons who are otherwise unavailable, assuming that these sources do, in fact, have such information (an assumption which should be checked via error estimates). Likewise, combining measurements from multiple error-prone sources requires appropriate statistical modelling, as sources may vary greatly both in overall accuracy and in the types of errors generated. Arc sampling designs are thus very effective tools for producing high-quality estimates at the complete network level, but require a greater investment of resources than do simpler approaches.

Common measurement instruments

Although networks may be obtained from archival materials, sensors, observation, or many other sources, much network data is gleaned from human informants via survey instruments. The most common instruments used in the field are of two basic types: prompted recall or 'roster' instruments, and free list or 'name generator' instruments. Both instrument types have particular strengths and weaknesses, and we consider each in turn.

Rosters. Perhaps the most common type of instrument for measuring interpersonal networks is the roster. Roster instruments typically consist of a stem question (e.g. 'To whom do you go for help or advice at work?') followed by a list of names. Subjects are instructed to mark the names of those with whom they have the indicated relation, leaving the others blank. Such an instrument is simple to use, and minimizes false negatives due to forgetting (as it automatically prompts for all alters). On the other hand, instrument length grows linearly with the number of possible alters, and generally becomes unwieldy when more than 30–50 names are involved. Likewise, a roster instrument can only be used where the set of potential alters is known in advance, and where that set can be divulged to the subjects without creating a breach of confidentiality. In a context such as Heckathorn's (1997) study of ties among intravenous drug users in New Haven, Connecticut, provision of a roster instrument would be both impractical and unsafe: impractical due to the difficulty of knowing the (hidden) population of intravenous drug users before administering the instrument, and unsafe due to the potential legal consequences of compiling and disseminating such a list within the study population. Despite such concerns, roster instruments can be effectively deployed in many contexts, and should generally be the preferred to name generators (see below) where feasible.

Name generators. The primary alternative to roster instruments for the collection of interpersonal network data is the use of name generators. A name generator consists of a question which asks the subject to produce from memory a list of individuals, generally those with whom the subject has some relationship. The name generator therefore differs from the roster instrument only in employing a free list protocol, as opposed to prompted recall. False negatives due to forgetting and subject fatigue are of concern here, particularly for relations for which ego has a large number

of ties (Brewer, 2000). However, this approach can be deployed where supplying a roster would be impossible, impractical, or would pose an unacceptable risk to subjects. As a result, name generators are often used in large-scale network studies, and in studies of sensitive and/or hidden populations. Although rosters are generally preferred to name generators where possible, both methods are likely to produce fairly similar results provided that the questions being asked do not pose an excessive mnemonic challenge, and that the number of alters for each ego is reasonably small.

Visualization

Networks are commonly depicted via displays in which each vertex is represented by a polygon or other shape (frequently a circle), with lines connecting the shapes associated with adjacent vertices. (Arrows are generally used to display directed edges, with the arrowhead pointing in the direction of the receiving vertex.) The introduction of such displays in the social sciences is generally credited to Moreno (1934), who coined the term sociogram to describe them. Unlike other data displays commonly used in scientific contexts, the specific location of points (vertices) in a sociogram is generally arbitrary, and is usually driven by communicative and aesthetic criteria: this is because the network is defined by the pattern of ties among vertices, a property which is not affected by the placement of vertices within the display. That said, some displays generally prove more effective than others in revealing network structure (McGrath, Blythe, & Krackhardt, 1997), and certain methods of placing vertices within a sociogram (known as layout algorithms) are more widely used than others. The most common layout algorithms are based on what are known as force-directed placement schemes, in which vertex placement is determined by a hypothetical physical process usually incorporating attraction between adjacent vertices balanced by a general tendency toward repulsion among all vertices. Examples of such schemes include the Fruchterman-Reingold (Fruchterman & Reingold, 1991) and Kamada-Kawai algorithms (Kamada & Kawai, 1989), both of which may be found in common network visualization and analysis packages (Butts, 2000; Batagelj & Mrvar, 2007; Borgatti, 2007). While other more exotic approaches are available, most layout algorithms share with these methods the common goals of placing vertices close to their network neighbours, preventing two vertices from occupying the same location, minimizing the number of edge crossings, and maintaining approximately constant edge length. With the exception of certain special classes of networks (e.g. the planar graphs (West; 1996)), these goals cannot generally be satisfied simultaneously. Different layout algorithms thus prioritize different visualization goals, as well as additional objectives such as scalability to extremely large graphs. The creation of such algorithms has spawned its own field within computer science (the field of graph drawing), and is a topic of active research.

In addition to layout methods designed to optimize aesthetic criteria, layout methods are sometimes used to convey specific structural information. Target diagrams, for instance, place vertices on a series of circular shells based on some specified criterion (e.g. centrality scores); although used in network analysis since before the dawn of computer-aided display (Freeman, 2000), they are now used infrequently due to their poor applicability to large and/or dense networks. Another popular method for determining vertex position is the use of multidimensional scaling (Torgerson, 1952) or eigenvector solutions (Richards & Seary, 2000), which can be used to superimpose network information on a more common multivariate display. A 'hybrid' approach which stands between purely aesthetic and data analytical layout methods are latent space models such as those of Hoff, Raftery, and Handcock (2002) and Handcock, Raftery, and Tantrum (2007). Although they can be viewed as proper stochastic models of network structure, a major application of latent space models is to produce informative layouts for network visualization. The line between visualization and analysis can hence be quite thin, and – as emphasized by Freeman (2004) – innovations in data display are often linked to other developments within the network analytical field.

In addition to purely configural properties, network visualization may also include information on edge values and vertex attributes. Vertex size and shape may be varied to indicate individual attributes and/or structural properties, line width may be used to denote edge strength, and colour or form may be used to distinguish between nominally distinct edges or vertices. There are few, if any, 'standard' rules for such techniques at this time, although obvious visual motifs such as proportional scaling of vertex radii or surface area, or edge widths, based on attribute magnitudes are frequently encountered. General references on the display of quantitative data (Tufte, 1983) maybe useful sources of guidance on effective methods for supplementing purely structural displays.

Measurement and modelling of structural properties

Many of the most basic questions in the study of social networks involve the measurement and modelling of particular structural properties. We may ask, for instance, which individuals serve as bridges between otherwise disconnected groups, or whether a given network shows signs of being more centralized than would be expected by chance. Structural properties have been shown to be predictive of work satisfaction and team performance

(Bavelas & Barrett, 1951), power and influence (Brass, 1984), success in bargaining and competitive settings (Burt, 1992; Willer, 1999), mental health outcomes (Kadushin, 1982), and a range of other phenomena; such investigations hinge on the ability to systematically measure the properties of social structure in a manner which facilitates modelling and comparison. Here, we review a widely used approach to the measurement of structural properties – the use of structural indices – and describe a range of measures that are frequently encountered in the network literature. We also consider basic methods for the testing of structural hypotheses, which can be used where classical procedures are not applicable. Finally, we briefly review one approach to the modelling of network structure, and describe its use in inferring underlying structural influences from cross-sectional data.

Structural indices

Upon obtaining network data, the analyst is immediately faced with a non-trivial problem: how can one extract interpretable, substantively useful information from what may be a large and complex social structure? Simple visualization of network data can be illuminating, but it is not sufficiently precise to serve as an adequate basis for scientific work. Rather, we require a means of specifying particular structural properties to be examined, quantifying those properties in a systematic way, and (ultimately) comparing those properties against some baseline model or null hypothesis. The oldest and most common paradigm for accomplishing these goals is what may be called the structural index approach. The basis of this paradigm is the development of descriptive indices – real-valued functions of graphs – which quantify the presence or absence of particular structural features. These indices may describe structure which is local to a particular entity (or group thereof), or may measure structural features of the network as a whole. Similarly, indices may be designed to be interpreted 'marginally' (i.e. as expressing the total incidence of some structural feature) or 'conditionally' (i.e. as expressing the relative incidence of some feature *vs* a 'baseline' determined by other features such as size or density). In addition to direct interpretation, structural indices may be used as covariates in statistical models, and are sometimes used as dependent variables (although, as we shall see, this is not always unproblematic). They can also serve as the 'building blocks' for more elaborate network models, such as the discrete exponential families which will be discussed below. Before considering modelling applications, then, we review some of the primary classes of structural indices, and highlight some of the most commonly used members of each class. Modelling and hypothesis testing for these indices will be discussed in the sections which follow.

Node-level indices. A frequent objective of social network analysis is the characterization of the properties of individual positions. We may seek to identify, for instance, persons in positions of prominence, or whose positions facilitate actions such as information dissemination. Alternately, we may also be interested in the social environment faced by a given individual, measuring features such as the extent to which his or her local environment is socially cohesive, or the diversity of his or her personal contacts. Such properties are generally summarized by means of node-level indices, real-valued functions which – for a given graph and vertex – express some feature of network structure which is local to the specified vertex. We may denote a node-level index (or NLI) by a function f such that $f(v, G)$ returns the value of the specified index at vertex v, within graph G. NLI are fairly well developed within the network literature, and a wide range of such indices exists. Here, we shall review two of the most common categories: centrality indices, and ego-network indices. As we shall see, there is much overlap between these two classes of NLI; we treat ego-network indices separately, however, because of their growing importance in survey research.

Centrality indices: The oldest and best-known descriptive indices within network analysis are those designed to capture the extent to which one vertex occupies a more central position than another (in any of several senses). There are many distinct notions of centrality, leading to a proliferation of measures – here, we focus on four of the most widely used. The first three of these were treated in Freeman's (1979) famous paper on centrality indices, which itself was a consolidation of previous work on the subject. We also add an additional measure (usually credited to Bonacich (1972), but also a refinement of existing indices) which is widely used in many applications.

The most basic centrality index is degree, defined in the undirected case as the size of the neighbourhood of the focal vertex. Formally $c_d(v, G) \equiv |N(v)|$. In the directed case, three notions of degree are generally encountered: outdegree $\left(c_{d^+}(v, G) \equiv \left|N^+(v)\right|\right)$; indegree $\left(c_{d^-}(v, G) \equiv \left|N^-(v)\right|\right)$; and total or 'Freeman' degree $\left(c_{d'}(v, G) \equiv c_{d^+}(v, G) + c_{d^-}(v, G)\right)$. There is, in fact, a fourth notion of degree corresponding to the degree of the focal vertex in G's underlying semigraph, specifically, $|N^+(v) \cup N^-(v)|$, but this does not seem to be explicitly named within the network literature. As this measure is equal to the total number of alters involved in any manner with v, it is nevertheless a useful tool in the analyst's arsenal. Regardless of their variations, the degree measures all capture the number of partners of v, and thus tend to serve as proxies for activity and/or involvement in the relation. In practice, degree also correlates strongly with most other measures of centrality, making it a powerful summary index. As degree is easily sampled and fairly robust to error (Borgatti, Carley, & Krackhardt, 2006) and missing data (Costenbader &

Valente, 2003), it is also a favoured index for use under adverse conditions. The counts of the number of vertices having degree 0, 1, . . . , $n-1$ (respectively) collectively comprise the degree distribution. Degree distributions have generated intense interest in recent years as easily modelled signatures for hypothetical network formation processes (Barabási & Albert, 1999; Ebel, Mielsch, & Bornholdt, 2002); we will revisit them briefly under the section on graph-level indices.

The second of the three 'classic' indices of Freeman (1979) is known as betweenness. As its name implies, betweenness quantifies the extent to which the focal vertex lies on a large number of shortest paths between various third parties; high-betweenness individuals thus tend to act as 'boundary spanners', bridging groups which are otherwise distantly connected, if at all. Formally, betweenness is defined in the directed case as $c_b(v, G) \equiv \Sigma_{(v',v'')\subset V/v} \frac{g'(v', v, v'', G)}{g(v', v'', G)}$, where $g(v, v', G)$ is the number of (v, v') geodesics in G, $g(v, v', v'', G)$ is the number of (v, v'') geodesics in G containing v', and $\frac{g'(v', v, v'', G)}{g(v', v'', G)}$ is taken equal to 0 where $g(v', v'', G) = 0$. Thus, betweenness considers only shortest paths, and weights paths inversely by their redundancy. (The stress centrality of Shimbel (1953) can be used where one seeks an index which is identical to betweenness, save in relaxing this latter condition.) As betweenness is based on the path structure of the graph, it is a truly global index.[3] Unfortunately, this means that it will be fairly non-robust to error and missing data in certain settings, and that it cannot be sampled from local network data (see, however, Borgatti *et al*., 2006 and Everett & Borgatti, 2005 for a counterpoint and some pragmatic approximations). Betweenness is also fairly expensive to compute, although algorithms such as those of Brandes (2001) produce reasonable performance on sparse networks. Despite these drawbacks, betweenness is a widely used measure, and is frequently invoked as an example of a positional property which cannot be reduced to simple local structural features.

