

A Survey of Methods for the Statistical Analysis of Network Data*

Ian M. Schmutte John M. Abowd

11/14/2006

Very Preliminary, Please do not cite

Abstract

This paper reviews several statistical models for network data with the aim of analyzing the relational structure of matched employer-employee data.

1 Introduction

Few economic relationships are as well documented as the exchange of labor. With the advent of matched employer-employee data, the identities of both parties in millions of labor contracts are known as well as important details of the participants and their relationship. This data has become available at the same time that interest and opportunity to study large-scale networks of social and economic interactions has become possible. The social relations between workers and firms at any point in time can be represented as a bipartite graph, with separate nodes in the network standing for either firms or workers [1]. The links between these nodes represent the existence of an employment relationship. Worker i is connected to firm j just when i is employed by j . In modern graph theory, all of the relevant attributes of the network can be summarized in the graph. Characteristics of the worker are attached to the workers node. These are the things that the worker carries with him from job to job, including education, skills, demographic characteristics, etc. Characteristics of the relationship between i and j , such as the wage, benefits, tenure, match-specific capital, attach to the edge linking the two.

From the perspective of labor economics, we are interested in modelling the exchange of labor services for policy purposes, and in order to understand the extent to which labor markets allocate labor efficiently. Thus, it is important to understand how labor market relationships form, how they end, and how the nature of the employment relationship is determined. More specifically,

*This preliminary and incomplete literature overview is only intended to be a discussion paper. Please contact the authors (ims28@cornell.edu) for further information about the research program for which this document was created.

there is reason to suspect that whether an employment relationship forms is not independent of the formation of other labor market relationships, and neither are the outcomes of those relationships. In sociological terms, there are structural features of labor markets that constrain the flow of labor. Workers might be constrained in the labor market because they are trapped in a segment of the labor market with low returns to human capital. In terms more common in economics, there may be peer effects that determine where people get to work, and what kinds of wages and tenure they earn.

Matched employer-employee data offer unprecedented information on such questions. Indeed, the hope is that these data are sufficiently rich to identify social interactions in labor market outcomes. However, the statistical and computational tools that exist for incorporating relational information about which employees work for particular employers with more conventional data on labor market outcomes are in their infancy. To develop statistical models and methods for inference that exhaust the relational information inherent in matched employer-employee data, we believe there is value in the methods developed by statistical sociologists and computer scientists for modeling network data.

This paper reviews a set of models that were developed by sociologists and statisticians for analyzing social network dynamics. There are several intersecting streams of literature that model random networks as generally as possible. The central model in this literature is the p^* logit model developed by Wasserman and Pattison [8] (1996), also known as the Exponential Random Graph Model (ERGM). The p^* model is a generalization of several earlier models; in particular the p_1 model developed by Holland and Leinhardt [5] described by Fienberg, Meyer and Wasserman [2] (1985) and the Markov Random Graph model devised by Frank and Strauss [3] (1986). Although the p^* is essentially a model for static social networks, there has also been a recent stream of literature extending the p^* to an agent-based model of network dynamics in a way that is similar to the classic RATE models of statistical sociology [4]. Snijders (2001) [7] is the essential reference, though Koskinen (2004) [6] claims to have generalized Snijders' model in a fully Bayesian framework that allows representation of a wider class of dynamic graph models, including, importantly, undirected bipartite graphs, and for which there are superior inferential methods.

2 Early Models: p_1 and Markov Random Graphs

2.1 p_1

2.1.1 The model

The data (adjacency) matrix, x , is a realization of a random matrix X , in which each dyad, $D_{ij} = (X_{ij}, X_{ji})$ is an independent bivariate random variable with

possible values

$$\begin{aligned} D_{ij} &= (1, 1) \\ &= (1, 0) \text{ or } (0, 1) \\ &= (0, 0) \end{aligned}$$

The object is to model the probability, $\Pr(D_{ij} = (k, h))$. This is a generalization of a Poisson random graph model in which each X_{ij} is taken to be an independent Bernoulli random variable. The extension to multiple relations simply replaces the dyad form above, in which the dyad outcome is a 2×1 vector with a $2 \times R$ dyad outcome.

