COMPSTAT

1982

physica verlag · wien

Applied Graph Analysis in the Social Sciences: The Software Project GRADAP

C.J.A. Sprenger, University of Amsterdam, The Netherlands
F.N. Stokman, University of Groningen, The Netherlands

In Section 1 it will be shown that all main types of graphs have been used to represent social systems. Section 2 gives a survey of the possibilities of the package GRADAP (GRAph Definition and Analysis Package) for graph definition, management and analysis. GRADAP provides facilities to analyze all major types of graphs. A direct data interface with SPSS provides extra data manipulation facilities and ample opportunities for analyzing multigraphs and relating social network results to other characteristics of points and lines. In Section 4 the planned extension of GRADAP is discussed.

Keywords: graph analysis, social sciences, software project, social networks, computer package

1. Introduction

The seventies have shown an increasing number of applications of graph theory in the social sciences. A graph is an object that contains points and lines, each line being incident with two - not necessarily different - points. Lines that are incident with two identical points are called loops. Two points incident with the same line are adjacent. Lines incident with the same points are called parallel. In applied graph analysis in the social sciences the points of the graph represent social entities, the lines certain social relations between the entities. It depends on the concrete empirical application whether a direction is given to the lines (directed graphs), whether the lines have a sign (signed graphs) or a certain value (valued graphs), or whether different kinds of lines exist (multigraphs). By using graph-theoretical concepts and theories such graphs can be analyzed and new graphs may be generated. The results of these analyses are again empirically interpreted in terms of the social network itself.

All main types of graphs have been used to represent social systems. Moreover, applied graph analysis in the social sciences has been used not only to represent social systems between persons, but between organizations, roles and variables as well, as the following survey shows. The survey is not intended to be complete, but aims to illustrate the wide range of applications in the past.

In small group analysis and anthropology the points represent persons and the (directed) lines social relations. In case of one social relation, e.g. friendship, digraphs (directed graphs without loops and parallel lines) are used to represent the social system (Hallinan and Felmlee, 1975).

In case of several types of social relations and time series data multigraphs and valued graphs have to be used (see White et.al., 1976). When positive and negative social relations are considered, signed graphs are used as representation (Cartwright and Harary, 1956; Hoede, 1981, a,b). For the analysis of communication relations sim-

ple undirected graphs without loops and parallel lines and undirected multigraphs can be used (Bavelas, 1968).
In elite studies again the points represent persons, whereas the lines represent communication or influence relations. Simple graphs, multigraphs and digraphs are used to represent these social relations between elites (Kadushin, 1974; Laumann and Pappi 1976). In studies of cognitive maps and knowledge graphs the points represent variables and the (directed) lines causal relations. Signed graphs are used to represent such cognitive maps (Axelrod, 1976). In legislative analysis multigraphs and valued graphs are used to represent common activities of delegates (Stokman, 1977). Simple (di)graphs, multigraphs and valued graphs have been used to analyze intra- and inter-organizational relations between organizations (Fennema and Schijf, 1979).
Recent applications of graph theory in the social sciences make increasing use of models from which certain restrictions for social structures can be derived to test the models empirically. This development can be accelerated by graph-theoretical elaboration of existing measurement and analysis models in the social sciences. Examples of such elaborations and applications are the graph theoretical elaboration of log-lineair models (Fienberg and Wasserman, 1979), multidimensional scaling techniques (Laumann, 1973; Laumann and Pappi, 1976), stochastic cumulative scaling models (Stokman, 1977, 1980) and logistic models (Scheiblechner, 1971; 1972; 1977; Holland and Leinhardt, 1981).

2. The GRAph Definition and Analysis Package GRADAP

The package GRADAP (GRAph Definition and Analysis Package) provides facilities to analyze the main types of graphs, up to 6000 points and 60000 lines (Stokman and Van Veen, 1980). GRADAP is a batch-oriented system that interprets a user program, written in the GRADAP (application) language, and performs the tasks specified in that program. The GRADAP language has been modelled after that of SPSS, mainly to simplify its learning and its use. A second reason for this close relationship with the SPSS language will become apparent later.
GRADAP contains a number of appropriate modes designed for the analysis of simple graphs, digraphs and valued graphs with special facilities and adjustments of coefficients for bipartite graphs. Burt (1978) summarizes the different modes of network analysis that can be distinguished in the enormous flow of recent literature on applied network analysis in a typology with six categories, based upon two fundamental dimensions. For the points he distinguished three levels of analysis: individual points, subsets of points and the whole set of points in the graph. For each of these levels two approaches can be distinguished: the positional or attributional approach on the one hand and the relational on the other. On a less abstract level Freeman (1978) offers another typology which is primarily concerned with point and graph centrality measures. In this typology three basic structural concepts are distinguished, which are the basis of the different modes of centrality measures. The concepts are: degree (i.e. neighbourhood relations), betweenness (i.e. characterization by the way a point links other points) and closeness (distance). GRADAP has analytic modes and

special facilities for all these kinds of analysis. Freeman's coefficients of graph centralization have certain undesirable properties, however. Therefore a new coefficient of graph centralization has been developed at the University of Groningen with considerable better properties than Freeman's coefficients (see Section 3).