The third 'classic' centrality measure is closeness, which captures the extent to which the focal vertex has short paths to all other vertices within the graph. In its standard formulation, $C_c(v, G) \equiv \frac{n-1}{\Sigma_{v' \in V} d(v, v')}$, where $d(v, v')$ is the geodesic distance from vertex v to vertex v'. Closeness is ill-defined on graphs which are not strongly connected, unless distances between disconnected vertices are taken to be infinite. In this case, $C_c(v, G) = 0$ for any v lacking a path to any vertex and, hence, all closeness scores will be 0 for graphs having multiple weak components. This rather unsatisfactory state of affairs greatly limits the utility of closeness in practical settings and, indeed, the index is much less widely used than betweenness or degree. (Some obvious alternatives to Freeman's closeness, such as $\frac{\Sigma_{v' \in V \setminus v} d(v, v')^{-1}}{n-1}$, avoid this problem. It is unclear why these measures remain largely unutilized.) Despite its limitations, closeness is useful in identifying vertices which can quickly reach others within a given network, and/or which can be quickly reached (in the undirected case). As maximum closeness vertices typically are (or are close to) vertices of minimum eccentricity (i.e. maximum distance from all other vertices), they correspond closely to intuitive notions of being in the 'middle' of the graph; indeed, vertices of minimum eccentricity are known as graph centres, and such vertices may be approximately identified using closeness scores. The closely related graph centrality of Hage and Harary (1995), based on inverse eccentricity, provides an exact identification.

The last centrality index to be presented here does not belong to the three 'classic' measures of betweenness, closeness, and degree, but is nevertheless of great importance for structural analysis. This is particularly true because of its surprising ubiquity: it arises from many different motivating arguments, and admits a number of seemingly distinct interpretations. The measure in question is the eigenvector centrality, defined by the principal solution to the linear equation system

$$\lambda \mathbf{c_e} = \mathbf{Y} \mathbf{c_e}, \tag{1}$$

where $\mathbf{c_e}$ is the vector of centrality scores, $\mathbf{Y}$ is the adjacency matrix of G, and λ is a scaling coefficient. Where the principal solution to Equation 1 is used, λ is equal to the first eigenvalue of $\mathbf{Y}$, and $\mathbf{c_e}$ is the corresponding eigenvector. Hence, $c_e(v, G)$ is v's score on the first eigenvector of G's adjacency matrix (whence comes the name of the index). The somewhat obscure meaning of these scores is elucidated by writing Equation 1 in another form:

$$c_e(v_i, G) = \frac{1}{\lambda} \sum_{j=1}^{n} Y_{ij} c_e(v_j, G). \tag{2}$$

Thus, we can see from Equation 2 that eigenvector centrality can be interpreted recursively as positing that the centrality of each vertex is equal to the sum of the centralities of its neighbours, attenuated by a scaling constant (λ). We might summarize this idea by the intuition that 'central vertices are those with many central neighbours.' As this is true of the neighbours, in turn, we can envision eigenvector centrality as reflecting the equilibrium outcome of a social process in which each individual sends some quantity (status, power, information, wealth etc.) to each of his or her neighbours, that quantity being determined by his or her current total (dependent upon incoming transfers from his or her neighbours) and an 'attenuation' effect. This can also be seen by writing the measure in terms of its series expansion:

$$c_e(v_i, G) = \sum_{\ell=1}^{\infty}\left[\left(\frac{1}{\lambda}\right)^{\ell}\sum_{j=1}^{N} Y_{ij}^{\ell}\right], \quad (3)$$

where $\mathbf{Y}^{\ell}$ is the ℓth power of $\mathbf{Y}$. As Y_{ij}^{ℓ} is equal to the number of walks of length ℓ from v_i to v_j, it follows that c_e composes v_i's centrality from the sum of its walks to other vertices, weighting those walks inversely by their length (via λ). As this implies, vertices are high on eigenvector centrality when they have many short paths to many other vertices in the network, whether or not those paths are necessarily geodesics. The simplest way to obtain such a state is to be deeply embedded in a large, dense cluster and, indeed, positions of this kind have the highest c_e scores. This can be taken yet farther by considering a simple core-periphery model of social interaction (Borgatti & Everett, 1999), in which we posit that the expected value of an interaction between any given pair v_i and v_j satisfies $\mathbf{E}\boldsymbol{Y}_{ij} \propto \beta_i\beta_j$ for some non-negative 'coreness' measure, β. The behaviour of this model is both simple and intuitive: high-coreness individuals are likely to have strong interactions with each other (high $\beta_i \times$ high β_j leads to high $\mathbf{E}\boldsymbol{Y}_{ij}$); high coreness individuals are likely to have only weak interactions with low-coreness individuals (high $\beta_i \times$ low β_j leads to low/medium $\mathbf{E}\boldsymbol{Y}_{ij}$); and low-coreness individuals are unlikely to have much interaction with each other at all (low $\beta_i \times$ low β_j leads to extremely low $\mathbf{E}\boldsymbol{Y}_{ij}$). Surprisingly, the optimal 'coreness' measure under this model (in a least squares sense) turns out to be eigenvector centrality – setting $\beta = \mathbf{c_e}$ minimizes the squared error between $\beta\beta^T$ and $\mathbf{Y}$. This means that eigenvector centrality is a core-periphery measure, in addition to its other interpretations. Furthermore, it is a well-known result of linear algebra (Strang, 1988) that $\lambda\mathbf{c_e}\mathbf{c_e}^T$ (where λ and $\mathbf{c_e}$ are the first eigenvalue/eigenvector pair of $\mathbf{Y}$) is the best one-dimensional approximation of $\mathbf{Y}$ in the least squares sense. Thus, eigenvector centrality also provides a set of scores which (in one sense, at least) best summarizes the entire structure of the network as a whole. These rather remarkable results demonstrate the deep connections between node-level concepts of centrality, global features such as core-periphery structure, structural summaries and dimension reduction, and social processes such as diffusion and influence. Eigenvector centrality turns up at the centre of many of these connections and, as such, is an index of great theoretical and methodological significance. (See Bonacich (1972), Seary and Richards (2003), and Baltz amd Kloemann (2005) for further discussion.)

Ego network indices: One family of node-level indices whose importance has grown in recent decades is that of measures for egocentric network (or 'ego net') properties. As mentioned above, the egocentric network of vertex v in graph G is defined to be $G[v \cup N(v)]$ (i.e. the subgraph of G induced by v together with its neighbourhood in G). v's ego net thus captures the local structural environment of v, in the sense of v's alters and any edges between them. (In some studies, a distinction is made between v's personal network, or local neighbours, and its 'complete' ego network as defined above. Our discussion here is concerned with the latter case.) Following this, an ego network index is formally defined as any function $f:(v,G)\mapsto\mathbb{R}$ such that $f(v,G') = f(v, G[v\cup N(v)])\ \forall v, G': G'[v\cup N(v)] = G[v\cup N(v)]$. Put less formally, an ego network index is a node-level index that depends only on v's ego net. This property is not only a defining condition for the ego network indices, but also accounts for their popularity: because these indices depend only on local structure, they can be used in settings for which only local network information is available. The classic example of such a setting is a conventional survey, in which an instrument is administered to members of a sample drawn from a larger population. Although reconstruction of complete networks is generally impossible in this case, respondents can be asked to provide information on their alters, as well as ties among those alters. The result of this elicitation scheme (introduced earlier in the context of complete egocentric sampling designs) is a collection of ego nets drawn from the larger network, which can, in turn, be studied using egocentric network indices. Given the widespread popularity of survey methods (and the great investment in infrastructure for such research), ego net studies have emerged as a popular means of integrating network measures into population research. Although very limited in scope, ego network indices thus play an important role in modern network research.

While it is obviously impossible to enumerate all members of the family of ego network indices, a number of frequently used measures are worth noting. The most popular index is one which has already been mentioned: degree. In addition to being an ego network index in its own right, degree also appears in the form of ego network size (often incorrectly shortened to 'network size') which is equal to one plus the degree of v (i.e. the number of vertices in v's ego net). Local cohesion is often measured by ego network density, which is generally defined as $|E(G[N(v)])|\binom{|N(v)|}{2}^{-1}$ in the undirected case and $2|E(G[N(v)])|\binom{|N(v)|}{2}^{-1}$ in the directed case. Somewhat confusingly, this definition excludes ties involving ego from the computation; the alternative measures $|E(G[v\cup N(v)])|\binom{|N(v)|+1}{2}^{-1}$ (undirected) and $2|E(G[v\cup N(v)])|\binom{|N(v)|+1}{2}^{-1}$ (directed) are sometimes used, and it is important to clear which version is used when interpreting the measure. Another useful index is local bridgeness (also referred to by Gould & Fernandez

(1989) as the total brokerage score), which measures the extent to which ego is a local mediator for ties among his or her alters. Specifically, the local bridgeness of v is the number of v', v'' pairs such that (v', v) $(v, v'') \in E$ and $(v', v'') \notin E$. In the undirected case, this happens to take the simple form $\binom{|N(v)|}{2} - |E(G[N(v)])|$, which highlights the measure's connection with both ego net size and ego net density. Gould and Fernandez (1989) further decompose the bridgeness/brokerage score based on nodal covariates, allowing for distinctions to be drawn regarding the specific types of brokerage in which v is implicated. This approach of combining local structural measures with nodal covariates has proven useful in a range of substantive settings, and is a common strategy within ego net research. A related family of indices due to Burt (1992) incorporates edge values to capture various aspects of local network structure related to brokerage and exclusion opportunities; these indices (stemming from Burt's popular 'structural holes' paradigm) have been widely used in organizational contexts.

In addition to these measures, it should be noted that almost all graph-level indices (which are discussed below) can be adapted to serve as egocentric network measures by restricting their computation to v's ego net. Formally, for graph-level index f, we can construct the ego net index f^* via the definition $f^*(v, G) \equiv f(G[v \cup N(v)])$. While such measures can be useful, it is important to remember that their behaviours will be constrained by the peculiar properties shared by all egocentric networks. For instance, all egocentric networks are connected with diameter less than or equal to two, contain at least one spanning star, and have a minimum density of $(|N(v)| + 1)^{-1}$ (under the 'alternate' measure in which ego is not excluded). These properties are artifacts of the manner in which ego nets are defined, and can affect otherwise familiar graph level indices in complex ways; comparison of graph-level indices (GLI) scores derived from ego nets with those derived from other networks is thus inappropriate in most cases. The same caveat applies to the use of conventional node-level indices on vertices within another's ego network: as only a constrained, typically biased sample of edges from such vertices are observed (much less higher order properties such as paths), alters' NLI within an ego network are not generally reflective of their NLI in the larger network structure. Researchers seeking to properly compare the structural properties of adjacent vertices are thus well advised to avoid egocentric network data in favour of more complete alternatives.

Graph-level indices. While node-level indices describe structure which is local to a particular vertex, GLI quantify structural properties of the network as a whole. Although such measures are especially important when comparing networks, they are also useful for determining the large-scale structural context in which behaviour occurs. GLI are extensively used in the modelling of network structure, where they serve to provide structural signatures for underlying dependencies among edges. By observing the particular pattern of GLI scores associated with a given network, it is thus possible in some cases to infer properties of the social process which gave rise to it; examination of such process/feature connections is an area of active theoretical research (Pattison & Robins, 2002; Robins, Pattison, & Woolcock, 2005).

Formally, a graph-level index is a real-valued function, f, such that $f(G)$ is the value of the index for graph G. There are many types of graph-level indices, measuring everything from counts of particular structural configurations to concentration of node-level features. Here, we review several major categories of GLI, along with well-known or otherwise instructive examples from each category. Later, we will see how these indices may be used in contexts such as network modelling and graph comparison.

Subgraph census statistics: An essential building block of graph-level analysis is the subgraph census statistic. Such statistics are defined as follows.[4] As usual, let $G = (V, E)$ be a graph on n vertices, and let H be a graph on $n' \leq n$ vertices. Let $S = \{s_1, s_2, \ldots\}$ be the set of all subsets of V having size n'. Then, the H-census statistic on G is $|\{s \in S: H \simeq G[s]\}|$ (i.e. the number of induced subgraphs of size n' which are isomorphic to H). This, in turn, is simply the number of copies of H which can be found in G. While it is possible to construct census statistics from any H, certain cases have particular importance within the existing literature. Chief among these are sets of census statistics corresponding to each of the isomorphism classes on the set of order-n' graphs. For instance, consider the case when $n' = 2$ – the order-two subgraphs, or dyads – and G is undirected. There are then two possible values of H: the empty or null dyad (two vertices without an edge); and the complete dyad (two vertices with an edge). The corresponding dyad census statistics for these graphs are the edge count of G and the 'hole count', or number of vertex pairs which are non-adjacent. (Clearly, the number of non-adjacent pairs is equal to $\binom{n}{2}$ minus the number of edges.) A slightly more interesting set of statistics arises when G is directed. In this instance, there are three possible forms which can be taken by H: the null dyad; the asymmetric dyad (two vertices with one edge between them); and the complete or mutual dyad (here, two vertices with two directed edges between them). Note that while there are two ways to draw the asymmetric dyad, each is isomorphic to the other; thus, the two forms are grouped together into one isomorphism class. Given the above, the directed dyad census of G consists of the numbers of mutual, asymmetric, and null dyads. These counts are conventionally indicated by the letters M, A, and

N, respectively. The dyad census is used to form many other measures of social structure, as described below.