Take the number of relations, $R = 1$. An intuitive way to model each dyad is to define a random variable Y_{ijkh} such that

$$Y_{ijkh} = 1 \text{ when } D_{ij} = (k, h)$$

This allows us to capture a very general model in which the probability of each dyad state depends on the actors' identities but also incorporates the dyad state itself.

Let $\pi_{ijkh} = \Pr(D_{ij} = (k, h)) = \Pr(Y_{ijkh} = 1)$. The Holland-Leinhardt p_1 -class specifies a log-linear model in which the probability of each dyad state depends on actor identities and the possibility of reciprocity.

$$\begin{aligned} \log \Pr(D_{ij} = (k, h)) &= \log \Pr(Y_{ijkh} = 1) \\ &= \lambda_{ij} + k(\alpha_i + \beta_j + \theta) + h(\alpha_i + \beta_j + \theta) + kh\rho_{ij} \end{aligned}$$

They impose the additional identifying restrictions

$$\begin{aligned} \rho_{ij} &= \rho \\ \sum_{i=1}^g \alpha_i &= \sum_{i=1}^g \beta_i = 0 \end{aligned}$$

2.1.2 Estimation

The p_1 model as specified above is fully identified. Estimation of p_1 is just the task of estimating a log-linear model for Y_{ijkh} with the particular set of two-way interactions. This is possible through an Iterative Proportional Fitting procedure on the $g \times g \times 2 \times 2$ matrix $Y = (Y_{ijkh})$.

2.2 Markov Random Graph Models

2.2.1 The Model

2.2.2 Estimation

3 p^*

3.0.3 The Model

The p^* model is a generalization of the Markov random graph model, and is able to account for completely general specifications of dependencies between features of the graph. Anderson, Wasserman and Crouch note that the p^* model can be derived either by specifying an autologistic regression model that keeps track of the dependency structure, or alternatively, as a generalization of the theory of Markov random fields via a result called the Hammersley-Clifford theorem. The presentation here is based on the autologistic regression approach and borrows from the presentation in Anderson, Wasserman and Crouch (1998). Again, g is the number of actors in the data, and $N = \{1, \dots, g\}$. We let \mathbf{x} be the observed data, which is a $g \times g$ adjacency matrix.

The function $z(x) = (z_1(x), z_2(x), \dots, z_r(x))$ is a map from the space of all possible $g \times g$ adjacency matrices to \mathbb{R}^r . These can be any functions of the data. Table 4 in Anderson, Wasserman and Crouch suggests some of the possible functions of the data that can be incorporated into $z(x)$. We assume that the probability over graphs is log-linear in the components of z .

$$\Pr(X = x) = \frac{\exp[\theta' z(x)]}{\kappa(\theta)}$$

where $\theta \in \mathbb{R}^r$ is a parameter vector to be estimated.

On the basis of this general model, we proceed to formulate a logit model for the individual links. Unlike the standard iid case for these models, it is necessary to work with the conditional distribution of each link. To facilitate this, some special notation is required. Let X_{ij}^c refer to the set of all random variables describing each edge, subtracting out the variable X_{ij} . This is the conditioning set for link X_{ij} . Let x_{ij}^+ be the matrix identical to x , but with $x_{ij} = 1$. Define x_{ij}^- as the data matrix identical to x but with $x_{ij} = 0$. Based on this, specify the conditional odds ratio:

$$\begin{aligned} \exp\{\omega_{ij}\} &= \frac{\Pr(X_{ij} = 1 | X_{ij}^c)}{\Pr(X_{ij} = 0 | X_{ij}^c)} \\ &= \frac{\Pr(X = x_{ij}^+)}{\Pr(X = x_{ij}^-)} \\ &= \frac{\exp[\theta' z(x_{ij}^+)]}{\exp[\theta' z(x_{ij}^-)]} = \exp(\theta'(z(x_{ij}^+) - z(x_{ij}^-))) \equiv \exp \theta' d_{ij} \end{aligned}$$

where $d_{ij} = (z(X_{ij}^+) - z(X_{ij}^-))$. So we end up with the system of equations

$$\omega_{ij} = \theta' d_{ij}$$

3.0.4 Estimation

The claim is that this model can, in general, be estimated using standard logit techniques. The catch is, apparently, that you have to assume that the logits, ω_{ij} , are mutually independent so that a pseudo-likelihood approach can be used.

