Social Network Analysis: An Overview by M. Pearson

Introduction

Social network analysis has roots as far back as the 1930's under Moreno (1932, 1938) and developed rapidly during the 1970's with the availability of fast computers and well developed graph theoretical concepts. Since then, to quote the INSNA (2013) website, "it has found important applications in organizational behaviour, inter-organizational relations, the spread of contagious diseases, mental health, social support, the diffusion of information and animal social organization".

Objectives

The objectives of this short paper are to identify and categorise important areas of current interest and research as well as data sources for social networks. I will begin by reviewing methodologies associated with health issues and, in particular substance use behaviour. This is because that is where I started my interest in social network analysis, though my interest in graph theory began some while ago when studying mathematics, and especially combinatorial theory, as an undergraduate. This review will cover the following areas:

1.) Complete data sets
2.) Longitudinal data sets with complete data
    a. Dynamic networks and graph visualisation
    b. Interdependence of network and behaviour
    c. Influence and selection
3.) Incomplete data sets
    a. Ego-centred (personal) network sampling
    b. Snowball sampling
4.) Longitudinal data sets with incomplete data
5.) Data sources
    a. Primary data sources
    b. Secondary data sources
6.) An Example of Ex-Offender Rehabilitation Survey Questions
7.) Conclusion

1. Complete Data Sets

A complete data set contains information about all of the participants involved in the study. The term derives from sociocentric network analysis and involves the quantification of relationships between people within a defined group on the assumption that members of a group interact more than would a randomly selected group of similar size. This enables the study of structural features of the network such as power, influence and selection which are often generalised from the particular study. Figure 1, for instance, illustrates the first wave of interviews carried out in a longitudinal survey of a year group of adolescents in a school in Glasgow. All (well, nearly all) the children in the year group were asked to complete questionnaires giving details of their behaviour patterns and named friends. The network connections (well, the first 6 connections named by the respondents) are all included in the diagram together with the behavioural and attribute characteristics of each respondent.

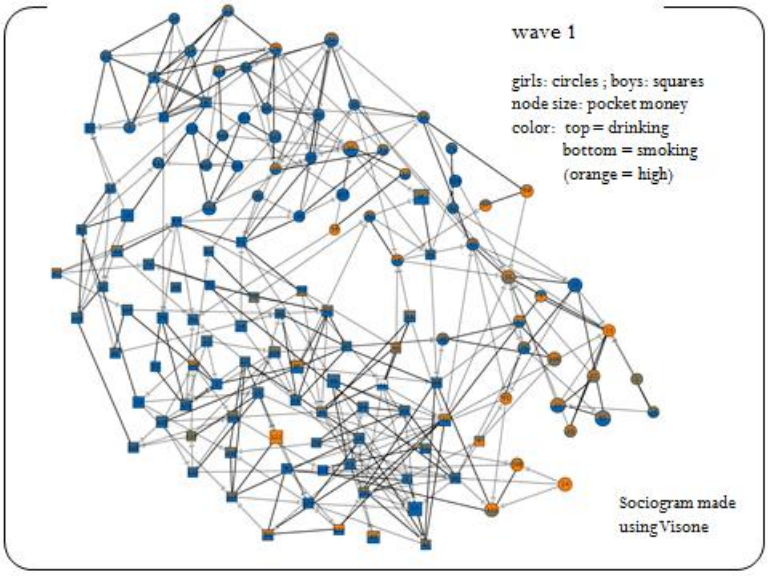

Figure 1

I quote from Steglich et al. (2010): "*Complete network* studies (i.e., measurements of the whole network structure in a given group) are preferable to personal (ego-centred) network studies, because selection patterns can best be studied when the properties of non-chosen potential partners are known, and because of the possible importance of indirect ties (two persons having common friends, etc.) that are difficult to assess in personal network studies. The interdependence of individual observations in complete networks, though, rules out the application of statistical methods that rely on independent observations".