Dyad census statistics reflect structural properties which are limited to the interactions among two individuals; the corresponding sets of statistics for sets of three individuals are those arising from the triad census. For G undirected, there are four H configurations which can potentially by observed, each determined entirely by the number of edges present (0–3 inclusive). Thus, the triad census of an undirected graph, G, consists of the counts of triads with 0, 1, 2, and 3 edges (respectively). This same simplicity, alas, does not hold in the directed case. There are 16 isomorphism classes for the directed triads, conventionally described (following Davis & Leinhardt, 1972) by their respective dyad census statistics, together with an extra letter designating orientation. The 16 numbers corresponding to census statistics for each of these isomorphism classes jointly constitute the directed triad census for G, and convey important information regarding local network structure. For instance, the related notions of transitivity (Holland & Leinhardt, 1972) and local clustering (Watts & Strogatz, 1998) can both be expressed in terms of the frequency of triadic configurations. In its most common form, the transitivity of a graph is the fraction of ordered (i, j, k) triads such that (i, j) and (j, k) are adjacent, for which i is adjacent to k. This quantity can be written as a function of the triad census using the weighting vector method described by Wasserman and Faust (1994, p. 574).

Beyond dyad and triad census statistics, the field becomes more ad hoc. The large number of tetradic isomorphism classes makes a complete enumeration unattractive, a problem which continues to worsen for larger vertex sets. Subclasses of census statistics which are sometimes used include the cycle census statistics (counts of cycles of specified length), and clique census statistics (counts of complete subgraphs of specified size). A statistically important family of census statistics is that of the k-stars (Frank & Strauss, 1986), which measure the number of configurations in which one vertex is adjacent to k others. k-stars exhibit a nested structure, in which every k-star necessarily contains $\binom{k}{k-1}$ k-1-stars; this creates strong dependence among k-star statistics. Interestingly, the complete k-star census exhibits a 1:1 relationship with the degree distribution. If $d_0, \ldots, d_{n-1}$ is the number of vertices with $0, \ldots, n-1$ edges (respectively) within G, then G contains $\Sigma_{i=k}^{n-1} d_i \binom{i}{k}$ k-stars. Obtaining the degree distribution from the k-star census is more complex, but can be accomplished by the recursion:

$$d_i = s_i - s_{i+1} + \sum_{j=1}^{n-1-i} d_{i+j} \binom{i+j}{i} \left(\frac{j}{i+1} - 1\right), \qquad (4)$$

where $s_1, \ldots, s_{n-1}$ are the k-star statistics of G, $d_{n-1} = s_{n-1}$, and $d_0 = n - \Sigma_{i=1}^{n-1} d_i$. Where G is directed, the k-star statistics are generalized into k-instars, k-outstars, and various mixed star configurations. These statistics collectively describe the joint indegree and outdegree distributions of G; due to the enumerative complexity of these statistics, they will not be discussed in detail here.

In addition to their use in modelling (which will be described presently), subgraph census statistics are important building blocks of other structural indices. For instance, network density (the ratio of observed to potential edges within a graph) can be written $M/(M + N)$ in the undirected case, or $(M + A/2)/(M + A + N)$ in the directed case. Another important family of measures based on the dyad census are the reciprocity measures, which will be discussed in detail below.

Centralization indices: One standard family of graph-level indices consists of those which measure the extent to which centrality is concentrated within a small number of vertices; these are known, appropriately enough, as centralization indices. The most commonly used of such indices are those belonging to the family introduced by Freeman (1979), which take the following form:

$$C(G) \equiv \sum_{i=1}^{n} \left[\left(\max_j c(j, G) \right) - c(i, G) \right] \qquad (5)$$

where c is a centrality index. Thus, C quantifies the difference between the centrality of the most central vertex and the centralities of all other vertices in the graph. This index clearly depends on graph size, and it is common to work with the corresponding family of normalized centralization indices,

$$C'(G) \equiv \frac{C(G)}{\max_{G' \in \mathcal{G}_n} C(G')} \qquad (6)$$

where $\mathcal{G}_n$ is the set of order-n graphs. The normalized measures vary from 0 to 1, and do not have an obvious dependence on n. Appearances can deceiving, however, as C' may still depend indirectly on graph size where the corresponding centrality measure is, in some way, size dependent. C' can also be constrained by network density, or other properties; for instance, Butts (2006b) has demonstrated that the range of possible degree centralization scores is approximately $[0, 1 - d]$ at density d, for large n.

Interestingly, it is not necessary to measure the entire centrality distribution to compute the Freeman centralization of a graph. From Equation 5,

$$C(G) = n \left[\frac{1}{n} \sum_{i=1}^{n} \left[\left(\max_j c(j, G) \right) - c(i, G) \right] \right] \qquad (7)$$

$$= n \left[\left(\max_j c(j, G) \right) - \frac{1}{n} \sum_{i=1}^{n} c(i, G) \right]. \qquad (8)$$

Thus, $C(G)/n$ is equal to the difference between the maximum observed centrality score and the average centrality of all vertices. For centralities which can be computed from sampled local network information (e.g. degree), this suggests that an estimator of the form $\hat{C}(G) = n(c_{\max} - \bar{c})$ (with $c_{\max}$ and $\bar{c}$ being the sampled maximum and mean centrality scores, respectively) may provide a reasonable approximation to $C(G)$ where c is not too heavily right-skewed.

One attractive feature of the Freeman centralization measures is that they obtain their maximum values under the star graph for most known centralities. Likewise, Freeman centralization is always zero for a graph in which all vertices are automorphically equivalent (e.g. a complete or empty graph). This provides a fairly strong intuition regarding the types of graphs which will be highly centralized (or decentralized), at least at the extremes. It should be noted, however, that the former condition is not true for all centrality measures. For instance, the graph which is of maximum centralization under eigenvector centrality is that composed of a single dyad together with $n - 2$ isolates. When applying C to a new centrality measure, then, it is important to verify that the maximum centralization actually occurs on the star graph before using the star graph centralization as the denominator for Equation 6.

Although Freeman's C is the most widely used measure of its kind, others have been proposed. Snijders (1981) proposes the variance of the degree distribution as a measure of centralization in that context, although (as he notes) this is really a measure of heterogeneity rather than centralization per se. Traditional upper-tail concentration measures, such as the Gini index, are also natural candidates for centralization indices. Inasmuch as these alternatives are somewhat less dependent on the extreme upper quantile of the centrality distribution, they may be more robust to measurement error than the Freeman measures. Thus far, however, most workers in the field have favoured the simplicity and intuitive power of the latter option.

Hierarchy and symmetry indices: Although frequently confused with centralization, hierarchy is a distinct and important structural phenomenon. While centralization is founded upon the notion of concentration (specifically, that some individuals are more central than others), hierarchy is based upon the notion of asymmetry. As such, hierarchy is only well defined within a directed context. When considering very local (i.e. dyadic) structure, hierarchy is more often encountered via the inverse concept of reciprocity. Reciprocity (the tendency of ties to be reciprocal rather than unidirectional) is measured in a number of ways, all of which can be computed from the dyad census. The simplest measure of reciprocity is the fraction of reciprocal dyads (here denoted r_1), which is given by $r_1(G) \equiv (M + N)/(M + A + N)$ in *MAN* dyad census notation. r_1 is a global measure of symmetry, and has the attractive property that $r_1(G) = r_1(\bar{G})$; however, r_1 does not distinguish between graphs which are symmetric due to having many reciprocated edges, versus graphs which are extremely sparse (and therefore contain many null dyads). One measure which does make such a distinction is the fraction of symmetric non-null dyads, or $r_2(G) \equiv M/(M + A)$, although this does not lead to a very natural interpretation. A more natural index is the fraction of reciprocated edges, or $r_3(G) \equiv M/(M + A/2)$, which can be thought of as the probability that a randomly selected edge within the graph will be reciprocated. While r_3 is very intuitive, it is still important to evaluate it against a known baseline, such as the background density of the graph. An example of a slightly more sophisticated index with such properties is $r_4(G) \equiv \ln \frac{M(M + A + N)}{(M + A/2)^2}$, or the logged relative risk of a reciprocating edge versus the baseline risk. Note that, with the exception of r_1, these measures are not well defined on empty graphs; empty graphs are generally taken to be fully reciprocal by definition, but this convention is not universally accepted.

Clearly, the r measures are measures of reciprocity; each is dual, however, to a measure of hierarchy. With the exception of r_4, hierarchy can be measured by $h_i(G) = 1 - r_i(G)$, translating (respectively) to the fraction of asymmetric dyads, the fraction of asymmetric non-null dyads, and the fraction of unreciprocated edges. In the case of r_4, some adjustment is necessary – the natural parallel is the logged relative risk of an unreciprocated edge, versus the corresponding baseline. This change leads to the corresponding index $h_4(G) \equiv \ln \frac{A(M + A + N)}{(2M + A)(A/2 + N)}$. As with the r indices, the h measures are local, and depend only on the dyad census. This makes them easy to estimate where G has been sampled (Frank, 1978), and relatively robust to measurement error. However, there are other aspects of hierarchy which cannot be captured via dyadic structure alone.

Beyond the local hierarchy measures derived from the dyad census, researchers have defined a number of global measures for quantifying asymmetry. Possibly the simplest of these is given by Krackhardt (1994), whose hierarchy measure is equal to the fraction of weakly connected dyads which are not strongly connected. Formally, we may express this measure in terms of the reachability graph of G, which is defined as the digraph $R = (V(G), E')$ such that $(v, v') \in E'$ iff there exists a path from v to v' in G. If R is the reachability graph of G, then Krackhardt's hierarchy measure is given by $h_2(R)$; intuitively, this corresponds to the fraction of pairs who can interact at some distance, but for whom this capacity to interact is not mutual. A more complex measure is given by Hummon and Fararo (1995), whose hierarchy index generalizes the notion of 'level'. Consider a simplified hierarchical structure, in which we have $v_1 \rightarrow v_2$, $v_2 \rightarrow v_3$, . . . $v_{n-1} \rightarrow v_n$ and no other edges.

Such a structure is said to have n levels, each level consisting of a position which sends an edge to the one immediately below it (for levels above the last) and receives an edge from the one immediately above it (for levels below the first). Such a strict case could be generalized by allowing each position to consist not only of a single vertex, but rather a set of vertices which are mutually reachable from one another (i.e. which are strongly connected). In this case, we can think of the levels as forming a partial rank structure on the graph, such that $v \leq v'$ iff G contains a (v', v) path. The more levels within the graph, the finer the ranking distinctions which it admits. Of course, real structures may not decompose neatly into levels: there may be multiple 'chains' of strong components which are asymmetrically connected. Hummon and Fararo's hierarchy measure deals with this by considering the finest range of rank-order distinctions which can be made using the given structure. Specifically, let G' be the graph minor formed by condensing the strong components of G into single vertices. Clearly, G' contains no strong component of size greater than 1 (as, if so, it could be further reduced); thus, the vertices of G' are asymmetrically connected. The Hummon-Fararo hierarchy of G is then the longest path in G'. To the extent that G approximates a 'clean', multilevel structure, the H-F hierarchy will approach $n - 1$. At the opposite extreme, in which G is strongly connected, the H-F hierarchy is equal to 0. The H-F hierarchy thus goes beyond the mere extent of local or global asymmetry, quantifying the extent to which that asymmetry is linearly organized. (The relative incidence of transitive versus cyclic triads (mentioned above) can be used in a similar fashion.)

Connectivity indices: A final class of indices we shall consider are those which describe the connectivity properties of a network (i.e. the extent to which the individuals within the network can reach one another via direct or indirect connections). Density, which we have already seen, can be thought of as the most primitive index of this form: as density can be interpreted as the marginal probability of an edge from any given vertex v to some other vertex v', it is necessarily a measure of local connectivity. However, density per se does not tell us about non-local connections between vertices, and is thus not a very satisfying index in this regard. Various alternatives have been developed which provide a more refined view of network connectivity, and we consider several of these here.

At the opposite extreme from density, one obvious connectivity index is the number of components in a graph. As there are four basic component types in the directed case (weak, unilateral, strong, and recursive), four such counts are possible for a given digraph (*vs* one in the undirected case). Intuitively, the more components within a given graph, the less well connected the associated network; normalizing by n to obtain the number of components per vertex gives a less order-dependent measure of fragmentation. To map this measure to the [0, 1] interval, a connectivity index such as $(n - K(G))/(n - 1)$ (where $K(G)$ is the number of components of G, and $n \geq 2$) may prove useful. This index is equal to 1 in the fully connected case (i.e. G has one component) and takes a value of 0 when G is fully disconnected (i.e. G is composed entirely of isolates). Although global in character, this index has the disadvantage of not permitting fine distinctions regarding degrees of connectivity, especially in small groups. For this purpose, it may be useful to consider the fraction of dyads which can reach one another by some criterion or another. This is the intuition behind Krackhardt's (1994) connectedness index, which is equal to the fraction of weakly connected dyads in G. When Krackhardt's connectedness is equal to 0, no vertex can reach any other via a semipath in the underlying network; as the number of pairs which are connected by semipaths increases, the measure approaches 1. While this index is more refined than the simple connectivity index described above, it is still unable to distinguish among weakly connected graphs (all of which have Krackhardt connectedness scores of 1). A simple modification of Krackhardt's index for directed graphs would thus be to consider the fraction of vertex pairs that are unilaterally, strongly, or recursively connected in G. By using a more stringent definition of connectedness, it is possible to distinguish between levels of connectivity even among weakly connected digraphs.