4 Longitudinal Models

4.1 Snijders (2001) Model for Longitudinal Social Network Data

Snijders (2001) considers the dynamic evolution of social interactions between a set of n actors. He models the directed social network at any point in time, t , as a $n \times n$ adjacency matrix, $x(t)$. Letting \aleph be the space of all such matrices, the network evolution is modeled as a continuous time Markov chain whose states are the adjacency matrices in the set \aleph . Transition probabilities are taken to depend on actor attributes, edge attributes, and various network statistics that are taken to matter because they determine how actors in the model behave to alter the network by changing their outgoing links. Estimation of the parameters in the transition probabilities is based on matching simulations of the model to the observed data.

4.1.1 The Model

Each actor in the model has an objective function, $f_i(\beta, x)$ which represents preferences over the various network configurations, $x \in \aleph$, which vary with the parameter β . At any point in time, at most one actor is selected to make a change to his outgoing links. We will adopt the following notation: let $x_{ij}^c(t)$ be the adjacency matrix that is identical to $x(t)$ except that the link from i to j has changed. Therefore, $x_{ij}(t) = 1$ if and only if $x_{ij}^c(t) = 0$.

The agent's objective function can depend upon his own characteristics, the characteristics of other agents, and on characteristics of their relationship. It can also depend on arbitrary characteristics of the complete network. Thus, the calculation by the agent of the benefit of extending or retracting any particular link can be based on any features of the network one cares to specify, as in the p^* model.

The various configurations of social network can be thought of as states in a stochastic process, $X(t)$ that evolves continuously over time. Transition probabilities in the continuous time Markov process are given by the "transition intensity", which is defined as

$$q_{ij}(X) = \lim_{dt \downarrow 0} \frac{1}{dt} \Pr\{X(t+dt) = x_{ij}^c | X(t) = x\}$$

This is the probability of changing from state x to state x_{ij}^c at any particular moment in time. The matrix $[q_{ij}(x)]$ is the continuous time analogue to the usual transition matrix. In Snijders setup, the transition intensity is

$$q_{ij}(x) = \lambda_i(x)p_{ij}(x)$$

where $\lambda_i(x)$ is the waiting time between each change made by i , and $p_{ij}(x)$ determines whether i changes his link to j conditional on the fact that he actually makes a change at t .

$\lambda_i(x)$ can either be specified as coming from a particular distribution, or modelled as a function of node and edge covariates. In terms of the modelling, all that matters is that the distribution of waiting times is such that the probability that any two agents make a change at the same time is zero. The idea is that the model evolves through "ministeps", of which we observe only a few. This assumption is more crucial in the limited panel data available in most social network research, where only a few timepoints are observed.

Snijders (2001) assumes that the rate function is identical for all agents, given by λ_k , which means that for any time point, $t \in (t_k, t_{k+1})$ the waiting time until the next change made by any actor has negative exponential distribution with parameter $N\lambda_k$. When an event occurs, the probability that it is made by any particular actor is $1/N$. That agent chooses whether to alter its links according to the objective function:

$$f_i(x) = \sum_l \beta_l s_{il}(x)$$

Here, $s_{il}(x)$ are arbitrary statistics of the network that can either capture individual attributes, match attributes, or more general aspects of the network's current configuration. Given that i makes a change, he chooses to change the link to j that maximizes

$$f_i(x_{ij}^c) + U_i(j, t, x)$$

$U_i()$ is unobserved heterogeneity that is assumed to be independent of the deterministic term, f_i , and that depends on j, x , and t . If U_i has the Gumbel distribution, then the probability that i changes his link to j is given by the multinomial logit distribution