2. Longitudinal Data Sets (Panel Data) with complete data

This arises when the survey is repeated at regular (or not so regular) intervals of time to enable research into underlying processes that occur. We will illustrate with a complete data set, though it is possible to extract valuable information from incomplete data sets using longitudinal data. Figures 2 and 3 show sociograms for waves 2 and 3 of the panel data extracted from the Teenage Friends and Lifestyle Study (Pearson & Michell, 2000; Pearson & West 2003) making use of the network visualisation package, VISONE.

2a. Dynamic Networks and Graph Visualisation

Figures 1, 2 and 3 together give a moving picture of the way the network and behaviour change together. The software package SIENA together with the visualisation package VISONE can provide an evolving picture of the most likely pattern of network and behaviour changes that give rise to the events recorded at each observation.

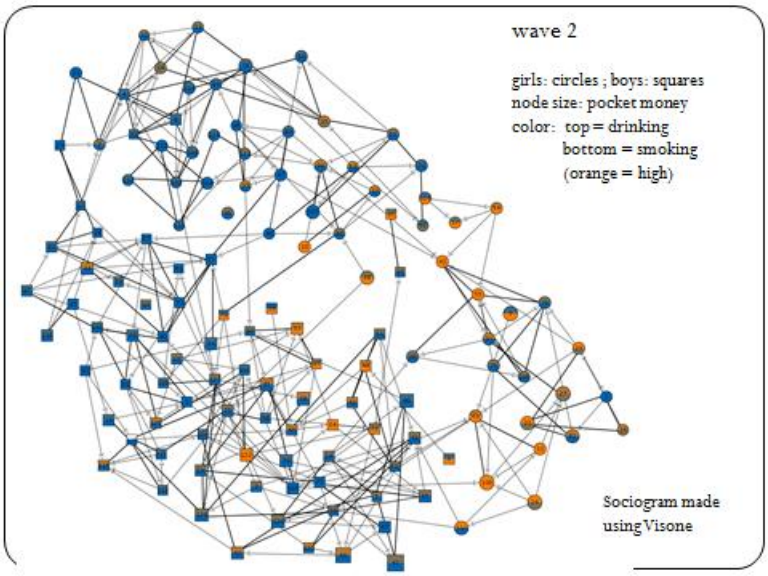

Figure 2

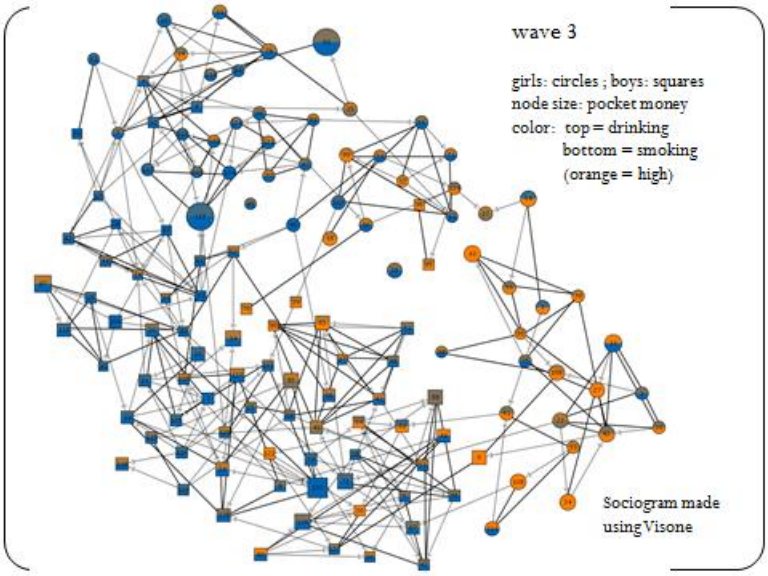

Figure 3

2b. Interdependence of Network and Behaviour

In social groups, there generally is interdependence between the group members' individual behaviour and attitudes, and the network structure of social ties between them. Figure 4 shows the way in which individuals change both network ties as well as behaviour and the way in which these can be inter-related.