Yet another approach to connectivity comes from the notion of cutsets. A subgraph $H \subset G$ is said to be a cutset of G if removing H increases the number of components in G. A vertex v which is a cutset for G is said to be a cut vertex (or cut point) of G, and an edge which is a cutset for G is similarly known as a cut edge. (Note that when a vertex is removed, all of its associated edges are removed as well – this is not the case when removing edges, whose end-points are left intact.) Intuitively, we may think of a graph as being better connected when it takes the removal of many elements to break it into smaller components. Such graphs are also said to be 'robust', an expression which highlights the fact that the potential for communication among elements in such networks is resistant to disruption via the failure of individual network elements (see Klau & Weiskircher, 2005 for an in-depth review). The extent to which a graph exhibits such robustness may be measured by the sizes of its minimum edge or vertex cut (i.e. the minimum number of edges or vertices, respectively, needed to increase the number of components in G). These numbers are, respectively, known as the edge and vertex connectivities of G, and can be considered graph-level connectivity indices. Conventionally, a graph is said to be k-connected if its minimum vertex cut is of size k, with higher values of k clearly indicating more robust (and better connected) networks. In the undirected case, it is

known that $k \geq h$ if and only if G contains at least h spanning cycles (Berge, 1962), and connectivity is thus related to other structural properties such as the incidence of long-range cycles. When applied to subgraphs (rather than to the graph as a whole), connectivity has also been taken to be an indicator of cohesion (Moody & White, 2003); the concept has thus proved to be useful at multiple levels of analysis.

Conditional uniform graph tests

In evaluating graph-level indices, it is frequently useful to compare observed index values against those which would be obtained by a baseline model with known substantive properties (see Mayhew, 1984a, b, for a forceful articulation of the baseline modelling approach). By noting the extent and direction of deviation of indices from their baseline distributions, we may detect the presence of structural biases within the networks under study; these, in turn, may provide useful clues regarding the mechanisms underlying the data in question. One important family of baseline models for network data is the family of conditional uniform graph (CUG) distributions. A CUG distribution may defined as follows. Let $\mathbb{G}$ be the set of all graphs, let $\mathbf{t} = (t_1, \ldots, t_n)$ be a tuple of real-valued functions on $\mathbb{G}$, let $\mathbf{x} \in \mathbb{R}^n$ be a known vector, and let $I_A(x)$ be an indicator function returning 1 if $x \in A$ and 0 otherwise. Then the distribution

$$\Pr(G = g|\mathbf{t}, \mathbf{x}) = [|\{g' \in \mathbb{G} : \mathbf{t}(g') = \mathbf{x}\}|]^{-1} I_{\{g' \in \mathbb{G}:\mathbf{t}(g')=\mathbf{x}\}}(g) \quad (9)$$

is said to be the conditional uniform graph distribution with sufficient statistic $\mathbf{t}$ taking value $\mathbf{x}$. As Equation 9 implies, the CUG distribution fixes certain properties of G (specified by $\mathbf{t}$) at particular values (specified by $\mathbf{x}$), and treats all graphs meeting those criteria as equally probable. CUG distributions are among the oldest and most widely used models for network data, and are used for their simplicity as well as for their statistical properties.

One the simplest families of CUG distributions is the family of order-conditioned uniform graphs. These distributions are defined by setting $\mathbf{t} = (|V|)$ and, hence, treat all graphs of a specific size as equiprobable. Although mathematically interesting, these models are generally very poor approximations of social network structure and, as such, are of limited scientific value. A slightly more sophisticated model is the so-called '*N*, *m*' family popularized by Erdös and Rényi (1960), which is defined by setting $\mathbf{t} = (|V|, |E|)$. This model conditions on both size and density, and is a rather better approximation to real-world networks (which tend to be fairly sparse). Other familiar models include the *U|MAN* family (which conditions on the dyad census; see Holland & Leinhardt, 1975), and the family of degree-conditioned uniform random graphs (Snijders, 1991). The former model family builds on the *N*, *m* model by capturing biases towards or away from reciprocity (a very important effect in real-world networks), while the latter allows for features such as excess degree centralization which are frequently encountered in social settings. We note that while CUG distributions need not condition on graph size, all distributions currently in active use do so. It should also be noted that the distribution of Equation 9 is only well defined where there exists $G \in \mathbb{G}$ such that $\mathbf{t}(G) = \mathbf{x}$. Careless choice of conditioning statistics may result in distributions that are degenerate (admitting only one isomorphism class), and/or ill-defined (admitting no graphs at all).

While conditional uniform graph distributions are used for a number of purposes (including baselines for simulation studies, and minimally informative priors for Bayesian analysis (Butts, 2003)), one of the most important is the conditional uniform graph test (or CUG test) procedure.[5] Formally, the CUG test is a test of the hypothesis that an observed statistic, $s(g)$, was drawn from the distribution of s arising from the CUG distribution specified by $\mathbf{t}$, $\mathbf{x}$. Such hypotheses are generally one-sided; the p-value for the upper tail test is then $\Pr(s(G) \geq s(g)|\mathbf{t}, (\mathbf{x})$, with $\Pr(s(G) \leq s(g)|\mathbf{t}, (\mathbf{x})$ providing the p-value for the corresponding lower tail test. Frequently, the value of x used is that associated with the observed graph (i.e. $\mathbf{x} = \mathbf{t}(g)$). For instance, if one wanted to determine whether the degree of centralization of a given structure was greater than would be expected from its size and density alone, one might perform an upper tail CUG test of the centralization score against the *N*, *m* distribution (with *N* and *m* set to match their values in the observed graph). A low p-value for the associated test would suggest that the observed graph is more centralized than would be anticipated from its size and density and, hence, that some additional process or constraint might be at work. Further tests based on additional constraints (e.g. reciprocity, number of isolates etc.) could, in turn, be used to provide clues as to the nature of the bias giving rise to the high level of observed centralization. Indeed, the simultaneous use of tests against multiple (often nested) models is a powerful means of discriminating among competing explanations for the sources of structural biases, and is strongly recommended. A common strategy is to begin with a simple baseline (e.g. the order-conditioned model), experimenting with various constraints until one arrives at a minimal set of conditioning statistics which are sufficient to account for the observation in hand. These statistics are then used to localize the deviations from uniformity found within the observed graph. For a more detailed quantitative analysis of how these biases interact, it is generally necessary to turn to a more elaborate modelling strategy; we now proceed to a discussion of one such approach.

Exponential random graph models

As we have seen, the essential logic of the conditional uniform graph lies in evaluating the quantile of a structural statistic with respect to a baseline distribution on the set of possible structures. It is immediate to ask whether that extremity might be directly parameterized, rather than simply used to perform a dichotomous statistical decision (as in the case of null hypothesis tests). The affirmative answer to this question was provided by a line of work originating with Holland and Leinhardt (1981), and later extended by Frank and Strauss (1986), Wasserman and Pattison (1996), and others. Following the development of conditional uniform graph tests above, let $\mathbf{t}$ be a vector of sufficient statistics, and let $\mathcal{G} \subseteq \mathbb{G}$ be a countable graph set. We may then write a probability mass function (PMF) on $\mathcal{G}$ in the form

$$\Pr(G = g|\mathbf{t}, \theta) = \frac{\exp(\theta^T \mathbf{t}(g))}{\sum_{g' \in \mathcal{G}} \exp(\theta^T \mathbf{t}(g'))} I_{\mathcal{G}}(g), \qquad (10)$$

where $\theta \in \mathbb{R}^n$ is a known parameter vector and $I_{\mathcal{G}}$ is an indicator function for $\mathcal{G}$. Intuitively, Equation 10 expresses the probability of observing any particular graph as being proportional to an exponentiated linear predictor, itself a weighted combination of structural characteristics. Graphs with higher values of t_i thus become increasingly probable as $\theta_i \to \infty$, or (by turns) become less probable as $\theta_i \to -\infty$. In the special case of $\theta = \mathbf{0}$, $\mathbf{t}$ receives no weight, and the CUG distribution on $\mathcal{G}$ is recovered.

It should be emphasized that any probability distribution on $\mathcal{G}$ can be written in the form of Equation 10;[6] thus, the above is less a probability model than a method for parameterizing such models. More properly, Equation 10 describes a discrete exponential family of random graphs. Models written in this form are referred to more succinctly as exponential random graph (ERG) models, or (in older literature) '*p**' models. The fact that all existing graph distributions (including, as noted, the CUG families) can be written in exponential family form allows the ERG framework to serve as a 'lingua franca' for models of network structure per se; although there do exist extended models (e.g. networks with endogenous nodal covariates (Robins, Pattison, & Elliott, 2001)) which do not belong to this class, it is nevertheless broad enough to have wide utility in practice. Much of the value of this unifying framework lies in its facilitation of tasks such as estimation of structural biases or prediction of network properties. Given methods for performing such tasks in the general ERG case, application to specific modelling scenarios becomes (in principle) a simple matter of writing the new model in ERG form and using the method in question. In practice, matters are not always so simple; in particular, the computational difficulties associated with simulation and model fitting can be severe for certain subfamilies (Handcock, 2003), but the approach is broadly effective in many settings. Beyond these considerations, the large body of statistical literature on discrete exponential family models in other contexts aids in the development of new insights regarding the behaviour of network models. Important examples of such cross-application of findings from the statistical literature to the literature on network methods include work on the use of dependency graphs in constructing network models (Frank & Strauss, 1986; Pattison & Robins, 2002) and phenomena such as 'degeneracy' (Strauss, 1986; Handcock, 2003).

While the literature on exponential random graph methods is too large to be easily summarized here (see Wasserman & Robins, 2005 for a recent review), a few important points are worth mentioning. First, the ERG framework provides a natural way to extend the conditional uniform graph concept described earlier. Rather than comparing observed graph statistics to a CUG distribution, the parallel ERG approach involves fitting parameters corresponding to the statistics in question. These parameters are then inspected to determine the strength and direction of structural biases which are inferred to have given rise to the observed graph. As zero-valued parameters may always be interpreted as reflecting no (conditional) bias on the associated statistic, it follows that null hypothesis tests on the parameters may be used in much the same manner as CUG tests. Unlike CUG tests, however, ERG modelling allows for the evaluation of a wider range of hypotheses (including those interactions between biases on multiple statistics). A second important feature of the ERG framework is that it provides a basis for likelihood-based inference. Maximum-likelihood based estimates for θ given $\mathbf{t}$ can be calculated using a number of methods (Crouch, Wasserman, & Trachtenburg, 1998; Snijders, 2002), and Bayesian approaches are also possible. One particularly useful result with respect to the former is the fact that $\mathbf{E}_{\hat{\theta}}\mathbf{t}(G) = \mathbf{t}(g)$ where $\hat{\theta}$ is the maximum-likelihood estimator (MLE) of θ given observed graph g. Thus, first-order method-of-moments estimators correspond to MLE for ERG; while this is not the most efficient method of computation, it is a useful fall-back method in many settings. This relationship hints at another important insight regarding the ERG parameterization: models in this form can be understood as providing distributions of maximum entropy over their support, conditional on fixing the expected sufficient statistics (as determined by θ) (Brown, 1986; Strauss, 1986). Thus, ERG can be used to construct extended baseline models of network structure, in which it is assumed that realized networks are maximally 'random' given the average values of their sufficient statistics. (Compare this to the CUG approach of assuming maximum entropy conditional on the exact values of selected sufficient statistics.) A third (and related) aspect of the ERG parameterization is that it facilitates the construc-

tion of network models which implement specific forms of dependence among edges. This construction is performed principally by application of the Hammersley-Clifford Theorem (Besag, 1974) to the dependence graph corresponding to the desired model (see Wasserman & Robins, 2005 for a discussion), although additional 'parameter filtering' methods are sometimes required (Pattison & Robins, 2002; Butts, 2006a). A rather remarkable result of this work has been the discovery of a deep duality between structural features (as measured by various indices) and dependence among edges. In particular, each potential choice of $\mathbf{t}$ implies a certain class of dependencies, and vice versa. The realization of this connection greatly facilitates the development of empirically grounded theory regarding social interaction (see Robins & Pattison, 2005, for a discussion), and is likely to be the basis for a great deal of research in the years ahead. Finally, it should be noted that the ERG form of Equation 10 can be extended in a number of ways to incorporate nodal covariates, multiple networks etc. One of these extensions (to multiple networks) will be considered further below.

Network comparison

Although much of the literature on social networks is focused on the measurement and modelling of features within particular networks, another important class of problems involves comparing structure across networks. Such problems naturally arise when we ask whether a particular intervention affects team structure, whether participation in one relation affects participation in others, or whether a particular collection of relations (e.g. expert mental models) reflect variations on a single underlying 'theme'. Here, I review three general approaches to network comparison (conditional uniform graph tests, linear subspace methods, and exponential family models), and describe some of the relative strengths and weaknesses of each approach.