$$p_{ij}(x) = \frac{\exp(f_i(x_{ij}^c))}{\sum_{h \neq i} \exp(f_i(x_{ih}^c))}$$

4.1.2 Model with application to the labor market

Imagine a set of I individuals, $A(t)$, and a set of J employers, $F(t)$ arranged in a bipartite graph. There is a link between $i \in A(t)$ and $j \in F(t)$ if and only if i is employed by j at date t . The totality of these links can be represented by the $I \times J$ adjacency matrix $B(t)$. This graph is changing over time, so it makes sense to refer to $B(t)$ as the adjacency matrix of the bipartite graph representing the individual-employer matches at time t . Since the

employment relations between firms and workers change at any time, it is reasonable to think of t as a continuous variable. Furthermore, we will distinguish primary employment from other forms of employment. This assumption puts constraints on the row degree distribution in $B(t)$. Specifically, assume that $j = 0$ refers to the non-employment state. Including the column $j = 0$ ensures that every individual in the population at date t has exactly one “employer.” Hence, $B(t) e_J = e_J$, where e_J is the $J \times 1$ column vector of 1s. Given this setup the column degree distribution, $e'_J B(t)$, is what is known in labor economics as the size distribution of employers (technically only the columns 1 to J are included in this distribution). We note that the (very hard) problem of entry and exit of employers can be included in this formalism by including columns in F for potential and defunct employers. For the moment, we are not going to worry about this complication.

The existing data are snapshots of the labor market at points in time, $B(t_1), \dots, B(t_T)$, where T is the total number of available time periods. These adjacency matrices describe outcomes sampled at discrete points in time from the $I \times (J + 1)$ potential outcomes at each moment of time. The objective is to use these snapshots of the labor market to test various assumptions about how the labor market evolves over time. The basic setup is as follows.

At any point in time, an individual has the objective function, $u_i(B(t))$ that assigns a value to every possible network configuration. At every feasible point in time, the individual can opt to alter his link in the graph by changing employers. Let b_i be the row vector of the adjacency matrix corresponding to the i^{th} worker. We assumed that a worker has at most one employer, so $\sum_j b_{ij} = 1$, since non-employment is included in the columns. The notation $B(i \rightarrow j)$ means that i changes his link with j . If j is i 's current employer, and i takes a job with j' , we denote this by $B(i \rightarrow j')$. If i leaves his job with j without taking a new job, we denote this by $B(i \rightarrow 0)$. If i leaves his current employer to non-employment, we denote this by $B(i \rightarrow j)$. [This notation, which was introduced by Snijders, doesn't seem to be helpful in the bipartite case. It is meant to represent the assumption that at any instant at most one link in the graph can change. The resulting change dissolves one employment relation and initiates another, including in either case the possibility of non-employment.]

The objective function for the individual-employer graph can be represented by a match function that depends upon the graph $B(t)$, characteristics of the individual $X(t)$, an $I \times k$ matrix, characteristics of the employers $Z(t)$, a $((J + 1) \times q)$ matrix, and characteristics of the match $W(t)$, a $(I(J + 1) \times p)$ matrix. Note that while B , X and Z are observable, at least in principle, W is (mostly) latent since it contains data for all potential matches. Express the match function at a point in time as $F(B(t), X(t), Z(t), W(t))$, an $I \times (J + 1)$ matrix function with elements f_{ij} .

The match function can, in principle, incorporate many features of the graph including any statistics about the graph structure that are relevant, covariates associated with the nodes, $X(t)$ and $Z(t)$, and covariates associated with the edges, $W(t)$. Theories of labor market equilibrium are descriptions of special-

izations of the match function f .

For modeling purposes, we now adopt Snijders' formalism of assuming that the state of the labor market adjacency matrix can only change a single match at a time. The configurations of employer-employee matches can be thought of as states in a stochastic process that evolves continuously over time. The states are characterized by relevant adjacency matrices, $B(t)$. The "transition intensity" is defined as

$$q_{\ell m}(B) = \lim_{dt \downarrow 0} \frac{1}{dt} \Pr\{B(t+dt) = B(i \rightarrow j) | B(t) = B\}$$

where $\ell, m = 1, \dots, I(J+1)$. The intensity $Q(t)$ matrix is the continuous analogue of a transition probability matrix. In this application $Q(t)$ is $(I(J+1) \times I(J+1))$ and its dependence on $B(t)$ means that only certain rows and columns, which depend upon the current state of the labor market, have non-zero transition rates. Snijders simplification, adapted to the bipartite graph case, restricts the the transition intensity intensity as follows

$$q_{\ell m}(B, X, Z, W) = \lambda_{ij}(B, X, Z, W) p_{\ell m}(B, X, Z, W)$$

where $\lambda_{ij}(B, X, Z, W)$ is the waiting time between each change made by ij , and $p_{\ell m}(B, X, Z, W)$ determines whether the pair represented by row ℓ changes to the pair represented by column m .