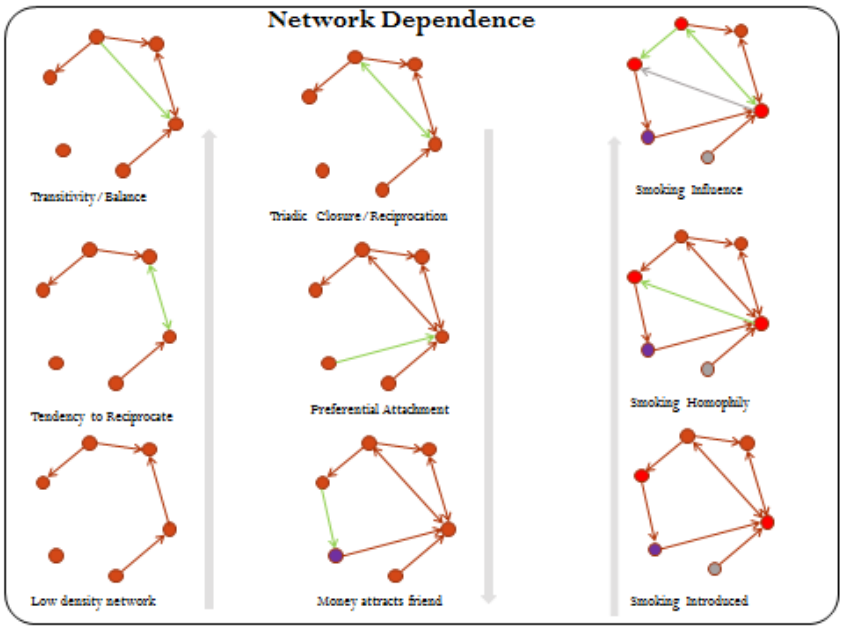

Figure 4

Figure 5 illustrates the situation where there is latent or hidden change between one observation and the next. It is difficult to model this sort of change as several unobserved changes may have taken place between actual observations. This leads to complexity in the modelling and the need, for example, to use Markov Chain Monte Carlo methods which help to model the changes between observations using random graph methods.

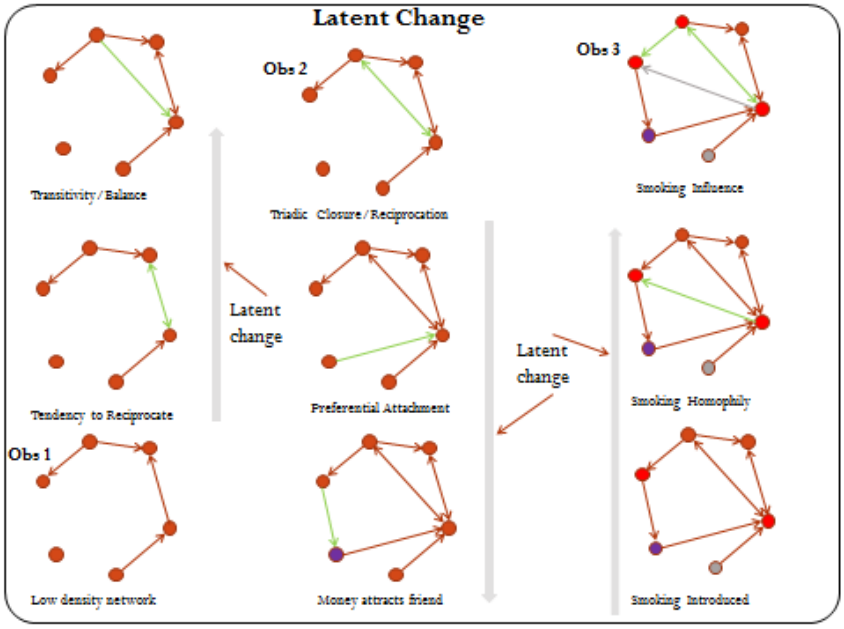

Figure 5

2c. Influence and Selection

Figure 6 illustrates the processes of influence and selection taking place in our illustrative example.

