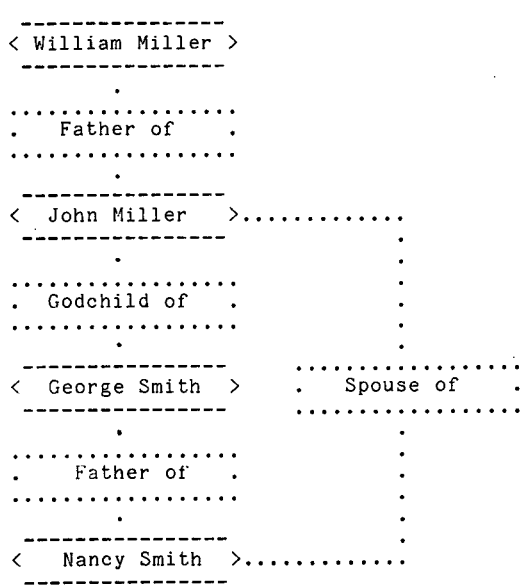
Historical Software Issue 6: Graph Definition and Analysis Package (GRADAP)

Thaller, Manfred

software

HISTORICAL SOFTWARE SECTION (1)

Using the Computer in itself does not necessarily have methodological implications for historial (or any) research; the availability of computer programs being capable of performing swiftly complex computations made necessary by some statistical approach most certainly has - as, to make the familiar quotation once more, the advent of systems like SPSS has shown quite convincingly.
Most of the systems used in that sense so far by historians implement a philosophy that centers around the wellknown "case - variable" concept. There are quite a number of research topics, though, which can not be adequately described by this approach. Take for example the following sketch:



What is important in this case is, that we have not one class of items relevant for analysis ("cases" like persons), but two of them: a set of items described by a number of variables (persons who have names, a date of birth, an occupation and so on) and linked by a set of other items, that is, relations between them, which themselves can be described by other variables as well (the most primitive one just being: "which kind of relationship - parental, marital and so on - is this?").

Objects like this one are called social networks. Networks are by no means restricted to the description of families: exactly the same kind of situation will arise if we do research upon, say, the internal structure of the decision processes in some legislative assembly, the exchanges between suitable objects of economic analysis and so on. As a matter of fact the tools developed for the analysis of such networks could be described as a distinct subfield of quantitative methodology in the Social Sciences.
The mathematical theory behind these methods is known as graph theory, as, in formal language, we call objects like the one depicted here "graphs", consisting of "nodes" or "points" ("Knoten" in German) linked by "edges" or "lines" ("Kanten").
The main problem for the analysis of such structures is obviously the detection of various classes of substructures and only later on the analysis of the causes that lead to their development - e.g. "is the selection of a mate from ones godfather's offspring influenced by the social status of the two families?"
That second part of the question can easily be reduced to familiar methods, as it is just another way of posing the eternal: "How and why are two variables related?". To express our interest in conventional statistical terms as can be understood by a computer we have to know, though, already at the beginning of our analysis exactly which relationships between which people we want to analyze. Finding out about the existence of such relationships would have to be done before and presumably by hand. To understand why this is not so easy to do, just try to enter data about a very simple kind of relationship ("who selected whom as godfather?") into SPSS, try what happens if you produce a CROSSTABS for that question within a population of 1000 persons (containing probably something like 100.000 cells) and consider that after this effort you have just illustrated the most obvious relationship - all the less trivial ones like: "Did pre-marital family links between spouses exist?" are not only not taken care of by your analysis, but with conventional statistical programs not even possibly be studied with the help of your data.
To overcome such difficulties, developments for the automatic detection of certain classes of substructures within social networks started rather early and quite a number of computer programs - asking for a varying degree of sophistication of the user - exist. Unfortunately, due to the limited facilities of the sixties and early seventies when most of such developments started of them were intended for the analysis of extremely small populations - finding out about what's going on within a small group of, say, 30 to 100 people. The vast statistical and computational effort going into somehow limited aspects of such research actually led to the accusation, that formal methods in network analysis attempted to "kill flies with dynamite" - an accusation probably not completely unfamiliar to most people trying to use formal methods at all.
Still, the problems just mentioned are there and, if computer programs capable of analysing really big networks existed, being usable on the same level of sophistication as SPSS, supporting a