The data that make up a social network and any further (documentary) information is specified to GRADAP in the graph definition phase of a user program. This information may be saved in a GRADAP system file and accessed in later user programs, a mechanism analoguous to that supported by SPSS and other packages.

Besides this system file GRADAP also allows users to request part of the social network data to be stored in an SPSS system file (for subsequent analyses in SPSS, STAP and other packages that can access such a file).

It is also possible to have GRADAP read additional network data from an SPSS system file. This interface allows users to combine social network data analysis techniques available in GRADAP with the more standard types of data analysis available in SPSS and STAP and to use the data management facilities of SPSS and SIR. This forms the second reason for keeping the GRADAP language "in line" with the languages of those packages.

2.1. Graph definition in GRADAP

The definition of a graph or network in GRADAP consists essentially of the definition of the elements of a graph: the points in the graph or network, the information associated to the points (pointinfos), the lines, and the information associated to the lines (lineinfos). A further extension to this basic scheme is the possibility to group the points and the lines (during their definition) into pointsets and linesets, respectively. These sets will in general be disjoint, but overlap is allowed. Sets play an important role in the GRADAP language and a number of statements exist to define new sets or redefine existing ones. These will be discussed in the next subsection.

Both points and lines may be defined, together with any associated info values, either free formatted as part of the user program or by fixed format input. In the latter case that data may be either embedded in the user program or reside on an alternate input medium.

It should be realized that the flexibility of GRADAP's data structure is particularly due to the fact that is is not based on the commonly used representation of a graph by its adjacency matrix between the points. GRADAP's data structure consists essentially of a point data matrix in which the points are the cases and the pointinfos the variables defined on the points, together with a line data matrix in which the lines are the cases and the lineinfos the variables defined on the lines. The possibility to define (overlapping) point- and linesets increases the flexibility even more, because it enables a user to submit any (partial) (sub)graph for analysis.

For identification in the user program the graph as a whole, point- and linesets, point- and lineinfos are identified by unique names of up to eight characters. Further documentation is possible by labels.

Points may be identified by such names and documented by labels, but in some situations GRADAP allows labels to be used for identification as well. In user programs lines can only by referred to by their sequence number, a possibility which also exists for points.

Once a graph is defined it may be saved and accessed in later programs, thus providing a "canned" definition. GRADAP also supports addition of new points and lines to an existing graph and addition of values of points and lines in new point and lineinfos. Both types of extension may be carried out via fixed format input or, in the case of new point- or lineinfos, through SPSS system files.

2.2. Data management

GRADAP has many facilities to generate new graphs from the original graph. These data management facilities may be divided in facilities that modify the basic structure of the graph or network, i.e. create new points or lines and/or delete existing points or lines, and facilities that modify the set structure by defining new pointsets or linesets, or redefining old ones. Modifications that affect the basic structure of the graph or network are:

- induction, where one group of points induces a network formed by the points in another group and the lines that are "carried" by points of the first group;
- combination, where parallel lines are replaced by single new lines;
- condensation, where points in user-defined equivalence classes are replaced by single new points;
- sampling of points and/or lines, where the points and/or lines that do not belong to the sample are deleted;
- reversion of the direction of lines.

New sets may be defined (and already existing sets may be redefined) by enumeration of their constituting elements, by giving their definition in the form of a set expression (using set operators as AND, OR, XOR, NOT and MINUS), or by specifying a selection clause. The latter form allows users to define pointsets on the basis of pointinfo values, and linesets on the basis of lineinfo values and on the basis of pointinfo values for the endpoints of the lines.

Direction is an attribute of linesets: each lineset may be declared DIRECTED or UNDIRECTED. In the latter case, GRADAP will treat lines in the lineset as having no direction. Although the endpoints of each line have a fixed meaning (one is the TAIL, the other is the HEAD and the line, if treated as directed, is directed from TAIL to HEAD), they are treated as equivalent when the line is part of an undirected lineset.