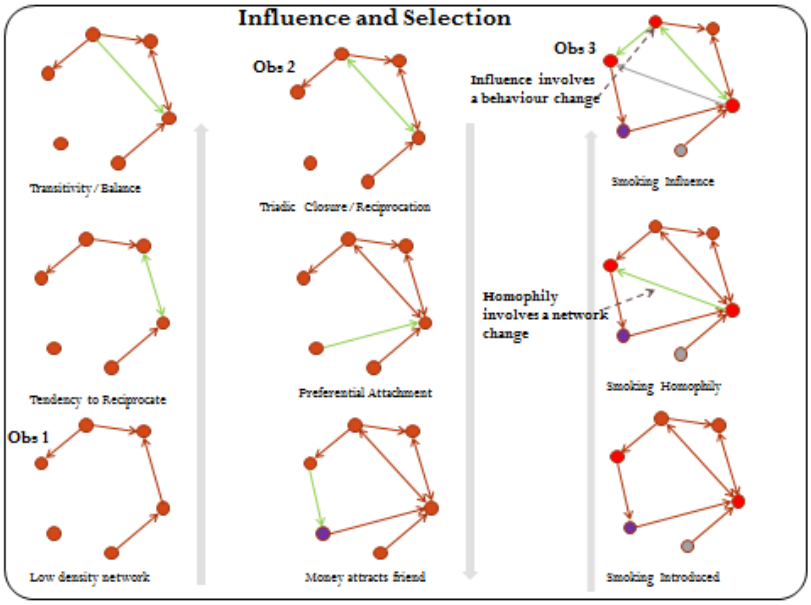

Figure 6

Homophily ('Birds of a feather flock together', McPherson et al., 2001) occurs when an individual makes a choice to form a tie with another individual with similar behaviour. This therefore involves a network change. Influence, on the other hand, occurs when an individual changes his or her behaviour to match that of another individual with whom they have an existing tie. This involves a behaviour change. In our study (Steglich et al., 2010) we investigated the degree to which influence and selection mechanisms account for observed coevolution of substance use and friendship ties in the data.

3. Incomplete data sets

We examine data sets where not all of the named individuals are respondents in the survey or where only partial data on networks of individuals is available.

3a. Ego-centred (personal) network sampling

The egocentric (personal) network approach arose from anthropology with its interest in people rather than groups. An egocentric network comprises the people (alters) known by an individual (ego). To gather this information it is only necessary to interview the 'ego' and so detailed information about the 'alters' is less reliable than it would be if they also had been interviewed. However the information provides a profile of the ego's personal network. In the case of a doctor, for instance, this may include colleagues at work, family members (including their friends), friends and patients. A senior consultant would be likely to have a different personal network pattern from a junior doctor with perhaps more work colleagues and fewer friends of family members. Egocentric networks enable research to be carried out into the personal features of networks such as the consumer behaviour or economic success of the participants and their relationship with their community.

Egocentric network analysis focusses on the diversity of the social environment and then applies survey sampling techniques to allow results to be generalized. For instance a teacher's egocentric network might include her doctor as well as her colleagues and their friends. The doctor's network might overlap with the teacher's and so have shared influences. Interviewers rely on respondents reports to get information about their relationships with the 'alters' and even, on occasion, between the 'alters' (e.g. does your friend Peter know Mary?). In social support studies respondents may be asked to name up to three or five individuals on whom they rely for help or advice or talk to about important matters such as health care decisions. In small world studies (Milgram, S., 1967), respondents may be told the name, occupation, and city of residence of some target person and are asked to mail a packet of papers to the that person only if they know the target personally. If respondents do not know the target personally, they are asked to send the packet to someone who they do know and who they believe has a chance of knowing the target. Tracking the path of the packets provides information on how people know each other and on the average number of links between pairs of randomly chosen people in a large society. For further information on this topic, such as identifying the structure of an egocentric network and small world phenomena visit http://www.bebr.ufl.edu/files/SNA_Encyclopedia_Entry_0.pdf.