convenient form of input and providing for an interface into a more conventional package to take care of the more simple computations without the need to input the data once again, the whole approach of analyzing with formal methods large sets of information about social interrelations might gain quite some importance for quantitative methodology in historical social research. - "Finding out about structures and processes" being the alleged aim of a very large amount of research, though those terms, implying at first look a very precise terminology, are often used in so haphazard a way, that they have in many cases to be considered as hermeneutic labels rather than as part of an hypothesis open to falsification.
Such a system has arrived and is available under the name of GRADAP (though an important restriction to that availability will have to be mentioned later).
GRADAP(2), an acronym that stands for "Graph Definition and Analysis Package" allows the handling of networks of up to 6000 points and 60.000 lines. While some of the community analyses under way currently consider many more than 6.000 persons, it should be quite possible to formulate all research tasks undertaken in a way that fits into the limits mentioned. The same holds true for the second class of limitations the package has: every point (or node or person or institution) can be described by up to 150 variables; the same limit applies to the number of variables that can be defined for the lines (or edges or relations) which connect the points.
The package has two outstanding features:

- it has a control language which is designed so closely along the lines of SPSS that a user can simply understand it as an extension of that package. As a matter of fact some of the commands like RUN NAME or VALUE LABELS are identical; the TO convention can be used in the same sense as within SPSS. Most of the other commands should be understandable immediately when one has understood the concepts of a graph consisting of points and lines described by "informations" as the variables are called. So instead of GET FILE/SAVE FILE you use GET GRAPH/SAVE GRAPH and instead of a VARIABLE LIST you have to define a POINTINFO LIST and a LINEINFO LIST. This possibility to understand GRADAP as a valuable extension to a package used by many people in social historical research already, is further enhanced by a very well developed interface with SPSS, which allows to read SPSS system files - or parts thereof - to get the necessary "point- and lineinfo's" for the description of the entities within a given graph. This process can be reversed as well, that is, you can output those informations as SPSS system files.
- GRADAP offers very convenient facilities for the input of graphs consisting of very many points, each of which is only connected to a relatively small number of the others (as will usually be the case in imaginable studies in social historical research).

To describe the simple network given as initial example we might proceed as follows:

```
RUN NAME          DEFINITION OF A SIMPLE GRAPH
GRAPH NAME        GODMATES.
POINT IDENTS      NAME
POINTINFO LIST    BIRTH, JOB, SOMETHING
POINT LIST        WMILLER (1775,12,14)
                  JMILLER (1797,1,24)
                  GSMITH  (1768,53,19)
                  NSMITH  (1795,0,0)
LINEINFO LIST     TYPE
LINE LIST         WMILLER:JMILLTER('F')/
                  JMILLER:GSMITH('G')/
                  GSMITH:NSMITH('F')/
                  JMILLER:NSMITH('S')
MISSING VALUES    BIRTH TO SOMETHING(0)
PRINT FORMATS     TYPE(A)
SAVE GRAPH
```

This is an example that should give you a feeling for the control language; its about as exhaustive as an attempt to describe SPSS within 16 lines would be. While not going into detail the following points should be emphasized:

- for clarity's sake we have choosen free field input here; fixed field is available as well (and considerably more efficient).
- points can (but need not) be identified by mnemonic abbreviations.(Though in the case of large nominative datasets an identification produced automatically by a record linkage system together with a "label" that can consist of the full name, will be more sensible.)
- lines are defined by the identification of the point the start from (the "head") and the point the go to (the "tail"). Facilities for abbreviating long lists of lines with the same head and/or tail exist.
- lines can have a numerically expressed "multiplicity", i.e. it can be announced to the system that there exists more than one connection between two given points.
- it can be indicated if lines are "directed" or not. ("'A' being the father of 'B'" is a relation that can not be reversed; when we define it, it is important who is the father and who is the son. "'B' being the spouse of 'C'" implies on the other hand that 'C' is the spouse of 'B', so it is irrelevant which of the two points is considered the starting point or head of the line.)
- points as well as lines can be subdivided into "pointsets" and "linesets" - roughly equal to the "subfiles" in SPSS but of more analytical relevance, so the ANALYSE SETS, while being somewhat analogous to RUN SUBFILES could be explained only after a more thorough discussion of the principles of graph analysis as implemented in GRADAP.

