What Is Special About Social Network Analysis?

Marijtje A. J. van Duijn[1] and Jeroen K. Vermunt[2]

[1]University of Groningen, the Netherlands
[2]Tilburg University, the Netherlands

Abstract. In a short introduction on social network analysis, the main characteristics of social network data as well as the main goals of social network analysis are described. An overview of statistical models for social network data is given, pointing at differences and similarities between the various model classes and introducing the most recent developments in social network modeling.

Keywords: complete network data, exponential random graph models, personal network data, random effect models, social network analysis

Introduction

Social network analysis is an interdisciplinary field of research with a long history of input from sociology, anthropology, statistics, mathematics, information sciences, education, psychology, and other disciplines. It still is a very active research area, as is evident from the many recent publications on social network analysis.

The large interest in social networks can be understood in view of the important theoretical and intuitively appealing research questions connected with social networks and the challenging methodological problems associated with the collection and analysis of social network data. This fruitful combination of content and methodology has stimulated lots of research in the past, described, for example, by Wasserman and Faust (1994, chap. 1). Both aspects of social network analysis involve theoretical as well as empirical problems, which makes the challenge even greater and the research more rewarding.

Wasserman, Scott, and Carrington (2005) note the rapid recent developments in the analysis of social network data but do not offer an explanation. We think that the increase in computer and computing facilities is an important factor. We expect the advances in social network analysis to continue and present examples of some of the most recent developments in this special issue of *Methodology*.

Social Network Analysis

Social network analysis aims at understanding the network structure by description, visualization, and (statistical) modeling. Social network data consist of various elements. Following the definition by Wasserman and Faust (1994, p. 89), social network data can be viewed as a social relational system characterized by a set of actors and their social ties. Additional information in the form of actor attribute variables or multiple relations can be part of the social relational system.

Several types of social network data are distinguished. The two main types, which are both present in this special issue, are *ego-centered* or *personal networks*, and *complete* or *one-mode networks*. Ego-centered network data are usually collected from a sample of actors (egos) reporting on the ties with and between other people (alters). The relational system is then assumed to be composed of the sampled egos and reported alters and their ties, as well as possible additional actor and tie information. Complete network data, on the other hand, concern a well-defined group of actors who report on their ties with all other actors in the group. The ties reported by actors can usually not be assumed to be independent, which makes personal and complete network sampling nonstandard. Rather than discussing network sampling in more detail here, we refer to Wasserman and Faust (1994) and Frank (2005) and the references therein.

Social network data can be collected in various ways. The most common approach is by means of questionnaires, but interviews, observations, and secondary sources are also frequently used network data collection methods (see also Marsden, 2005). In research utilizing egocentered network data, it is important to obtain as complete a picture of the respondents' networks as possible, which requires special tools for helping respondents to delineate their networks. A commonly used tool for this purpose is the so-called name generator, which provides a clear definition of which persons known by ego qualify as a network member (or alter) of ego. In a recent study, Van der Gaag and Snijders (2005) developed a new instrument for the measurement of ego's social capital that was named the *resource generator*. In this special issue, Gerich and Lehner (2006) show how computer-assisted self-administered interviews (CASI) with name generators can provide an improvement in personal network data collection.

Tie variables are often, though not necessarily, dichotomous, indicating the presence or absence of a relationship. This facilitates a nice depiction of the network in a graph or sociogram. The accompanying mathematical represen-

DOI 10.1027/1614-1881.2.1.2

tation is an adjacency matrix with 0s and 1s, where the diagonal is usually not defined (actors do not indicate ties with themselves). If the graph is undirected (for instance when relations between actors are observed instead of self-reported) the adjacency matrix is symmetric.

From mathematical graph theory a variety of concepts describing the properties of the network are available, such as reciprocity (when two actors indicate the existence of a tie between them), stars (when one central actor is connected to a number of other, unconnected actors—personal network data can be viewed as a collection of stars), and cliques (when there is a group of at least three actors that are all connected to each other). Many more concepts are available, and the reader is referred to Wasserman and Faust (1994) for a complete overview.