As an illustrative example suppose a survey was carried out to discover where individuals in a community find out information about events and activities. A questionnaire asks individuals in Southpond to name up to five information sources for local events and activities. The responses were coded into UCINET (Borgatti et al.) and NETDRAW (Borgatti) and exported to GEPHI, which is a network visualisation package. The results are illustrated in the sociogram of Figure 7.

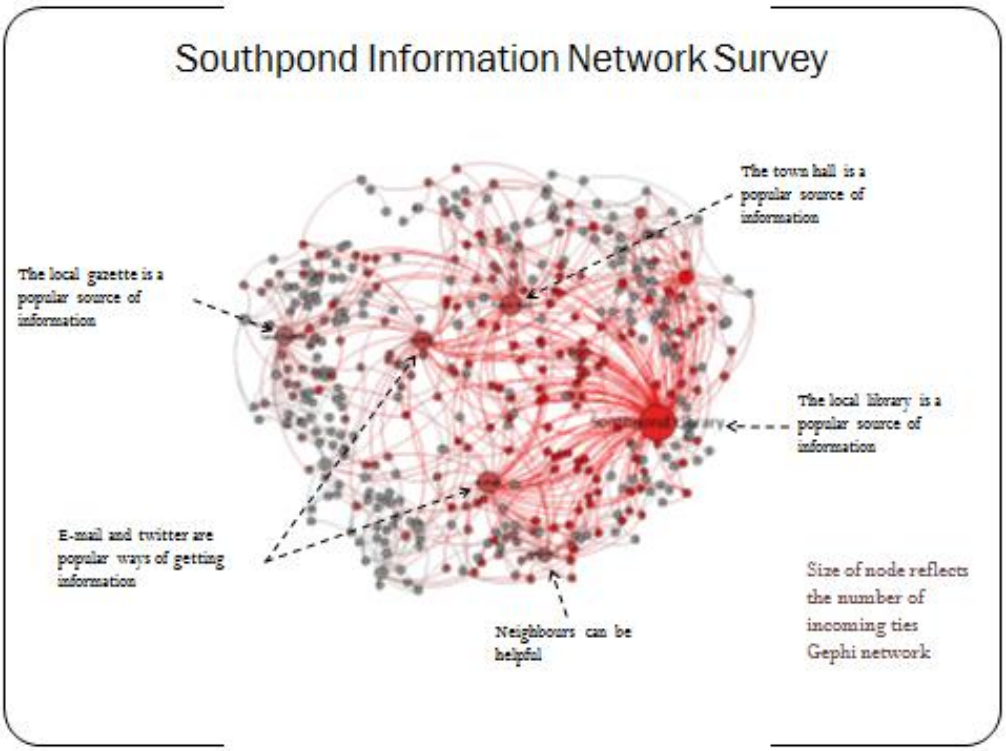

Figure 7

We can see that the local library, the town hall and the local gazette are popular sources of information with many ties while email and twitter are popular ways of getting information.

3b. Snowball Sampling

In social networks analysis (as in other methodology) snowball sampling is a technique where existing subjects in a study recruit future subjects from among their acquaintances. Further samples are then obtained by recruiting from the acquaintances which makes the original sample appear to grow like a snowball. Snowball sampling, therefore, uses existing social networks between members of a target population to build a sample. It is more directed than many other non-random sampling techniques, such as convenience sampling that focuses only on the most easily identified and reachable members of a population. To acquire a sample a 'seed' is needed usually from known individuals engaging in the behaviour under scrutiny. Miller (2013) describes an example of snowball sampling to identify users of Landsat imagery in the USA (http://www.fort.usgs.gov/landsatsurvey/SnowballSampling.asp) which started with a web search for organisations using Landsat and then followed up with requests to these organisations to provide contact lists of members or send out snowball sampling requests to those members. One disadvantage of snowball sampling is that it suffers from bias such that a person with many friends is more likely to be recruited into the sample. Methods, such as respondent-driven sampling (Heckathorn, D., 1997; Snijders, T., 1992) can assist in reducing the bias.