- extensive labelling facilities exist.
- graphs stored can be expanded at any later stage by the use of ADD POINTS and ADD LINES (roughly equivalent to ADD CASES) and of ADD POINTINFOS and ADD LINEINFOS (roughly equivalent to ADD VARIABLES) with the additional facility, already mentioned, of adding complete SPSS system files as lineinfos and/or pointinfos.

So much for input and structuring of data.
GRADAP provides an arsenal of commands for transformations of stored graphs as well. It provides only those transformations, which are specific for the handling of transformations of graphs though - so there are no direct equivalences to the COMPUTE's, RECODE's and IF's of SPSS which might be applied to the lineinfos and pointinfos as such. I would not consider this as a shortcoming of GRADAP: as the interface between the system and SPSS is very good, one should always be able to convert with minimal effort (one command) the point- and/or lineinfos into SPSS system files, transform them as one is up to and read them back into GRADAP with another command. Knowing that the reinvention of the wheel is a vice indulged in by all to many producers of software already, one might even recommend this as an extremely wise decision in the design of the package - we will nevertheless have to mention a somewhat unfortunate consequence of it later.
Graphspecific transformations exist in six ways:

- new subgraph (or "pointsets" and "linesets" in the GRADAP language) can be created by either enumerating their constituent points or lines or by their deduction from existing point- or linesets based on expressions which define them as sub- or supersets of the already existing structure of sets, employing expressions which use settheoretical operators. (It's considerably easier as it may sound in abbreviated description.) Furthermore one can define new point- or linesets by formulating logical conditions (as in IF and SELECT IF of SPSS) which a point or line has to fulfill to qualify for the newly introduced set.
- random selection of new graphs based on the random selection of a subset of points or lines of the stored one is possible.
- the points within a graph can be "condensed", that is, (talking somewhat simplifying) a number of points which fulfill a logical criterion can be drawn together into a kind of "superpoint". So you might start with analyzing the existing relationships between persons, perform a CONDENSE IF and continue with analyzing a new network that describes the relationships between e.g. categories of jobs. (Avoiding the bright new possibilities for ecological fallacies remains in the responsibility of the user.)
- similiarly the lines within a graph can be "combined", that is, lines which connect the same pair of points can be reduced to one.
- furhtermore new graphs can be "induced" from the existing one. This principle is hardest to explain without more graphtheoretical explanations. Just for illustration: a possible induction might

be to create in our initial example a network of the parental generation, where the "lines" between two persons represent the marriage between a son and a daughter who, being not in the same generation, do not appear in the network as points in their own right any more.

- finally a graph (or a part thereof) can be reverted: when you start with a graph containing directed lines from father to son you can turn it into another one, containing directed lines from son to father.

All transformations can be executed permanently or temporarily (in the same sense these terms have within SPSS).
Remain the analytical procedures. As we mentioned already the primary aim of graph analysis is the detection of structures within a graph, the discovery, which kinds of relationships exist within a social network. This statement is about as broad and "exhaustive" as the possible other one that the purpose of table and regression type analysis is to discover relationships between variables. While skipping all the details we can simply say that GRADAP provides a sound number of analytical procedures to detect a variety of structures within a network, to find out successively as well about the relative importance of a given class of points (persons, institutions) within the network as about the typical ways in which points interact. Still, as the existence of REGRESSION in SPSS doesn't make the user an expert in the interpretation of regression equations, the existence of SUBGRAPHS within GRADAP does not in itself explain, how to interpret the frequency of a given type of "n-clique" in one's data. - And both problems go beyond the evaluation of the potential usefullness of a software product.
What is relevant in our context, is the quèstion how well documented the particular system is. The manual decribing GRADAP's use is generally quite well written, contains a large number of illustrative sketches and many reproductions of exemplifying printouts. It even tries to give an introduction into graph theory and is certainly adequate to describe what is meant by the output produced by the system. Still one has to emphasize quite strongly that the manual will suffice as explanation what is meant by a particular type of analysis only for someone, who is used at least slightly to formal statistical language. If one intends to use the system, who can consider himself a reasonably experienced user of SPSS, it should take him or her a day to produce the first meaningful output. To interpret that output correctly, one has nevertheless to invest at least a week or two to raise one's level of statistical expertise beyond Chi Square and the rest.
So what remains as general picture? GRADAP is a software product that is of considerable potential value for historical social research; it implements methods which are rather refined and complex in a way that makes them available almost without technical knowledge - but the methods as such ask the user still for a greater investment of understanding than the ubiquitious crosstabulation.