Multivariate CUG tests

An immediate method of comparing networks is via their respective graph-level index values. A difficulty with this approach, however, is the fact that many GLI vary in nontrivial ways with the size and density of the networks under comparison. To determine whether differences in GLI values reflect substantive structural effects – as opposed to differences stemming from background features such as size – it is necessary to invoke a baseline model of some sort. Anderson, Butts, and Carley (1999) suggest using a variant of the conditional uniform graph approach discussed above as such a baseline when comparing graphs. In particular, let $\mathbf{t}$ be a real-valued vector of conditioning statistics, let $\mathbf{x}_1, \ldots, \mathbf{x}_m$ be real-valued vectors, and $G_1, \ldots, G_m$ be a set of graphs. We then posit a multivariate generalization of the conditional uniform graph distribution of Equation 9,

$$\mathcal{G}^m_{\mathbf{t},\mathbf{x}} = \{(g'_1, \ldots, g'_m) \in \mathbb{G}^m : (\mathbf{t}(g'_1), \ldots, \mathbf{t}(g'_m)) = (\mathbf{x}_1, \ldots, \mathbf{x}_m)\} \quad (11)$$

$$\Pr((G_1, \ldots, G_2) = (g_1, \ldots, g_m) | \mathbf{t}, \mathbf{x}_1, \ldots, \mathbf{x}_m) = \left[\left|\mathcal{G}^m_{\mathbf{t},\mathbf{x}}\right|\right]^{-1} I_{\mathcal{G}^m_{\mathbf{t},\mathbf{x}}}(g_1, \ldots, g_m). \quad (12)$$

As in the univariate case, $\mathbf{t}$ may consist of statistics such as network size, number of edges, the dyad census etc. Here, however, these statistics are specified for all graphs in the set (as opposed to a single graph).

To use the multivariate CUG distribution in the context of graph comparison, we first identify the multivariate statistic s on $\mathbb{G}^m$ to be tested. In the bivariate case, s will usually be a difference in GLI values for the two input graphs (or the absolute value of such a difference); other functions are possible, however. We then set $\mathbf{x}_1, \ldots, \mathbf{x}_m$ to form the hypothesis which is to be tested. Typically, we will seek to condition on the values of $\mathbf{t}$ in the observed networks $g_1, \ldots, g_m$, and, hence, will require that $(\mathbf{x}_1, \ldots, \mathbf{x}_m) = (\mathbf{t}(g_1), \ldots, \mathbf{t}(g_m))$. The one-tailed p-values for $s(g_1, \ldots, g_m)$ under the corresponding multivariate CUG test are then

$$\Pr(s(G_1, \ldots, G_m) \geq s(g_1, \ldots, g_m) | \mathbf{t}, \mathbf{x}_1, \ldots, \mathbf{x}_m)$$

for the upper tail, and

$$\Pr(s(G_1, \ldots, G_m) \leq s(g_1, \ldots, g_m) | \mathbf{t}, \mathbf{x}_1, \ldots, \mathbf{x}_m)$$

for the lower tail. Note that a 'two-tailed' test of GLI differences can be implemented here by defining $s(G_1, G_2) = |f(G_1) - f(G_2)|$ (for GLI f) and using the p-value associated with an upper-tail test. This last test can be interpreted as assessing the extent to which the absolute difference between GLI scores is large compared to the distribution of absolute differences which would be expected to arise, given the choice of conditioning statistics. A low p-value for such a test suggests that the difference in GLI scores for the graphs in question is larger than would be expected under the baseline model, suggesting the possibility that more subtle structural mechanisms may be at work. A large p-value, however, indicates that the difference in observed statistics is not particularly large compared to the baseline model, and calls into question whether additional explanations are needed.

Linear subspace methods

In some cases, we may wish to compare two graphs G, G' on some common vertex set, V. For instance, let us imagine that G_1 represents a network of positive interpersonal evalu-

ations and G_2 represents a network of event coparticipation for the members of some group; we might then seek to test the hypothesis that coparticipation is positively associated with positive interpersonal evaluations among group members. As the vertex sets for G_1 and G_2 are shared, this is properly seen as a problem of edge set comparison, which is a special case of the more general graph comparison problem. Hubert (1987) postulated a simple approach to edge set comparison based on the use of matrix product-moment statistics, which was further developed in the social network context by Krackhardt (1987b, 1988). As pointed out by Butts and Carley (2001, 2005), this approach is properly regarded as the application of linear subspace methods to graph sets, in direct analogy with the use of such methods in conventional multivariate data analysis; these authors also explore the use of closely related distance-based methods (following Banks & Carley, 1994), which will not be treated here.

The central element of the linear subspace methods for graph comparison is the graph covariance, which is defined as

$$\mathrm{cov}(G, G') = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(Y_{ij}-\mu)(Y'_{ij}-\mu') \qquad (13)$$

where $\mathbf{Y}$ and $\mathbf{Y}'$ are the respective adjacency matrices of G and G', and μ and μ' are the respective means of these adjacency matrices. (Note that diagonals elements should be treated as missing if loops are not allowed; for simplicity, we use the notation for the general case.) Intuitively, the graph covariance is simply the covariance of the two adjacency matrices, taken as a collection of edge variables. As one would then expect, $\mathrm{cov}(G, G) = \mathrm{var}(G)$ is the graph variance of G, leading to the graph correlation $\rho(G, G') = \mathrm{cov}(G, G')/\sqrt{\mathrm{var}(G)\,\mathrm{var}(G')}$. Graph correlations/covariances can be used directly to compare graphs, in the manner discussed by Krackhardt (1987b): tests for the observed magnitude of these comparison statistics can be conducted using the quadratic assignment procedure (QAP) of Hubert (1987), which controls for the effects of row, column, and block autocorrelation (all of which are common in network data). The QAP test is a simple matrix permutation test, in which the observed graph statistic (here, correlation or covariance) is compared to the distribution of such statistics arising from the simultaneous row/column permutation of the respective adjacency matrices. Specifically, let ℓ be a random permutation of the integers $1, \ldots, n$, and let t be a bivariate graph statistic. Then the null distribution of t under the QAP hypothesis is the distribution of $t(\mathbf{Y}, \mathbf{Y}'_{\ell\ell})$, where $\mathbf{Y}'_{\ell\ell}$ is the adjacency matrix $\mathbf{Y}'$ row/column reordered by ℓ; this is equivalent to the distribution of $t(G, \ell(G'))$, using the graph permutation notation developed earlier. This procedure controls for all purely structural properties of the graphs being compared and, as such, effectively tests the hypothesis that the degree of association induced by the observed labelling of the two networks can be explained by their underlying structure. Rejection of this hypothesis suggests the possibility that the elements of each network have been positioned in a way which specifically induces a stronger degree of association (or disassociation) between the two networks than would be expected given their respective structures. More intuitively, the bivariate QAP test can be thought of as comparing the degree of observed association between networks to that which would be expected to arise from a process in which individuals were randomly assigned to positions within the two networks, holding the structure constant. If ties between positions coincide more (or less) frequently than this process would indicate, this may suggest that some other social process is at work.

In addition to tests for bivariate association, the graph covariance/correlation can be used for multivariate analysis of graph sets. Given a graph set $G_1, \ldots, G_m$, one can construct a graph covariance or correlation matrix in precisely the same manner as one would construct a covariance or correlation matrix for conventional variables. These matrices can then be used to obtain solutions for linear regression, principal component analysis, canonical correlation analysis, or other linear subspace analyses, just as in conventional multivariate analysis (Mardia, Kent, & Bibby, 1979). Of these solutions, linear regression has been the most widely used (following the early incorporation of the approach of Krackhardt (1988) into software packages such as UCINET (Borgatti, Everett, & Freeman, 1999)); alternatives such as canonical correlation analysis have been available in some software packages for several years (e.g. the sna package for R (Butts, 2000)), but have not thus far seen extensive use. As most network data are dichotomous, linear analyses are rarely plausible as data models – however, they can be highly effective as tools for exploratory data analysis. Given a large collection of networks, linear subspace methods such as principal component analysis can identify associations among structures, and can identify underlying 'structural factors' which can parsimoniously explain variation in a larger set of relations. The insights resulting from such analyses can then be used in constructing more principled data models, such as those discussed below.

Exponential family models

While exponential family parameterizations have been most frequently used in the modelling of single networks, this is not a fundamental restriction. In fact, this framework can be easily extended to encompass multiple relations, either on shared or distinct vertex sets. Here, we briefly review two approaches to the use of discrete exponential

family models to the problem of graph comparison. The first (based on graph permutations, or re-assignment of individual positions) can be seen as a model-based extension of the philosophy of Hubert (1987). The second involves the direct modelling of multiple networks from a common set of sufficient statistics. Both approaches are relatively recent entrants to the literature, and it is expected that this area will see much development over the next several years.

Permutation models. A major limitation of the linear subspace models described above is that they are poorly suited to dichotomous data: this makes coefficients difficult to interpret, and effectively negates the plausibility of the associated models as data-generation mechanisms. Similarly, such models provide little principled basis for inference, as they do not posit a likelihood for the set of observed networks. A recently developed approach which overcomes these limitations is the use of permutation models to compare graphs or graph sets (Butts, 2007). Let us consider a case in which we have two sets of graphs, $G_1, \ldots, G_m$ and $G'_1, \ldots, G'_p$ on common vertex set V. For convenience, I will represent the adjacency structures of each graph set by the respective arrays $\mathbf{Y}$ and $\mathbf{Y}'$, such that Y_{ijk} is the j, kth entry of the adjacency matrix of G_i, and Y'_{ijk} is the j, kth entry of the adjacency matrix of G'_i. As in the discussion of the QAP test, let ℓ be a permutation vector on $1, \ldots, n$ (reflecting a potential vertex ordering), and let $\mathbf{Y}'_{\cdot\ell}$ reflect the adjacency array for $G'_1, \ldots, G'_p$, with all vertices permuted by ℓ. I then posit a model for the assignment of vertices to positions in one graph set relative to the other (i.e. for the vector ℓ) using the following discrete exponential family PMF:

$$\Pr(\ell = l|\mathbf{t}, \theta, \mathbf{Y}, \mathbf{Y}') = \frac{\exp\left(\theta^t \mathbf{t}(\mathbf{Y}, \mathbf{Y}'_{\cdot l})\right)}{\sum_{l' \in \mathcal{L}} \exp\left(\theta^t \mathbf{t}(\mathbf{Y}, \mathbf{Y}'_{\cdot l'})\right)} I_{\mathcal{L}}(l). \quad (14)$$

$\mathcal{L}$ here defines the support of ℓ, and is known as the set of accessible permutations. θ and $\mathbf{t}$ are both assumed to take values in $\mathbb{R}^h$, as with the ERG model of Equation 10. The primary difference here is that we are modelling not the network structures per se, but the 'assignment' of individuals to positions within the existing networks. Indeed, we condition on the network structures themselves and, in so doing, control for all sources of within-graph (and within graph set) autocorrelation. The cost of this manoeuvre is some loss of information, as the support of ℓ is generally much smaller than the support of $G_1, \ldots, G_m$ and $G'_1, \ldots, G'_p$ would be in the absence of conditioning (see below). One compensation for this loss, however, is that the model can be easily applied to arbitrarily valued data, something which is not true of conventional exponential random graph models.[7]

As Equation 14 defines an exponential family on a set of graph permutations, Butts (2007) refers to this as the 'exponential random graph permutation' (ERGP) family of models. Although Butts's treatment is restricted to the product moment statistics $\Gamma(\mathbf{Y}_{i\cdot\cdot}, \mathbf{Y}'_{i\ell}) = \Sigma_{j=1}^n \Sigma_{k=1}^n Y_{ijk} Y'_{i\ell j\ell k}$ (better known as Hubert's Gamma), $\mathbf{t}$ can be chosen to be any statistic which is not invariant to ℓ. This includes the cross-graph statistics derived by Pattison and Wasserman (1999), but excludes statistics which depend only on single graphs (or on graphs chosen strictly from within the two comparison sets). Butts provides methods for simulation and inference for ERGP models, and discusses connections with procedures such as the QAP test. Butts also notes that the ERGP has a non-empty intersection with the general family of multivariate exponential random graph models, which can be used to model general joint distributions on graph sets. We thus turn next to this family of models.