In the analysis of complete networks, a distinction can be made between (a) descriptive methods, also through graphical representations (see Freeman, 2005); (b) analysis procedures, often based on a decomposition of the adjacency matrix; and (c) statistical models based on probability distributions. Visualization by displaying a sociogram as well as a summary of graph theoretical concepts provides a first description of social network data. For a small graph this may suffice, but usually the data and/or research questions are too complex for this relatively simple approach.

Often it is of interest to compare actors on the basis of their tie variables, possibly also taking into account other actor characteristics. The identification of subgroups is an important area in social network analysis, for which visualization tools can be very helpful. In this special issue, Brandes, Kenis, and Raab (2006) make a strong case for the explanatory power of network visualization, based on complex mathematical algorithms implemented in specialized network visualization software called visone (Brandes & Wagner, 2003).

Another important goal of social network analysis is the modeling of ties between actors, so as to explain and/or predict the observed network. Several modeling approaches exist that are discussed in more detail in the next section. As is shown by four articles in this special issue, the modeling becomes more complex for richer social network data—that is, when actor attributes, multiple networks, and/or multiple observations of the same network are available.

The availability of software is an important condition for the feasibility of social network analysis, especially for applied researchers. Some specialized software has been available for a long time. As noted earlier, the development of new or improved software for social network analysis seems continuous. A recent overview of the state of the art in software for social network analysis—focusing on analysis through computation rather than visualization—is provided by Huisman and Van Duijn (2005). Important free software packages include Pajek (Batagelj & Mrvar, 2004; see also De Nooy, Mrvar, & Batagelj, 2005) and StOCNET (Boer, Huisman, Snijders, & Zeggelink, 2003). The Web site of the International Network on Social Network Analysis (INSNA) contains a page with short descriptions of and links to a broad range of available software for social network analysis (see INSNA, 2004).

Statistical Models for Social Network Analysis

The statistical models applied in social network analysis are typically nonstandard because the common assumption of independent observations does not hold: The multiple ties to and from the same actor are related. Moreover, the popular assumption of continuous normally distributed variables does not hold when tie variables are binary, nominal, ordinal, or count variables. Below, three sometimes interrelated classes of statistical models for complete network data are discussed: (a) dyadic interaction models, (b) exponential random graph models, and (c) stochastic actor-oriented models. But first we introduce several statistical models for the analysis of ego-centered network data.

The dependence structure of personal network data is least complex if there is no information on ties between alters, and if at the same time egos have neither overlapping networks nor are members of other egos' networks. Then, the data have a neat hierarchical structure, where alters are ordered (or nested) in egos. If the tie variable can be treated as continuous, standard linear two-level models (also known as hierarchical, random effects, or mixed effects models) for personal network data can be defined in a straightforward way (Van Duijn, Van Busschbach, & Snijders, 1999), distinguishing an ego-specific error (or random) term in addition to the usual (dyadic) error term. For binary data, a multilevel logistic model can be used, as was done by Wellman and Frank (2001). The advantages of the multilevel modeling framework are that it is extremely flexible and that it does not require balanced data, that is, the same number of alters for each ego.

In this special issue, Vermunt and Kalmijn (2006) propose a related conditional logit modeling approach for the modeling of categorical (nominal) personal network data in the presence of categorical covariates. It deals with the dependence between alters within egos using either a parametric or nonparametric random effects approach. The parametric formulation leads to a model with a high-dimensional random structure, which Vermunt and Kalmijn (2006) restrict by superimposing a factor-analytic model. In the nonparametric formulation, egos are assigned to latent classes.

Models for complete directed network data can be obtained by extending the multilevel approach. In that case, two observations instead of one are available for each pair of actors, the dyad, and we no longer have a convenient hierarchical structure but a cross-nested structure where the two dyadic observations are nested in two actors. Snijders and Kenny (1999) formulated the social relations model (SRM) as a random effects model for continuous directed tie variables for possibly incomplete social networks. The multilevel SRM contains, apart from an intercept (called *density* in social network analysis) and possibly fixed covariate effects, random sender and receiver effects. The random effects represent the two different roles actors have in dyadic relations and are assumed to be actor-wise correlated, taking into account the interdependence of these roles through the dependence between the relation to and from the same actor (*reciprocity* in social network analysis terminology). Because of the random effects formulation, extensions to the analysis of multiple observations of complete networks, possibly with specific configurations (teams, families, or couples), are straightforward. The special case of the SRM for networks consisting of only two persons, say for couples, is known as the actor-partner interdependence model (APIM; Kashy & Kenny, 2000).