4. Longitudinal data sets with incomplete data

Longitudinal data sets with incomplete data occur when the list of members involved in the survey is too large to collect all the data on ties between respondents. Various methods are used to analyse such data depending on the detail of the tie structure and strength. These methods mostly result in egonet measures such as changes in egonet centrality, density and structure over time each measure being attached to the 'ego' (respondent) reporting details of the social network information. Standard statistical analysis, such as linear or logistic regression, are commonly used to establish a relationship between a dependent variable, such as employability, and a number of explanatory variables including network variables such as density of friendship network or number of ties in advice ego-network. Hence most analyses of egocentric network data summarize the composition of the network as a set of variables that become attributes of the respondent. These variables are can then be examined over time giving a general picture of the progress of change. It is not, however, possible to make complete sociograms, such as those displayed in Figures 1-3, since the data on all ties is not available. Instead a picture of network progress can be developed from the egonet ties as they change over time.

5. Data sources

Primary data refers to data gathered by the researcher while secondary is gathered from previous work carried out by another source.

5a Primary data sources

Primary data is sourced from carrying out a survey/questionnaire (either person to person or telephone or internet/web based) asking individuals for information on network ties and other information such as attribute data.

5a(i) Example of Primary Data Source

Table 1(values calculated using a random number generator) illustrates the sort of data that might be gathered from carrying out a survey of adolescent teenagers (see, for example, Figure 1). The first column contains the ID numbers of the respondents answering the questionnaire. The next three columns give data on the respondents smoking, drinking and pocket money (the column on gender is not given here) with 1 representing low level and 5 high level. The remaining columns provide information on the identity of the respondents

ID	Smoke	Drink	Money	Friend1 ID	Best Friend	Friend2 ID	Best Friend	Friend3 ID	Best Friend	Friend4 ID	Best Friend	Friend5 ID	Best Friend
12124	4	4	5	12171	3	12213	2	12005	3	12092	2	11983	2
12117	2	5	4	11985	2	11979	2	11863	3	12014	2	12128	1
12070	2	3	2	11940	2	12388	2	11924	2	12118	1	12008	3
14135	1	5	5	14276	1	14113	2	14112	2	13962	2	13736	3
14053	5	1	3	14236	2	14196	3	14023	3	14016	2	13971	1
14082	2	3	1	13952	3	14157	1	14056	3	14062	3	14093	1
14126	4	2	3	14314	3	14381	3	14284	1	14101	2	14190	2
24001	2	4	4	24284	1	24265	1	23758	1	24086	3	24043	2
32201	2	3	5	32333	2	31866	3	31820	1	32090	2	31830	3
32032	1	1	3	31947	1	31879	3	32012	2	32141	1	31840	2
32209	1	4	1	32313	3	31979	1	32223	2	32204	3	32128	1
32172	3	2	4	32285	3	32369	2	32324	1	32172	2	32126	3
32043	1	2	3	32294	3	32090	3	32148	3	32133	2	32012	1
32214	2	3	3	32371	3	32161	2	32224	2	32123	3	32026	3
32001	4	5	4	32197	3	32022	2	32063	3	32099	3	31935	1
32138	3	2	4	32047	1	32430	2	32084	1	32139	1	32427	3
34032	1	2	1	33877	3	33939	1	34234	1	33924	2	34284	2
34151	5	1	3	34112	3	34087	3	34024	2	34125	2	34269	2
34210	4	4	4	34283	2	33968	3	34021	2	34022	1	33881	3
42125	3	1	3	42060	1	42385	1	42117	3	42090	1	42220	1
42091	3	4	2	42290	2	42147	3	42270	3	42054	3	42365	3
42176	5	2	5	41968	1	42100	2	42159	1	42216	1	42416	1
42134	1	4	3	42244	3	42010	1	42372	1	42245	1	42158	1
42250	5	1	3	42256	1	42488	3	42236	2	42216	1	42124	2
42211	2	3	2	42433	2	42436	1	42143	2	42201	3	41995	1