Multivariate ERG models. Just as the 'univariate' model of Equation 10 expressed a probability model for a single network in terms of a set of sufficient statistics, so too can we construct 'multivariate' exponential family models (MERG) for sets of graphs (Pattison & Wasserman, 1999). Formally, let $G_1, \ldots, G_m$ be graphs drawn from a distribution with finite joint support $\mathcal{G}_1 \times \cdots \times \mathcal{G}_m$, let $\theta \in \mathbb{R}^h$ be a parameter vector, and let $\mathbf{t}$ be a vector of sufficient statistics taking $\mathcal{G}_1 \times \cdots \times \mathcal{G}_m$ into $\mathbb{R}^h$. Then we may write a PMF for the joint distribution of $G_1, \ldots, G_m$ of the form

$$\Pr((G_1, \ldots, G_m) = (g_1, \ldots, g_m)|\mathbf{t}, \theta) = \frac{\exp\left(\theta^t \mathbf{t}(g_1, \ldots, g_m)\right)}{\sum_{(g'_1, \ldots, g'_m) \in \mathcal{G}_1 \times \cdots \times \mathcal{G}_m} \exp\left(\theta^t \mathbf{t}(g'_1, \ldots, g'_m)\right)} \times I_{\mathcal{G}_1 \times \cdots \times \mathcal{G}_m}(g_1, \ldots, g_m), \quad (15)$$

(with I being, as in Equation 10, a dichotomous indicator function for membership in the support). The MERG family is a direct generalization of the ERG family, and can be interpreted in the same manner. In particular, the model posits that graph sets with larger values of t_i become more probable as $\theta_i \to \infty$ (ceteris paribus), and less probable as $\theta_i \to -\infty$. As $\mathbf{t}$, in this case, is a function of the graph set as a whole, the MERG can directly parameterize arbitrary dependence between (as well as within) graphs; note that this is not necessary, however, as any given statistic can be made to depend on only a single graph. As such, the MERG takes the 'univariate' ERG as a special case, and a product of disjoint ERG distributions is equivalent to a corresponding MERG in which no sufficient statistic depends on more than one input graph. Simulation and inference for MERG is conducted exactly as for the ERG case, with the complication that the support involves multiple graphs. Thus, the computational cost of working with MERG models may be

substantially higher than ERG, although the underlying methods are the same. The simultaneous treatment of multiple networks does offer the possibility of a range of new forms of dependence, each corresponding to new sets of sufficient statistics. Pattison and Wasserman (1999) and Koehly and Pattison (2005) have demonstrated a number of distinct statistics for multivariate exponential family models, based on such dependence hypotheses, offering rich opportunities for model construction in this area.

Analysis of nodal covariates

Although the foregoing has focused on the measurement and modelling of network structure per se, nodal covariates are also of interest in many settings. In the case of social influence, for instance, we may be interested in how individuals' attitudes affect one another through a social network. Similarly, we may seek to determine having large numbers of ties to close friends and family is predictive of mental health outcomes, or, alternatively, whether such outcomes may impact one's social position. Although analysis of nodal covariates may, in some cases, be carried out using traditional statistical methods, the interdependence of structural properties (and, in the case of influence processes, the covariates themselves) sometimes require the use of alternative methods. Here I briefly review some of these approaches, and provide suggestions regarding their effective use.

Node-level indices and node-level attributes

An enduring line of inquiry within the social network field concerns the relationship between node-level attributes, and the contrasting properties of structural positions. Such questions arise naturally from theories which posit differences in social behaviour and/or positional attainment due to exogenous covariates, differences in outcomes due to differing social position etc., and can take many forms. While not all position/attribute questions fall into this category, many such queries lead naturally to analyses which directly relate node-level indices to nodal covariates.

Where one's objective is the prediction of nodal covariates from node-level indices and where conditional independence of covariate values can be assumed, traditional methods (e.g. generalized linear models) may usually be used without special difficulty. More serious concerns arise where node-level indices are taken as dependent variables, or where measures of symmetric association (e.g. correlation) are to be evaluated. The primary difficulties here are two-fold: the fact that node-level index values typically exhibit intrinsic dependence, and the fact that conditional normal models are often poorly suited to describing index distributions. In a regression context, standard transformation methods and/or modified models such as tobit or quantile regression (Tobin, 1958; Koenker & Bassett, 1978) can prove helpful in alleviating the latter problem. The issue of dependence is, in some ways, more complicated and cannot be entirely resolved without the use of exponential family models (see above). In many contexts, however, it is possible to test simple hypotheses of association by means of permutation tests (much like the QAP case described above). In particular, the observed value of an association statistic for a vector of node-level index values versus a vector or matrix of nodal covariates can be compared with the value of the statistic arising from repeated permutations of the index distribution. As this procedure preserves the joint distribution of the indices (effectively 'moving' individuals while keeping network structure fixed), it is nonparametric with respect to the index distribution per se. Standard considerations regarding the use of (vector) permutation tests apply here; a reasonable general-purpose reference is Good (2000).

As a final cautionary, it must be stressed that the node-level index/nodal covariate approach can easily be overused. Many social process theories, in particular, argue that the properties of one's alters are as important as the configural aspects of one's network position, and may make no direct predictions regarding the effect of the latter quantities per se. For example, most theories of social influence (e.g. Latané, 1981; Butts, 1998; Freidkin, 1998) posit that individuals will tend to adopt the attitudes and/or beliefs of their alters; thus, the predicted effect of features such as centrality or ego network density cannot be specified independent of alters' attributes. Use of purely structural measures to assess covariate-based theories is incorrect, and will yield misleading inferences.

Network autocorrelation, influence, and diffusion

Frequently, nodal covariates are not socially exogenous, but are, at least partially, the result of interaction between individuals. Even where one's primary interest is in the impact of covariates which are hypothesized to have socially exogenous effects, failure to control for social endogeneity can lead to extremely misleading results. An important family of regression-like models which can be used to capture and/or control for such effects is the family of (linear) network autoregressive/moving average (ARMA) models. Network ARMA models (Doreian, 1989, 1990) treat individual nodes' covariate values as potentially dependent upon the values of neighbours' covariates, as well as upon exogenous covariates and (possibly dependent) shocks. In this they can be seen as a natural generalization of models for temporal and spatial dependence. (In fact, the network ARMA model is formally identical to the spatial ARMA

(SARMA) model (Cliff & Ord, 1973; Anselin, 1988) which is widely used in geographical settings. The two differ only in terminology and application.)

Network ARMA models comprise a standard regression model combined with two components: an autoregressive (AR) component, which models the direct dependence of observations upon one another; and a moving-average (MA) component, which models the dependence among the exogenous perturbations, or errors. These two components act in distinct ways, and one or both may be used in any given setting. At the same time, the substantive difference between the AR and MA processes can be subtle, and are a frequent source of confusion. MA processes, for instance, are sometimes said to be applicable only when measurement errors are correlated across individuals; this is an important issue in spatial settings, but is less common with interpersonal networks. In general, however, it is appropriate to use an MA process wherever one has reason to expect the presence of exogenous shocks which are transmitted through the social network independent of any covariate effects. As an illustrative example, consider a model for self-reported coping success, in the context of life difficulties. Naturally, we expect that each individual will have his or her share of good and bad luck, which enters the system as an exogenous shock. Clearly, such shocks will interact with each person's individual attributes to determine his or her success in coping; however, we may also hypothesize that this process is not independent of the experiences of friends and family members. One example of such a process is one in which each person feels not only his or her own shocks, but some weighted average of the total shocks felt by his or her peers. Thus, good fortune on the part of a given family member will aid the entire family (to some extent, at least), whereas a corresponding misfortune will have a negative impact. If these shocks diffuse independently of each person's actual success in coping with them, then the result will behave as a network MA process. Alternatively, consider the possibility that each person's coping success depends not on his or her neighbours' shocks alone, but directly on his or her neighbours' own levels of coping success. In this case, the process in question is autoregressive, and a network AR component is implicated. Note that a key difference between the two cases is that neighbours' covariates themselves have a diffusive effect in a network AR process, whereas it is only the shocks or deviations which diffuse in the MA case. In terms of our example, being tied to someone with very poor coping skills will tend to drag you down where coping is autoregressive, even if his or her luck has been fairly good. By contrast, if coping is a moving average process, it is only his or her luck which will impact you. In many cases, it will not be obvious *ex ante* which process is the correct one (or if both are active). By fitting AR, MA, and ARMA models, this question can be resolved empirically.

Frequently, it is assumed that any AR and/or MA effects act through a single adjacency structure; this need not be the case, however, and the generalization to network ARMA models with multiple channels of dependence is quite immediate. Specifically, let $\mathbf{W}_1, \ldots, \mathbf{W}_w$ be the set of adjacency matrices governing the AR process, and let $\mathbf{Z}_1, \ldots, \mathbf{Z}_z$ be the corresponding adjacency matrices for the MA process. These adjacency matrices need not be dichotomous and, indeed, often should be valued (see below); we interpret the j, k cell of each matrix as giving the weight placed on node j by node i in the corresponding social process. We also allow for the presence of a real-valued covariate matrix, $\mathbf{X}$, which is assumed to act directly on the dependent variable, $\mathbf{y}$. The network ARMA model may then be defined as follows:

$$\mathbf{y} = \left(\sum_{i=1}^{w} \theta_i \mathbf{W}_i \right) \mathbf{y} + \mathbf{X}\beta + \varepsilon \tag{16}$$

$$\varepsilon = \left(\sum_{i=1}^{z} \psi_i \mathbf{Z}_i \right) \varepsilon + v, \tag{17}$$

where $\mathbf{E}(v) = \mathbf{0}$, $v_i \perp v_j \forall i, j$. Positing a parametric form for v (typically iid normal with unknown constant variance σ^2) permits model estimation using maximum likelihood using standard methods. A more useful form for this purpose is obtained by solving Equations 16 and 17 for $\mathbf{y}$ and ε, respectively. Specifically, we have

$$\varepsilon - \left(\sum_{i=1}^{z} \psi_i \mathbf{Z}_i \right) \varepsilon = v \tag{18}$$

$$\left(\mathbf{I} - \left(\sum_{i=1}^{z} \psi_i \mathbf{Z}_i \right) \right) \varepsilon = v \tag{19}$$

$$\varepsilon = \left(\mathbf{I} - \left(\sum_{i=1}^{z} \psi_i \mathbf{Z}_i \right) \right)^{-1} v, \tag{20}$$

and, similarly,

$$\left(\mathbf{I} - \left(\sum_{i=1}^{w} \theta_i \mathbf{W}_i \right) \right) \mathbf{y} = \mathbf{X}\beta + \varepsilon \tag{21}$$

$$\mathbf{y} = \left(\mathbf{I} - \left(\sum_{i=1}^{w} \theta_i \mathbf{W}_i \right) \right)^{-1} (\mathbf{X}\beta + \varepsilon). \tag{22}$$

Hence, by substitution,

$$\mathbf{y} = \left(\mathbf{I} - \left(\sum_{i=1}^{w} \theta_i \mathbf{W}_i \right) \right)^{-1} \times \left(\mathbf{X}\beta + \left(\mathbf{I} - \left(\sum_{i=1}^{z} \psi_i \mathbf{Z}_i \right) \right)^{-1} v \right). \tag{23}$$

These solutions exist only when the aggregated weight matrices $\mathbf{W}^* = \Sigma_{i=1}^{w} \theta_i \mathbf{W}_i$ and $\mathbf{Z}^* = \Sigma_{i=1}^{z} \psi_i \mathbf{Z}_i$ are invertible. This, in turn, amounts to the condition that each of the $\mathbf{W}$ and $\mathbf{Z}$ matrices is invertible. Setting θ and/or ϕ to zero leads to the network MA, network AR, or standard regression models, respectively; these can thus be considered submodels of the joint ARMA process.

Extensive research on the network AR model as a model for social influence in small group settings has been carried out by Friedkin and Johnsen (Friedkin & Johnsen, 1990; Friedkin & Cook, 1991; Freidkin, 1998). Results from a large body of experiments performed by these researchers have suggested that (in the case of influence processes for attitudes) the aggregate AR weight matrix $\mathbf{W}^*$ should nearly always be non-negative and quasi-convex (i.e. that $\Sigma_{j=1}^{n} \mathbf{W}_{ij}^* \leq 1$) in typical settings. In practical terms, this condition corresponds to a process in which final opinions are contained within (or exist on the boundary of) the convex hull of initial opinions. Given that this constraint appears to be satisfied by observed discussion groups, it seems reasonable to posit quasi-convexity in similar settings. This and a number of other theoretical issues regarding network AR models for social influence are discussed in Freidkin (1998).

In addition to influence, network autocorrelation models have been proposed as potentially useful tools for the study of diffusion (see Valente, 2005 for a discussion). A great deal of caution is advised here, however, as the linear process on which the network ARMA model is based may not be satisfied by such data. Another significant concern is the choice of potential weight matrices in practical settings. A review of many potential options, and a discussion of the relevant issues can be found in Leenders (2002).

Discussion

In the preceding pages, we have considered a brief overview of common and useful methods for network analysis. The scientific fruitfulness of such techniques, however, is dependent upon the power of the theoretical framework in whose service they are employed, and the match between theory and method. Here I comment on a few related issues that affect the use of social network analysis in practical settings.

Choosing the right network

Whenever one engages in network analysis, it is important not to lose sight of the fact that the relations being studied are only a subset of those within which the associated individuals are embedded. Essentially all persons live within networks of physical interaction, material transactions (e.g. exchange), interpersonal communication, mating and sexual contact etc. Along with these, we have more culturally specific networks of friendship and affiliation, social support, ascribed kinship, and the like; persons living within complex societies will additionally have non-trivial networks of institutional affiliation, collaborative task performance, advice and information sharing, training and mentorship, and technologically mediated contact (among many others). Further, this short enumeration says nothing of the many networks which may be defined among concepts, texts, organizations, or other non-human entities. Given this diversity, it is highly misleading (at best) to speak of 'the' social network in which a person or other entity resides. An individual to whom no one comes for professional advice may nevertheless have many friends, and vice versa – it is unwise to jump to the conclusion that an individual is generally socially isolated on the basis of isolation in one relation, just is it is similarly unwise to presume that an individual who is highly central in one setting is highly central in all settings. Likewise, the global properties of one relation on a given group may or may not be reflective of other relations' properties. For instance, an organization with highly centralized reporting structures may have very decentralized structures of informal communication (perhaps to the chagrin of senior management). Although the structures of multiple relations on the same individuals may tend to coincide, this coincidence cannot be taken for granted: social structure is rarely reducible to a single network.