For binary tie variables, the equivalent of the SRM is known as the p_2 model (Van Duijn, Snijders, & Zijlstra, 2004), which is a random effect variant of the p_1 model of Holland and Leinhardt (1981). The p_1 model—defining the first probability distribution for binary dyadic data—is a model for the four possible dyadic outcomes, one mutual, one null, and two asymmetric. It distinguishes density, sender, receiver, and reciprocity effects, but unlike the SRM, all effects are modeled as fixed effects. Holland and Leinhardt (1981) showed that the p_1 model is identified with suitable restrictions and that it implies conditionally independence between dyads given the sender and receiver parameters. As in the SRM, in the p_2 model, the sender and receiver parameters are modeled as correlated random effects, a formulation that makes it possible to include actor- and dyad-specific covariates as fixed sender, receiver, density, or reciprocity regression parameters.

In the current special issue, Zijlstra, Van Duijn, & Snijders (2006) propose a multilevel extension of the p_2 model for the analysis of multiple social networks, which may have different sizes. Instead of estimating a separate p_2 model for each network, a single identical model is defined for all networks, whose parameters are allowed to vary across networks by introducing random effects at the network level. Other interesting extensions of the SRM and p_1 model have been proposed (see Wasserman & Faust, 1994, sec. 15.5), several of which are developed in combination with a Bayesian estimation framework (see, e.g., Hoff, 2005).

Next to the Bernoulli distribution of independent ties, the p_1 model can be regarded as one of the simplest forms of a family of exponential random graph distributions (see Wasserman & Faust, 1994, p. 528). Many of the extensions of the p_1 model belong to the family of exponential random graph distributions. A very important category is formed by the Markov random graph (Frank & Strauss, 1986) or p^* model (Wasserman & Pattison, 1996; see also Wasserman & Robins, 2005). This model defines the probability of a complete—directed or undirected—social network as an exponential family parameterized with sufficient statistics based on the network, possibly conditional on covariates. The dependence among network ties is implied by the choice of the sufficient network statistics. For instance, the p_1 model is defined by an exponential random graph distribution with the indegrees, outdegrees, and number of mutual dyads as sufficient statistics (Wasserman & Faust, 1994, p. 633). For the Markov random graph or p^* model, which assumes dependence between ties involving the same actor, the number of possible statistics and accompanying number of parameters to be estimated is enormous. This number can, however, be restricted by using suitable homogeneity constraints.

It has turned out to be extremely difficult to get reliable estimates for parameters of the p^* models. The early pseudolikelihood estimation approach proposed by Wasserman and Pattison (1996)—following Strauss and Ikeda (1990)—based on a conditional logistic model formulation has been shown to yield incorrect estimates and underestimated standard errors because of disregarding the dependence between dyads. As shown by Snijders (2002), an alternative estimation method in the form of Markov chain Monte Carlo maximum likelihood estimation may also cause problems because certain, degenerate, model configurations cannot be sampled from adequately. Recently, advances have been made in the understanding of these estimation difficulties, which has resulted in improved model definitions and estimation methods (Snijders, Pattison, Robins, & Handcock, in press), and more developments in exponential random graph modeling and estimation are to be expected.

The latent space model used by Shortreed, Handcock, and Hoff (2006)—which also belongs to the exponential random graph family—handles the dependence between ties to and from the same actor in a different way. More specifically, actors are assumed to have latent positions, conditional on which ties can be assumed to be mutually independent, leading to a logistic regression model for the tie probabilities. Instead of using fixed or random sender and receiver effects as in the p_1 and p_2 models, respectively, the (Euclidian) distances between actors (possibly conditional on actor attributes) are the model parameters to be estimated. The main difficulty of the latent space model is not the estimation of the distances between actors or the regression parameters corresponding to covariates, which can be obtained with Markov chain Monte Carlo algorithm, but the derivation of the latent positions of the actors from the distances. Shortreed et al. (2006) propose using the posterior mean of the Kullback-Leibler divergence instead of more obvious choices such as the posterior mean or posterior mode. An attractive feature of the latent space model is that, in addition to estimates for actor attribute effects, it provides a model-based visualization of the network under study.