Table 1 Friendship data with five best friends

five best friends together with the strength of the friendship (1 for just a friend and 3 for very close friend). The data can be used to draw sociograms and carry out social network analysis on the structure of the friendship networks.

5b Secondary data sources

Secondary data is sourced from previous data gathering and research exercises. Some examples are given below. These can be used to carry out social network research into areas such as wellbeing, employability, social capital and health.

5b(i) Scottish Household Survey

Data is available from the Scottish Household Survey and can be obtained by a SHS special data request (http://www.esds.ac.uk/findingData/snDescription.asp?sn=6361). There is longitudinal data (2005-2011) available for the Scottish Household Survey since it is continuous and based on a sample of the general population in private residences in Scotland. The aim of the survey is to provide representative information about the composition, characteristics and behaviours of Scottish households, both nationally and at a

more local level. The survey covers a wide range of topics to allow links to be made between different policy areas, with a particular focus on information to aid policy decisions on transport and social inclusion. Some of the methods outlined in sections 3 & 4 can be used to analyse this as a longitudinal data set.

5b(ii) Scottish Social Attitudes Survey

The Scottish Social Attitudes Survey (http://www.scotcen.org.uk/series/scottish-social-attitudes) focuses on how the public's views have changed over time and could be used in combination with other surveys to carry out a longitudinal investigation of social capital and other issues. Findings from the study inform policies aimed at changing or challenging attitudes, as well as policies that seek to change behaviours that may stem from these attitudes.

5b(iii) Growing up in Scotland Survey

Another survey which could be used is the Growing up in Scotland (GUS) Survey (http://www.crfr.ac.uk/gus/index.html ; http://www.scotland.gov.uk/Publications/Recent), which also has a longitudinal dimension. GUS follows the lives of thousands of children right across Scotland from infancy through to their teens. As one of the largest studies ever done in Scotland it will provide information to help develop policies and plan services for children and their families.

5b(iv) National Child Development Study (Centre for Longitudinal Studies)

(http://www.esds.ac.uk/longitudinal/access/ncds/l33004.asp). The NCDS is a continuing longitudinal study that seeks to follow the lives of all those living in Great Britain who were born in one particular week in 1958. The aim of the study is to improve understanding of the factors affecting human development over the whole lifespan.

5b(v) British Cohort Study (http://www.esds.ac.uk/longitudinal/access/bcs70/l33229.asp)

The British Cohort Study 70 (BCS70) began when data were collected about the births and families of babies born in the UK in one particular week in 1970. The first wave, called the British Births Survey, was carried out by the National Birthday Trust Fund in association with the Royal College of Obstetricians and Gynaecologists. Its aims were to examine the social and biological characteristics of the mother in relation to neonatal morbidity, and to compare the results with those of the National Child Development Study (NCDS), which commenced in 1958 (held separately at the UK Data Archive under GN 33004 – see above).

6. An Example of Ex-Offender Rehabilitation Survey Questions

The questions depend on the objectives of the study. Suppose we want to find out about employment prospects for ex-prisoners. We would first need to identify a method for identifying a sample of ex-offenders who have been (or are about to be) rehabilitated. We might acquire these from records with the cooperation of the relevant authority. The sections of the questionnaire might then be arranged as follows:

(i) *About the respondent*
This section would include information on gender, age group, family details, who the respondent lives with, how long in prison and/or been rehabilitated,

rental/property arrangements, work and/or study pattern (including voluntary work), qualifications, ethnicity/religion. It may also include some questions on satisfaction with life/health and optimism about progressing in the new environment. Some open questions could be appropriate in this section.