Given the reality of overlapping, multiplex structures in social life, it is important that analysts select their networks with the same care that they apply to selecting other variables of substantive interest. In particular, the networks that are chosen for a particular application should be those indicated by applicable substantive theory, and not simply those that happen to be close at hand. While it is possible to use one network as a proxy for another, unobserved relation, the reliability and validity of such a solution should be empirically demonstrated rather than assumed on an a priori basis. Similarly, it is important to ensure that the network boundary which is used for a given analysis is substantively justifiable. It may be reasonable to assume that an individual living in a total institution (in the sense of Goffman, 1961) will rely primarily on other members of his or her organization for affective support, for instance, but such an assumption would be very questionable within a setting such as a voluntary interest group with infrequent meetings. If an individual were to appear an isolate with respect to support as measured in one of these settings, the implications would hence be quite different. In particular, it would be unreasonable to presume that a lack of support within the voluntary group implies a lack of social resources, as this population reflects only a small subsample of potential alters. The naïve analyst may be

tempted to conclude exactly that, however, falling prey to a type of 'tunnel vision' which regards the network at hand as a complete census of its members' social interactions. Careful attention to the substantive meaning of network ties and the sampling process by which they are measured is needed to avoid such errors.

Social process and structural signatures

Although our focus here has been on the analysis of 'snapshots' reflecting either instantaneous or time-averaged structure, it should be emphasized that network analysis can also play a role in the understanding of social processes. Stable networks can serve as the context in which phenomena such as social influence (Freidkin, 1998) and bargaining (Willer, 1999) occur and, hence, interact with low-level dynamics to shape social outcomes; such processes have been explored through simulation studied (Krackhardt, 1997; Butts, 1998), and are a target of ongoing research. Likewise, there is a growing literature on the time evolution of networks themselves (including both agent-based (Carley, 1991) and statistical (Snijders, 1996) approaches), which builds on the 'static' methods reviewed here. Beyond these, however, it is also important to emphasize that even static snapshots can contain the structural signatures of the microprocesses giving rise to them and can, hence, be used in many cases to test hypotheses regarding such processes. Although this tradition extends back at least to Rapoport (1949a, b, 1950) and Davis and Leinhardt (1972), it has been considerably enhanced by recent work on dependence graphs by Robins and Pattison (2005) and others, and has stood behind much of the interest of the physical science community in degree distributions (Newman, 2003). As cross-sectional network data are much more easily obtained than longitudinal data, there is much to be said for its use in this regard. It is thus hoped that the coming years will bring further innovations in linking social dynamics to cross-sectional structure.

When networks are not enough

As has been emphasized, effective network analysis depends as much on knowledge of the phenomenon at hand as any other area of scientific study. An important component of that knowledge is the recognition of where non-network data are needed to resolve a question of substantive or methodological importance. Although social networks provide a powerful tool for understanding social processes – and are of great scientific interest in their own right – it is naïve to presume that all social scientific questions can be answered with network data alone. Information on individual attributes, contextual variables, and social processes can and should be combined with network data in drawing conclusions regarding social phenomena, as required by the theories being tested.

Conclusion

Social network analysis is a powerful family of tools for the representation and analysis of relational data. I have here reviewed some of the basic methods in this area, along with the rudiments of study design and data collection. As an area of active interest, the techniques of social network analysis are likely to see considerable development in the years ahead. By making use of these innovations, researchers in psychology and allied sciences can better predict and account for the structural dimensions of social processes.

Acknowledgements

The author would like to thank Garry Robins for his helpful comments on this manuscript. This work was supported in part by NSF award CMS-0624257.

End notes

1. For an insightful treatment of the latter, see Sussman and Wisdom (2001).
2. Note that, where loops are not meaningful, most authors permit adjacent vertices to be structurally equivalent despite the fact that they do not belong to their own neighbourhoods.
3. Or, more accurately, it is local only to v's component.
4. The use of these concepts within the social network literature extends back at least to Holland and Leinhardt (1970) and related papers, and accompany a corresponding history within the mathematical literature in graph theory; recent reinventions under names such as 'motifs' or 'graphlets' do not always recognize this prior work.
5. Although terminology differs widely by author, this method has been used at least since the work of Katz and Powell (1953). See, for example, Holland and Leinhardt (1970, 1975); Wasserman (1987); Snijders (1991); Anderson *et al*. (1999); and Pattison, Wasserman, Robins, and Kanfer (2000) for variants.
6. To see this, let $t_i(g)$ be an indicator for the ith element of $\mathcal{G}$, and $\theta_i = \text{logit Pr}(G = g_i)$.
7. Exponential families for valued graphs are possible, but are considerably less trivial to parameterize than non-valued ERG. Robins, Pattison, and Wasserman (1999) provide one such application, but a comprehensive treatment is not currently available.

References

Anderson, B. S., Butts, C. T. & Carley, K. M. (1999). The interaction of size and density with graph-level indices. *Social Networks, 21* (3), 239–267.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Norwell, MA: Kluwer.

Baltz, A. & Kloemann, L. (2005). Spectral analysis. In: U. Brandes & T. Erlebach, eds. *Network Analysis: Methodological Foundations*, pp. 373–416. Berlin: Springer-Verlag.

Banks, D. & Carley, K. M. (1994). Metric inference for social networks. *Journal of Classification, 11* (1)*,* 121–149.

Barabási, A.-L. & Albert, R. (1999). Emergence of scaling in random networks. *Science, 206,* 509–512.

Batagelj, V. & Mrvar, A. (2007). Pajek – program for large network analysis. Ljubljana: Vlado Networks. Electronic data file. Available from http://vlado.fmf.uni-lj.si/pub/networks/pajek/.

Batchelder, W. H. & Romney, A. K. (1988). Test theory without an answer key. *Psychometrika, 53* (1)*,* 71–92.

Bavelas, A. & Barrett, D. (1951). An experimental approach to organizational communication. *Personnel, 27,* 366–371.

Berge, C. (1962). *The Theory of Graphs*. London: Methuen and Company.

Bernard, H. R. & Killworth, P. (1977). Informant accuracy in social network data II. *Human Communication Research, 4* (1)*,* 3–18.

Bernard, H. R., Killworth, P., Kronenfeld, D. & Sailer, L. (1984). The problem of informant accuracy: The validity of retrospective data. *Annual Review of Anthropology, 13,* 495–517.

Bernard, H. R., Killworth, P. & Sailer, L. (1979). Informant accuracy in social networks IV: A comparison of clique-level structure in behavioral and cognitive network data. *Social Networks, 2,* 191–218.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B, 36* (2)*,* 192–236.

Bollobás, B. (1998). *Modern Graph Theory*. New York: Springer.

Bonacich, P. (1972). Factoring and weighting approaches to clique identification. *Journal of Mathematical Sociology, 2,* 113–120.

Borgatti, S. P. (2007). NetDraw: Network visualization software. Software package. Harvard: Analytic Technologies.

Borgatti, S. P., Carley, K. & Krackhardt, D. (2006). Robustness of centrality measures under conditions of imperfect data. *Social Networks, 28,* 124–136.

Borgatti, S. P. & Everett, M. G. (1999). Models of core/periphery structures. *Social Networks, 21,* 375–395.

Borgatti, S. P., Everett, M. G. & Freeman, L. C. (1999). *UCINET 5.0*, Version 1.00. Natick, NJ: Analytic Technologies.

Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology, 25* (2)*,* 163–177.

Brandes, U. & Erlebach, T., eds. (2005). *Network Analysis: Methodological Foundations.* Berlin: Springer-Verlag.

Brass, D. J. (1984). Being in the right place: A structural analysis of individual influence in an organization. *Administrative Science Quarterly, 29,* 519–529.

Brewer, D. (2000). Forgetting in the recall-based elicitation of personal networks. *Social Networks, 22,* 29–43.

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families, with Applications in Statistical Decision Theory*. Hayward, CA: Institute of Mathematical Statistics.

Burt, R. S. (1992). *Structural Holes: The Social Structure of Competition*. Cambridge, MA: Harvard University Press.

Butts, C. T. (1998). A Bayesian model of panic in belief. *Computational and Mathematical Organization Theory, 4* (4)*,* 373–404.

Butts, C. T. (2000). The sna package for the R statistical computing system. Software library. Pittsburgh, PA. Available from http://erzuli.ss.uci.edu/R.stuff.

Butts, C. T. (2003). Network inference, error, and informant (in)accuracy: A Bayesian approach. *Social Networks, 25* (2)*,* 103–140.

Butts, C. T. (2006a). *Cycle census statistics for exponential random graph models* (IMBS Technical Report MBS 06-05). Irvine, CA: Institute for Mathematical Behavioral Sciences, University of California, Irvine.

Butts, C. T. (2006b). Exact bounds for degree centralization. *Social Networks, 28* (4)*,* 283–296.

Butts, C. T. (2007). Permutation models for relational data. *Sociological Methodology, 37,* 257–281.

Butts, C. T. & Carley, K. M. (2001). *Multivariate methods for interstructural analysis.* CASOS Working Paper. Carnegie Mellon University, Pittsburgh, PA: Center for the Computational Analysis of Social and Organization Systems.

Butts, C. T. & Carley, K. M. (2005). Some simple algorithms for structural comparison. *Computational and Mathematical Organization Theory, 11* (4)*,* 291–305.

Butts, C. T. & Pixley, J. E. (2004). A structural approach to the representation of life history data. *Journal of Mathematical Sociology, 28* (2)*,* 81–124.

Carley, K. M. (1991). A theory of group stability. *American Sociological Review, 56* (3)*,* 331–354.

Carley, K. M. (1997). Extracting team mental models through textual analysis. *Journal of Organizational Behavior, 18,* 533–538.

Carrington, P. J., Scott, J. & Wasserman, S., eds. (2005). *Models and Methods in Social Network Analysis.* Cambridge: Cambridge University Press.

Choudhury, T. & Pentland, A. (2003). Sensing and modeling human networks using the sociometer. *Proceedings of the 7th IEEE International Symposium on Wearable Computers*, pp. 216–222. New York: White Plains.

Cliff, A. D. & Ord, J. K. (1973). *Spatial Autocorrelation*. London: Pion.

Costenbader, E. & Valente, T. W. (2003). The stability of centrality measures when networks are sampled. *Social Networks, 25,* 283–307.

Crouch, B., Wasserman, S. & Trachtenburg, F. (1998). *Markov chain Monte Carlo maximum likelihood estimation for* p* *social network models.* Paper presented at the XVIII International Sunbelt Social Network Conference; April 1998, Sitges, Spain.

Davis, J. A. & Leinhardt, S. (1972). The structure of positive interpersonal relations in small groups. In: J. Berger, ed. *Sociological Theories in Progress*, Vol. 2, pp. 218–251. Boston, MA: Houghton Mifflin.

Davis, J. A. & Smith, T. W. (1988). *General Social Survey, 1988.* Chicago, IL: National Opinion Research Center.

Degenne, A. & Forsé, M. (1999). *Introducing Social Networks*. London: Sage.

Doreian, P. (1989). Two regimes of network autocorrelation. In: M. Kochen, ed. *The Small World*, pp. 280–295. Norwood: Ablex.

Doreian, P. (1990). Network autocorrelation models: Problems and prospects. In: I. D. A. Griffith, ed. *Spatial Statistics: Past, Present, and Future*, pp. 369–389. Ann Arbor, MI: Institute of Mathematical Geography.

Doreian, P., Batagelj, V. & Ferlioj, A. (2005). *Generalized Blockmodeling*. Cambridge: Cambridge University Press.

Ebel, H., Mielsch, L. I. & Bornholdt, S. (2002). Scale-free topology of e-mail networks. *Physical Review E, 66,* 035103.

Erdös, P. & Rényi, A. (1960). On the evolution of random graphs. *Public Mathematical Institute of Hungary Academy of Sciences, 5,* 17–61.

Everett, M. G. & Borgatti, S. P. (2005). Ego-network betweenness. *Social Networks, 27* (1)*,* 31–38.

Festinger, L., Schachter, S. & Back, K. (1950). *Social Pressures in Informal Groups*. Stanford, CA: Stanford University Press.

Frank, O. (1978). Sampling and estimation in large social networks. *Social Networks, 1* (1)*,* 91–101.

Frank, O. (2005). Network sampling and model fitting. In: P. J. Carrington, J. Scott & S. Wasserman, eds. *Models and Methods in Social Network Analysis*, Chapter 3, pp. 31–56. Cambridge: Cambridge University Press.

Frank, O. & Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association, 81,* 832–842.

Freeman, L. C. (1979). Centrality in social networks: Conceptual clarification. *Social Networks, 1* (3)*,* 223–258.

Freeman, L. C. (2000). Visualizing social networks. *Journal of Social Structure, 1* (1).

Freeman, L. C. (2004). *The Development of Social Network Analysis: A Study in the Sociology of Science*. Vancouver: Empirical Press.

Freeman, L. C., Fararo, T. J., Bloomberg, W. J. & Sunshine, M. H. (1963). Locating leaders in local communities: A comparison of some alternative approaches. *American Sociological Review, 28,* 791–798.