The final class of models discussed here are models for longitudinal network data, that is, models for the evolution of social networks over time. These models require multiple observations of a complete social network, where actors may join and/or leave the network. An important model for longitudinal network data—the stochastic actor-oriented model—was proposed by Snijders (1996); see also Huisman and Snijders (2003) and Snijders (2005). This is a continuous-time Markov model for network "events" of single actors who may change their ties with other actors at a certain rate, which induces changes in the overall network structure. A multinomial logit model expressing the preferences of the actors determines the probability corresponding to a tie change with one of the other actors. Note that the model is estimated using discrete-time observations of the network. An extension of this model for the analysis of the joint development of social networks and behavior, acknowledging the simultaneous influence of individual behavior on the network structure and of network structure on individual behavior, was recently proposed by Snijders, Steglich, and Schweinberger (in press). In this special issue, Steglich, Snijders, and West (2006) provide a very nice, completely formula-free introduction to the main model ingredients and assumptions of this complex model using an application to teenager friendship networks and social behavior defined by alcohol use and music taste. This illustration and the substantive applications described in the other articles of this special issue provide excellent examples of the usefulness of social network methodology for the advancement of social and behavioral sciences.

The powerful combination of methods and applications makes social network analysis special. This is not to say that there are no unresolved issues left. Apart from the estimation difficulties outlined earlier, one important issue is the assessment of goodness-of-fit of statistical models for social networks. This entails both the issue of model selection and the question of prediction accuracy of the selected model. The complexity of the models used for analyzing social network data makes both issues difficult, as recently noted by Hoff (2005, p. 295) and Zijlstra, Van Duijn, and Snijders (2005).

We expect that social network analysis will continue to encourage researchers from different fields to explore its special methodological and empirical challenges.

References

Batagelj, V., & Mrvar, A. (2004). Pajek: Package for Large Networks (Version 1.00) [Computer software]. Ljubljana, Slovenia: University of Ljubljana. Retrieved August 30, 2004, from http://vlado.fmf.uni-lj.si/pub/networks/pajek/.

Boer, P., Huisman, M., Snijders, T. A. B., & Zeggelink, E. P. H. (2003). StOCNET: An Open Software System for the Advanced Analysis of Social Networks (Version 1.4) [Computer software]. Groningen, the Netherlands: ProGAMMA/ICS. Retrieved January 20, 2004, from http://stat.gamma.rug.nl/stocnet/

Brandes, U., Kenis, P., & Raab, J. (2006). Explanation through network visualization. *Methodology*, *2*, 16–23.

Brandes, U., & Wagner, D. (2003). Visone—Analysis and visualization of social networks. In M. Jünger & P. Mutzel (Eds.), *Graph drawing software* (pp. 321–340). Berlin: Springer.

De Nooy, W., Mrvar, A., & Batagelj, V. (2005). *Exploratory social network analysis with Pajek*. New York: Cambridge University Press.

Frank, O. (2005). Network sampling and model fitting. In P. J. Carrington, J. Scott, & S. Wasserman (Eds.), *Models and methods in social network analysis* (pp. 31–56) New York: Cambridge University Press.

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

Freeman, L. C. (2005). Graphic techniques for exploring social network data. In P. J. Carrington, J. Scott, & S. Wasserman (Eds.), *Models and methods in social network analysis* (pp. 248–269) New York: Cambridge University Press.

Gerich, J., & Lehner, R. (2006). Collection of ego-centered network data with computer-assisted interviews. *Methodology*, *2*, 7–15.

Hoff, P. D. (2005). Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical Association*, *100*, 286–295.

Holland, P. W., & Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. (With discussion.) *Journal of the American Statistical Association*, *76*, 33–65.

Huisman, M., & Snijders, T. A. B. (2003). Statistical analysis of longitudinal network data with changing composition. *Sociological Methods and Research*, *32*, 253–287.

Huisman, M., & Van Duijn, M. A. J. (2005). Software for social network analysis. In P. J. Carrington, J. Scott, & S. Wasserman (Eds.), *Models and methods in social network analysis* (pp. 270–316). New York: Cambridge University Press.