2.3. Analytic procedures

The analysis phase of a GRADAP user program consists of an unlimited number of analysis blocks, each of which may be optionally preceded by temporary data management statements. These temporary modifications remain in effect until the analysis block is completed. Each analysis block starts with an ANALYSE SETS statement, that specifies the (partial) (sub)graphs to be analyzed in the subsequent block. The block is terminated by an END ANALYSE statement. Within an analysis block one or more analytical procedures may be activated, of which the following are currently available:

- ADJACENCY, that computes the adjacency matrix and optionally lists the isolated points and for each non-isolated point its neighbours.
- CENTRALITY, that computes point and graph centrality measures based on adjacency, distance or all sequences in the graph.
- DISTANCE, that computes the distance matrix and corresponding statistics based on distance and neighbourhood.
- REDUCE, that reduces the graph in an iterative process to the null graph, according to criteria based on adjacency and incidence.
- RUSH, that computes the centrality measure rush for points and optionally for lines. This measure is based on the structural property betweenness.
- SUBGRAPHS that detects subgraphs like (strong) (weak) components, blocks and N-cliques.
- VARIANCE DEGREE that computes measures of graph heterogeneity based on the variance of the degrees of the points.

Moreover several informative procedures are available, such as list/write graphinfo (cf list/write file info of SPSS) and list/write points/lines (cf list/write cases of SPSS).

3. An example of applied graph analysis with GRADAP

In a study on intercorporate structure in the Netherlands two types of relations between the largest 250 corporations and financial institutions are considered: financial participations and personal interlocks between the highest decision making boards (executives and directors) in these organizations. In the literature it is assumed that the tightness of an interlock depends on the combination of positions of the multiple director in the two corporations. To test this assumption three types of positions are distinguished in the Dutch corporations on the basis of involvement in decision making within a corporation. Chief executives are assumed to be involved in all major decisions within a corporation on a day to day basis; these positions are therefore considered to be inside (IN) positions. Chairmen of boards of directors, delegated directors and members of governing boards of cooperatives are involved only in major decisions, but on a more regular basis than the remaining positions; these positions are therefore considered to be intermediate (MED) positions. All other positions are considered to be outside (OUT) positions. Interlocks can now be classified into six types on the basis of all possible combinations of positions of a multiple director in two corporations: ININ, INMED, INOUT, MEDMED, MEDOUT and OUTOUTinterlocks.

We assume that the tightness of a type of interlock is reflected in the number of times that it coincides with other types of interlocks. For each type of interlock we therefore determined how often it co-occurred with another type of interlock or with a financial participation between the two corporations. With the help of the interface between SPSS and GRADAP together with several management facilities within GRADAP such as the possibility to combine parallel lines, the lines in the GRADAP file were transformed in such a way that each pair of corporations for which at least one relation existed, was connected by an undirected line with seven lineinfos, containing respectively the percentage of financial participation and the number of interlocks in each of the six types that were defined above. The results of these analyses are given in Tabel 1. The more outside the combination of positions, the looser the interlock, but INOUT interlocks are clearly tighter than MEDMED interlocks. This is supported, if we consider with which types of interlocks financial participations co-occur. Only in 3 out of 47 times, financial participations coincide solely with interlocks of a looser type than INOUT interlocks, whereas in 39 out of 47 times INOUT interlocks are among the co-occurring interlocks. Together with the results presented in Table 1 this is a strong argument to consider ININ, INMED and INOUT interlocks as primary interlocks and the other types of interlocks as secondary interlocks.

Table 1 (Co-)occurrence of financial participations and types of interlocks in the Dutch network in percentages (absolute numbers in parantheses)

	Fin.Part.	ININ	INMED	INOUT	MEDMED	INOUT	OUTOUT
Together with other type of relation	55(47)	67(2)	49(23)	37 (81)	31(19)	21 (52)	14 (71)
Sole relation	45(38)	22(1)	51(24)	63(136)	69(43)	79(200)	86(445)
Total	100(85)	100(3)	100(47)	100(217)	100(62)	100(252)	100(516)

Which consequences has the elimination of secondary interlocks for the centralization of the graph? Freeman's coefficients of graph-centralization are all based on the sum of the differences between the point-centrality of the most central point and all other points in the graph. The highest graph-centrality is attained for the graph that is a star. For graphs with different numbers of lines his coefficients are difficult to compare and are inadequate therefore. Snijders' index H of graph heterogeneity is superior for two reasons. It is based on the degree variance, which implies that differences of centrality of all pairs of points are taken into account. Moreover, it is normalized on the basis of the expected degree variance under the null model of the uniform distribution of all graphs with that number of points and lines. This makes it possible to compare heterogeneity of graphs with different numbers of points and lines (Snijders, 1981). Elimination of secondary interlocks reduces the graph heterogeneity from .25 to .22 in the large component of 190 directly or indirectly connected corporations. It gives however far more heterogeneity among the central corporations within the large component. In the network of all interlocks centrality of the corporations was determined on the basis of their degree. The 38 corporations with the highest degree were considered as the center, being the maximal subgraph of central corporations with diameter 2 in the graph as a whole. With a density of .36 the graph heterogeneity among these 38 corporations was .13. After elimition of secondary interlocks the density decreased to .09, whereas the graph heterogeneity increased substantially to .26.