GRADAP would be very useful for quite a lot of research projects - and, as the example of SPSS has been proving, methodological understanding spreads sometimes quite swiftly, if the technical problems of applying the method dwindle. Can the system be made available? It is currently qouted at 500 $ for the initial licence and 200 $ for the annual renewal. This should in principle be in the possibilities of all academic computing centres - provided there exists a group of interested users of more than just one or two historians. GRADAP might be a good occasion to drop in at the sociological and psychological institutes of your university, finding out about the possibilities of joint lobbying at your computing centre.

A serious restriction is that the system can currently be used at CDC computers only. So everybody interested in a non-CDC version would have to provide for the conversion first. This is certainly no easy task to start with; and the distribution policy of GRADAP will support a conversion only, if a computing centre is ready to implement it institutionally - so you will have to muster considerably more support for the system, as if there might be a chance to try its conversion within the context of some large scale research project.

Another word should be said about the relationship to SPSS: we should leave no doubt that the close link between the two systems is very commendable - but when calculating the probable cost of a conversion, it seems to be a handicap. As I have to admit GRADAP is, due to its being a very new system that could not be used at the computing centre I have access to, the first one that has been reviewed here, without my having actually used it or having had at least a possibility to talk longer to more experienced users. After quite some practical experience with CLUSTAN - another package without transformation facilities - I guess, though, that in practical work with the notoriously bad data available for historical social research, one will have to rely very much on the possibilities of SPSS for transformations within the point- and linesets. This can, if need arises, be done via formatted output and a definition of a new SPSS system file as well, but one should say that the GRADAP-SPSS interface is in all probability not only a fine thing to have, but a very essential one for GRADAP's operation. And SPSS system files change their internal format sometimes - so a GRADAP conversion could probably be undertaken only by an institution that can muster sufficient manpower to guarantee that changes in SPSS are taken care of at a reasonable speed.

And most serious of all: while upholding everything that was said in favour of the system, one has to express some doubts if GRADAP, as it is available right now, was very well prepared for conversion. As a matter of fact the first thing started after the initial release is an attempt to convert it from the original "old" FORTRAN it was written in, to FORTRAN 77 - and this conversion (with the additional aim of faciliating further ones, though) from one compiler to the next one - on the very same machine - seems to be seen as a major effort by the original producer of the system.

An additional point: the system uses assembly language routines for some purposes, where they are "unavoidable" according to the available documentation. Unfortunately it is not specified what these purposes are, but one has to guess, that beyond the usual circumvention of what FORTRAN considers an efficient I/O, there will be a lot of handling of bitstrings to be done and, even if this should be taken care of by "trick programming" in FORTRAN, it is bound to reflect the word length of the source computer. Even if CDC would not support the exotic length of 60 bits a word, this creates a very large number of pitfalls for "how not to program parametric and hamper transportability".
So the conclusion has probably to be that GRADAP, while of great potential value, can be considered at present only by people who have access to CDC's while everybody else will just have to wait. (Not too long - hopefully.)

FOOTNOTES

1 Address all communications to: Manfred Thaller, Max-Planck-Institut für Geschichte, Hermann-Föge Weg 11, D. 34 Göttingen

2 System, manual (2 volumes with approximately 400 pages) and further information are available from: The inter-university project group GRADAP, Technisch Centrum FSW, Universiteit van Amsterdam, Roetersstraat 15, 1018 WB Amsterdam.