International Network on Social Network Analysis (INSNA) (2004). Computer programs for social network analysis. Retrieved August 30, 2004, from http://www.insna.org/INSNA/soft_inf.html

Kashy, D. A., & Kenny, D. A. (2000). The analysis of data from dyads and groups. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social psychology* (pp. 451–477). New York: Cambridge University Press.

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). New York: Cambridge University Press.

Shortreed, S., Handcock, M. S., & Hoff, P. (2006). Positional estimation within a latent space model for networks. *Methodology*, *2*, 24–33.

Snijders, T. A. B. (1996). Stochastic actor-oriented models for network change. *Journal of Mathematical Sociology*, *21*, 149–172.

Snijders, T. A. B. (2002, April 19). Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*, *3*, 2. Retrieved September 26, 2005, from http://www.cmu.edu/joss/content/articles/volume3/Snijders.pdf

Snijders, T. A. B. (2005). Models for longitudinal network data. In P. J. Carrington, J. Scott, & S. Wasserman (Eds.), *Models and methods in social network analysis* (pp. 215–247). New York: Cambridge University Press.

Snijders, T. A. B., & Kenny, D. (1999). The social relations models for family data: A multilevel approach. *Personal Relationships*, *6*, 471–486.

Snijders, T. A. B., Pattison, P. E., Robins, G. L., & Handcock, M. S. (in press). New specifications for exponential random graph models. *Sociological Methodology*.

Snijders, T. A. B., Steglich, C., & Schweinberger, M. (in press). Modeling the co-evolution of networks and behavior. In K. Van Montfort, H. Oud, & A. Satorra (Eds.), *Longitudinal models in the behavioral and related sciences*. Mahwah, NJ: Lawrence Erlbaum.

Steglich, C., Snijders, T. A. B., & West, P. (2006). Applying SIENA: An illustrative analysis of the coevolution of adolescents' friendship networks, taste in music, and alcohol consumption. *Methodology*, *2*, 48–56.

Strauss, D., & Ikeda, M. (1990). Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, *85*, 204–212.

Van der Gaag, M. P. J., & Snijders, T. A. B. (2005). The resource generator: Measurement of individual social capital with concrete items. *Social Networks*, *27*, 1–29.

Van Duijn, M. A. J., Snijders, T. A. B., & Zijlstra, B. J. H. (2004). p_2: A random effects model with covariates for directed graphs. *Statistica Neerlandica*, *58*, 234–254.

Van Duijn, M. A. J., Van Busschbach, J. T., & Snijders, T. A. B.

(1999). Multilevel analysis of personal networks as dependent variables. *Social Networks*, *21*, 187–209.

Vermunt, J. K., & Kalmijn, M. (2006). Random-effects models for personal networks: An application to marital status homogeneity. *Methodology*, *2*, 34–41.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications.* New York: Cambridge University Press.

Wasserman, S., & Pattison, P. (1996). Logit models and logistic regressions for social networks: I. An introduction to Markov random 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. 148–161). New York: Cambridge University Press.

Wasserman, S., Scott, J., & Carrington, P. J. (2005). Introduction. In P. J. Carrington, J. Scott, & S. Wasserman (Eds.), *Models and methods in social network analysis* (pp. 1–8). New York: Cambridge University Press.

Wellman, B. A., & Frank, K. A. (2001). Network capital in a multi-level world: Getting support from personal communities. In N. Lin, R. Burt, & K. Cook (Eds.), *Social capital: Theory and research* (pp. 233–274) Chicago: Aldine De Gruyter.

Zijlstra, B. J. H., Van Duijn, M. A. J., & Snijders, T. A. B. (2005). Model selection in random effects for directed graphs using approximated Bayes factors. *Statistica Neerlandica*, *59*, 107–118.

Zijlstra, B. J. H., Van Duijn, M. A. J., & Snijders, T. A. B. (2006). The multilevel p_2 model: A random effect model for the analysis of multiple social networks. *Methodology*, *2*, 42–47.

Marijtje A. J. vanDuijn

University of Groningen
Faculty of Behavioral and Social Sciences
Department of Sociology
Grote Rozenstraat 31
NL-9712 TG Groningen
The Netherlands
Tel. +31 50 363 6195
Fax +31 50 363 6226
E-mail m.a.j.van.duijn@rug.nl