4. Extensions

At the end of 1982 the second release of GRADAP will be delivered with two additional analytic procedures:

- TRIAD COUNT, to perform the triad census, as developed by Holland and Leinhardt (1970, 1975), but with special adaptations in case bipartite graphs are analyzed;
- SPATIAL AUTOCORRELATION, to compute coefficients of spatial autocorrelation.

In 1983 and 1984 extensions are planned, specially focused on the analysis of signed graphs. Moreover, at that time, a FORTRAN 77 version of GRADAP may be expected, that is portable to all major computers.

We thank W. Molenaar and W.H. van Schuur for their comments on an earlier draft.

References

Axelrod, R. (ed.) (1976), Structure of decision. The cognitive Maps of political elites, Princeton University Press, Princeton N.J.
Bavelas, A. (1968), Communication patterns in task-orientated groups, Cartwright D. and Zander A. (eds.), Group dynamics. Research and theory, Third edition, Tavistock, London.
Burt, R.S. (1978), Applied network analysis: An overview, Sociological Methods and Research, 7, 123-31.
Cartwright, D. and Harary F. (1956), Structural balance: A generalization of Heider's theory, Psychological Review, 63, 277-93.
Fennema, M. and Schijf H. (1979), Analysing interlocking directorates: Theory and methods, Social Networks, 1, 297-332.
Fienberg, S.F. and Wasserman S.S. (1979), Categorical data analysis of directed graphs: models for single generator, A.S.A. Proceedings Social Statistics Section, 1979, 407-412.
Freeman, L.C. (1978), Centrality in social networks, Conceptual Clarification, Social Networks, 1, 215-239.
Hallinan, M. and Felmlee D. (1975), An analysis of intransitivity in sociometric data, Sociometry, 38, 195-212.
Hoede, C. (1981a), The integration of cognitive consistency theories. Memorandum nr. 353, THT, Enschede, The Netherlands.
Hoede, C. (1981b), Social atoms. Part I: Kinetics. Memorandum nr. 372, THT, Enschede, The Netherlands.
Holland, P.W. and Leinhardt S. (1970), A method for detecting structure in sociometric data, American Journal of Sociology, 76, 492-513.
Holland, P.W. and Leinhardt S. (1975), Local structure in social networks, Heisse, D.R. (ed.) Sociological Methodology, 1976, 1-45, Jossey-Bass, San Francisco etc.
Holland, P.W. and Leinhardt S. (1981), An Exponential Family of Probability Distribution for Directed Graphs, JASA, 76, 33-50.
Kadushin, C. (1974), The American intellectual elite, Little Brown, Boston.
Laumann, E.D. (1973), Bonds of pluralism: The form and substance of urban social networks, Wiley, New York etc.
Laumann, E.D. and Pappi F.U. (1976), Networks of collective actions. A perspective on community influence systems, Academic Press, New York etc.
Scheiblechner, H.H. (1972), Personality and system influences on behaviour in groups: frequency models, Acta Psychologica, 36, 322-36.
Scheiblechner, H.H. (1971), The separation of individual and system influences on behaviour in social contexts, Acta Psychologica, 35, 442-60.
Scheiblechner, H.H. (1977), The social structure of large groups, Kempf W.H. and Repp B.H. (eds.) Mathematical models for social psychology, Huber H., Bern etc.

Snijders, T.A.B. (1981), The degree variance: an index of graph heterogeneity, Social Networks, 3, 163-174.
Stokman, F.N. (1977), Roll calls and sponsorship. A methodological analysis of Third World group formation in the United Nations. Sijthoff, Leiden, The Netherlands.
Stokman, F.N. (1980), Graph theoretical elaboration of cumulative scaling techniques, Quality and Quantity, 14, 277-88.
Stokman, F.N. and Van Veen, F.J.A.M. (eds.) (1981), GRADAP, Graph Definition and Analysis Package USER'S MANUAL. Volume 1 and Volume 2. Interuniversity Project Group GRADAP: University of Amsterdam-Groningen-Nijmegen.
White, H.C., Boorman, S.A. and Breiger, R.L. (1976), Social Structure from Multiple Networks. I Blockmodels of Roles and Positions, A.J.S., 81, 730-780.