(ii) *Relationship to local area*
This section might include information on the types of local activity alliances and organisational involvement the respondent takes part in. The information might be scaled on level of importance. It may also include questions on the respondent's satisfaction with the local neighbourhood and whether or not the respondent feels empowered to act and influence decisions and what happens in the future.

(iii) *Social networks and types of support*
This section of the questionnaire may ask about who the respondent would turn to for friendship, advisory and emotional support, including practical help and employment prospects. The degree of closeness, mode of communication (e.g. face to face or phone) as well as type of relationship can be included and coded as an attribute of the network data. Information might also be gathered on the ties between alters by asking the respondent if any of the people he/she has mentioned know each other. A question may also be included on who the respondent knows who is good at organising things locally and who has assisted in the formation of valued ties for the respondent. Answers might be broad and include organisations such as a job centre or individuals such as a family member, a local councillor or a business person.

A showcard including details of, for example, local groups and activities, might be used to carry out the survey. This would assist the interviewer who shows the card with a list of numbered organisations to the respondent. This leads to a more rapid and accurate completion of the questionnaire provided that the showcard does not miss out on any important local groups and activities.

The data from the survey might then be coded for input to a social network package, such as UCINET (via Excel as shown in Table 1) for social network analysis and subsequently exported to Netdraw or as a VNA file to Gephi for visualisation purposes. The survey would assist in identifying the structure of ex-offender social and support networks and their relationship to employment prospects.

7. Conclusion

We have focussed on areas associated with health, wellbeing and employment in describing some methods used in social network analysis. These, and other, methods can also be applied in other areas such as business or innovation networks and econometrics. There may also be an overlap between methods as when a data set, though not complete, can be treated as being so for the purpose of the study.

8. References

Borgatti, S.P., Everett, M.G. and Freeman, L.C. 2002. Ucinet for Windows: Software for Social Network Analysis. Harvard, MA: Analytic Technologies.

Heckathorn, D. 1997. "Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations." *Social Problems*. 49, 11-34.

INSNA http://www.insna.org/ Accessed Feb 2013.

McPherson, M., Smith-Lovin, L. and Cook, J., 2001. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology.* Vol. 27: 415-444.

Milgram, Stanley. "The Small World Problem". *Psychology Today*, 1(1), May 1967. pp 60 – 67

Miller, 2013. http://www.fort.usgs.gov/landsatsurvey/SnowballSampling.asp. Accessed February 2013.

Moreno, Jacob L. Application of the group method to classification, New York, 1932. National Committee on Prisons and Prison Labor, New York, 1932.

Moreno, J. L. and H. H. Jennings, 1938. Statistics of social configurations. Sociometry, 1:342-374.

Pearson, M. and Michell, L., 2000 "Smoke Rings: Social Network Analysis of Friendship Groups, Smoking and Drug-Taking," *Drugs: Education, Prevention and Policy*, 7(1), 21-37.

Pearson, M. and West, P., 2003. 'Drifting Smoke Rings. Social Network Analysis and Markov Processes in a Longitudinal Study of Friendship Groups and Risk-Taking' *Connections* 25(2) 59-76.

SIENA http://www.stats.ox.ac.uk/~snijders/siena/ Accessed Feb 2013.

Snijders, T. 1992. Estimation on the basis of snowball samples: How to weight. Bulletin Methodologie Sociologique 36: 59-70.

Steglich, C., Snijders, T. and Pearson, M., 2010. 'Dynamic Networks and Behaviour: Separating Selection from Influence', *Sociological Methodology*, 40, 329-393.

VISONE http://visone.info/ Accessed Feb 2013