Freidkin, N. (1998). *A Structural Theory of Social Influence*. Cambridge: Cambridge University Press.

Friedkin, N. & Cook, K. S. (1991). Peer group influence. *Sociological Methods and Research, 19,* 122–143.

Friedkin, N. & Johnsen, E. C. (1990). Social influence and opinions. *Journal of Mathematical Sociology, 15,* 193–206.

Fruchterman, T. & Reingold, E. (1991). Graph drawing by force-directed placement. *Software – Practice and Experience, 21* (11)*,* 1129–1164.

Goffman, E. (1961). *Asylums: Essays on the Social Situation of Mental Patients and Other Inmates*. New York: Doubleday.

Good, P. (2000). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. New York: Springer.

Goodman, L. A. (1961). Snowball sampling. *Annals of Mathematical Statistics, 32,* 148–170.

Gould, R. & Fernandez, R. (1989). Structures of mediation: A formal approach to brokerage in transaction networks. *Sociological Methodology, 19,* 89–126.

Hage, P. & Harary, F. (1995). Eccentricity and centrality in networks. *Social Networks, 17,* 57–63.

Handcock, M. S. (2003). Statistical models for social networks: Inference and degeneracy. In: R. Breiger, K. M. Carley & P. Pattison, eds. *Dynamic Social Network Modeling and Analysis*, pp. 229–240. Washington, DC: National Academies Press.

Handcock, M. S., Raftery, A. E. & Tantrum, J. M. (2007). Model based clustering for social networks. *Journal of the Royal Statistical Society, Series A, 170* (2)*,* 301–354.

Heckathorn, D. D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems, 44* (2)*,* 174–199.

Heckathorn, D. D. (2002). Respondent-driven sampling ii: Deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems, 49* (1)*,* 11–34.

Heider, F. (1946). Attitudes and cognitive organization. *Journal of Psychology, 21,* 107–112.

Hoff, P. D., Raftery, A. E. & Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association, 97* (460)*,* 1090–1098.

Holland, P. W. & Leinhardt, S. (1970). A method for detecting structure in sociometric data. *American Journal of Sociology, 70,* 492–513.

Holland, P. W. & Leinhardt, S. (1972). Some evidence on the transitivity of positive interpersonal sentiment. *American Journal of Sociology, 72,* 492–513.

Holland, P. W. & Leinhardt, S. (1975). The statistical analysis of local structure in social networks. *Sociological Methodology, 6,* 1–45.

Holland, P. W. & Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs (with discussion). *Journal of the American Statistical Association, 76* (373)*,* 33–50.

Hubert, L. J. (1987). *Assignment Methods in Combinatorial Data Analysis*. New York: Marcel Dekker.

Hummon, N. P. & Fararo, T. J. (1995). Assessing hierarchy and balance in dynamic network models. *Journal of Mathematical Sociology, 20,* 145–159.

Kadushin, C. (1982). Social density and mental health. In: P. V. Marsden & N. Lin, eds. *Social Structure and Network Analysis*, pp. 147–158. Newbury Park, CA: Sage.

Kamada, T. & Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters, 31* (1)*,* 7–15.

Katz, L. & Powell, J. H. (1953). A proposed index of conformity of one sociometric measurement to another. *Psychometrika, 18,* 249–256.

Killworth, P. D. & Bernard, H. R. (1976). Informant accuracy in social network data. *Human Organization, 35* (8)*,* 269–286.

Killworth, P. D. & Bernard, H. R. (1979). Informant accuracy in social network data III: A comparison of triadic structure in behavioral and cognitive data. *Social Networks, 2,* 10–46.

Klau, G. W. & Weiskircher, R. (2005). Robustness and resilience. In: U. Brandes & T. Erlebach, eds. *Network Analysis: Methodological Foundations*, pp. 417–437. Berlin: Springer-Verlag.

Klovdahl, A. S. (1989). Urban social networks: Some methodological problems and possibilities. In: M. Kochen, ed. *The Small World*, pp. 176–210. Norwood: Ablex.

Koehly, L. M. & Pattison, P. (2005). Random graph models for social networks: Multiple relations or multiple raters. In: P. J.

Carrington, J. Scott & S. Wasserman, eds. *Models and Methods in Social Network Analysis*, pp. 162–191. Cambridge: Cambridge University Press.

Koenker, R. W. & Bassett, G. W. (1978). Regression quantiles. *Econometrica, 46,* 33–50.

Krackhardt, D. (1987a). Cognitive social structures. *Social Networks, 9* (2)*,* 109–134.

Krackhardt, D. (1987b). QAP partialling as a test of spuriousness. *Social Networks, 9* (2)*,* 171–186.

Krackhardt, D. (1988). Predicting with networks: Nonparametric multiple regression analyses of dyadic data. *Social Networks, 10,* 359–382.

Krackhardt, D. (1994). Graph theoretical dimensions of informal organizations. In: K. M. Carley & M. J. Prietula, eds. *Computational Organizational Theory*, pp. 88–111. Hillsdale, NJ: Lawrence Erlbaum Associates.

Krackhardt, D. (1997). Organizational viscosity and the diffusion of controversial innovations. *Journal of Mathematical Sociology, 22* (2)*,* 177–199.

Krackhardt, D. & Stern, R. N. (1988). Informal networks and organizational crises: An experimental simulation. *Social Psychology Quarterly, 51,* 123–140.

Latané, B. (1981). The psychology of social impact. *American Psychologist, 36* (4)*,* 343–356.

Lazega, E. (2001). *The Collegial Phenomenon: The Social Mechanisms of Cooperation Among Peers in a Corporate Law Partnership*. Oxford: Oxford University Press.

Leenders, T. T. A. J. (2002). Modeling social influence through network autocorrelation: Constructing the weight matrix. *Social Networks, 24* (1)*,* 21–47.

Lorrain, F. & White, H. C. (1971). Structural equivalence of individuals in social networks. *Journal of Mathematical Sociology, 1,* 49–80.

McGrath, C., Blythe, J. & Krackhardt, D. (1997). The effect of spatial arrangement on judgments and errors in interpreting graphs. *Social Networks, 19* (3)*,* 223–242.

Mardia, K. V., Kent, J. T. & Bibby, J. M. (1979). *Multivariate Analysis*. London: Academic Press.

Marsden, P. V. (1990). Network data and measurement. *Annual Review of Sociology, 16,* 435–463.

Marsden, P. V. (2005). Recent developments in network measurement. In: P. J. Carrington, J. Scott & S. Wasserman, eds. *Models and Methods in Social Network Analysis*, pp. 8–30. Cambridge: Cambridge University Press.

Mayhew, B. H. (1984a). Baseline models of sociological phenomena. *Journal of Mathematical Sociology, 9,* 259–281.

Mayhew, B. H. (1984b). Chance and necessity in sociological theory. *Journal of Mathematical Sociology, 9,* 305–339.

Moody, J. & White, D. R. (2003). Social cohesion and embeddedness. *American Sociological Review, 68,* 103–127.

Moreno, J. L. (1934). *Who Shall Survive?* Washington, DC: Nervous and Mental Disease Publishing Co.

Morris, M., ed. (2004). *Network Epidemiology: A Handbook for Survey Design and Data Collection.* Oxford: Oxford University Press.

Newcomb, T. (1953). An approach to the study of communicative acts. *Psychological Review, 60,* 393–404.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review, 45* (2)*,* 167–256.

Pattison, P. & Robins, G. (2002). Neighborhood-based models for social networks. *Sociological Methodology, 32,* 301–337.

Pattison, P. & Wasserman, S. (1999). Logit models and logistic regressions for social networks: II. multivariate relations. *British Journal of Mathematical and Statistical Psychology, 52,* 169–193.

Pattison, P., Wasserman, S., Robins, G. & Kanfer, A. M. (2000). Statistical evaluation of algebraic constraints for social networks. *Journal of Mathematical Psychology, 44,* 536–568.

Pool, I. D. S. & Kochen, M. (1979). Contacts and influence. *Social Networks, 1* (1)*,* 5–51.

Rapoport, A. (1949a). Outline of a probabilistic approach to animal sociology I. *Bulletin of Mathematical Biophysics, 11,* 183–196.

Rapoport, A. (1949b). Outline of a probabilistic approach to animal sociology II. *Bulletin of Mathematical Biophysics, 11,* 273–281.

Rapoport, A. (1950). Outline of a probabilistic approach to animal sociology III. *Bulletin of Mathematical Biophysics, 12,* 7–17.

Richards, W. D. & Seary, A. J. (2000). Eigen analysis of networks. *Journal of Social Structure, 1* (1).

Robins, G. & Pattison, P. (2005). Interdependencies and social processes: Dependence graphs and generalized dependence structures. In: P. J. Carrington, J. Scott & S. Wasserman, eds. *Models and Methods in Social Network Analysis*, pp. 192–214. Cambridge: Cambridge University Press.

Robins, G., Pattison, P. & Elliott, P. (2001). Network models for social influence processes. *Psychometrika, 66,* 161–190.

Robins, G., Pattison, P. & Wasserman, S. (1999). Logit models and logistic regressions for social networks III. Valued relations. *Psychometrika, 64,* 371–394.

Robins, G., Pattison, P. & Woolcock, J. (2005). Small and other worlds: Network structures from local processes. *American Journal of Sociology, 110* (4)*,* 894–936.

Romney, A. K., Weller, S. C. & Batchelder, W. H. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist, 88* (2)*,* 313–338.

Scott, J. (1991). *Social Network Analysis: A Handbook*. London: Sage.

Seary, A. J. & Richards, W. D. (2003). Spectral methods for analyzing and visualizing networks: An introduction. In: R. Breiger, K. M. Carley & P. Pattison, eds. *Dynamic Social Network Modeling and Analysis*, pp. 209–228. Washington, DC: National Academies Press.

Shimbel, A. (1953). Structural parameters of communication networks. *Bulletin of Mathematical Biophysics, 15,* 501–507.

Snijders, T. A. B. (1981). The degree variance: An index of graph heterogeneity. *Social Networks, 3* (3)*,* 163–223.

Snijders, T. A. B. (1991). Enumeration and simulation methods for 0–1 matrices with given marginals. *Psychometrika, 56,* 397–417.

Snijders, T. A. B. (1996). Stochastic actor-oriented models for network change. *Journal of Mathematical Sociology, 23,* 149–172.

Snijders, T. A. B. (2002). Markov Chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure, 3* (2).

Strang, G. (1988). *Linear Algebra and Its Applications*, 3rd edn. Fort Worth, TX: Harcourt Brace Jovanovich.

Strauss, D. (1986). On a general class of models for interaction. *SIAM Review, 28* (4)*,* 513–527.

Sussman, G. J. & Wisdom, J. (2001). *Structure and Interpretation of Classical Mechanics*. Cambridge, MA: MIT Press.

Thompson, S. K. (1997). Adaptive sampling in behavioral surveys. In: L. Harrison & A. Hughes, eds. *The Validity of Self-Reported Drug Use: Improving the Accuracy of Survey Estimates*, pp. 296–319. Rockville, MD: National Institute of Drug Abuse.

Thompson, S. K. & Frank, O. (2000). Model-based estimation with link-tracing sampling designs. *Survey Methodology, 26* (1)*,* 87–98.

Tierney, L. (1996). Introduction to general state-space Markov chain theory. In: W. Gilks, S. Richardson & D. J. Spiegelhalter, eds. *Markov Chain Monte Carlo in Practice*, pp. 59–74. London: Chapman & Hall.

Tobin, J. (1958). Estimation of relationships for categorical and limited dependent variables. *Econometrica, 26,* 24–36.

Torgerson, W. S. (1952). Multidimensional scaling: I, theory and method. *Psychometrika, 17,* 401–419.

Travers, J. & Milgram, S. (1969). An experimental study of the small world problem. *Sociometry, 32,* 425–443.

Tufte, E. R. (1983). *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.

Valente, T. W. (2005). Network models and methods for studying the diffusion of innovations. In: P. J. Carrington, J. Scott & S. Wasserman, eds. *Models and Methods in Social Network Analysis*, pp. 98–116. Cambridge: Cambridge University Press.

Wasserman, S. (1987). Conformity of two sociomatrices. *Psychometrika, 52,* 3–18.

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

Wasserman, S. & Pattison, P. (1996). Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and *p**. *Psychometrika, 60,* 401–426.

Wasserman, S. & Robins, G. (2005). An introduction to random graphs, dependence graphs, and *p**. In: P. J. Carrington, J. Scott & S. Wasserman, eds. *Models and Methods in Social Network Analysis*, pp. 192–214. Cambridge: Cambridge University Press.

Watts, D. J. & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393,* 440–442.

Wegner, D. M. (1995). A computer network model of human transactive memory. *Social Cognition, 13* (3)*,* 313–339.

West, D. B. (1996). *Introduction to Graph Theory*. Upper Saddle River, NJ: Prentice Hall.

Willer, D., ed. (1999). *Network Exchange Theory*. Westport, CN: Praeger.

Wimmer, A. & Min, B. (2006). From empire to nation-state: Explaining wars in the modern world, 1816–2001. *American Sociological Review, 71* (6)*,* 867–897.

Yancey, W. L. (1971). Architecture, interaction, and social control: The case of a large-scale public housing project. *Environment and Behavior, 3,* 3–21.