The waiting time matrix $\lambda_{ij}(B, X, Z, W)$ can either be specified as coming from a particular distribution, or modelled as a function of node and edge covariates. In terms of the modelling, what matters is that the distribution of waiting times is such that the probability that any employer-employee pair make a change "two at a time." The idea is that the model evolves through "ministeps," of which we observe the cumulative effect after a single "period" of time has elapsed, from t_k to t_{k+1} . This assumption is crucial in the limited panel data available in most social network research, where only a relatively few timepoints are observed.

Snijders (2001) has the rate identical for all agents, which in our application translates to all pairs λ_{ij} , hence for any time point, $t \in (t_k, t_{k+1})$ the waiting time until the next change made by employer-employee pair has negative exponential distribution with parameter $I(J+1)\lambda_{ij}$. When an event occurs, the probability that it is made by any particular pair is $1/I(J+1)$. That pair chooses whether to alter its links according to the objective function:

$$f_{ij}(B, X, Z, W) = \sum_{r=1}^R \beta_k s_{ijk}(B(t), X(t), Z(t), W(t))$$

where $s_{ijk}(B(t), X(t), Z(t), W(t))$ are specific characteristics of the network. Given that ij make a change, the resulting link maximizes

$$f_{ij}(B, X, Z, W) + u_{ij}(i, j, t, B)$$

$u_{ij}()$ is some unobserved heterogeneity term that is independent of the deterministic term, f_{ij} , and that depends on i, j, t , and B . If u_{ij} has the Gumbel distribution, then the probability that ij changes links is given by the multinomial logit distribution

$$p_{ij}(x) = \frac{\exp(f_{ij}(B, X, Z, W))}{\sum_{\ell m} \exp(f_{\ell m}(B, X, Z, W))}$$

4.1.3 Estimation

The parameters of the model are $\theta' = (\beta', \lambda')$ where $\lambda' = (\lambda_{11}, \dots, \lambda_{I(J+1)})$. If $\beta' = (\beta_1, \dots, \beta_K)$, so θ is $I(J+1) + K \times 1$. The idea is to use a $I(J+1) + K$ -dimensional statistic M such that

$$E_{\theta} M = m$$

where m are the observed moments. Snijders suggests a stochastic iterative algorithm for estimation of θ . The basic iteration step is

$$\hat{\theta}_{N+1} = \hat{\theta}_N - a_N I(M_N - m(B))$$

where a_N is some sequence converging to zero, M_N is the network statistic generated by simulating the model. Since the λ_{ij} are changing over time, they are estimated as

$$E_{\theta} [M_{ij}(B(t_{k-1}), B(t_k)) | B(t_{k-1})] = m_{ij}(B(t_{k-1}), B(t_k))$$

Since the β do not change over time, the relevant moment equation is

$$\sum_{r=1}^R E_{\theta} [M_{ij}(B(t_{k-1}), B(t_k)) | B(t_{k-1})] = \sum_{r=1}^R m_{ij}(B(t_{k-1}), B(t_k))$$

For estimating such coordinates of M , it is necessary to simulate the model. So, taking the parameter estimate $\hat{\theta}$, for each $k = 2, \dots, K$, simulate the process starting with initial state $B(t_{k-1})$ and let time run from $k-1$ to k . The network that results is $B^{sim}(t_k)$. Then, define

$$M = \sum_{r=1}^R m_r(B(t_{k-1}), B^{sim}(t_k))$$

While the observed outcome is

$$m(B) = \sum_{r=r}^K m_r(B(t_{k-1}), B(t_k))$$

References

- [1] John M. Abowd, Robert H. Creecy, and Francis Kramarz, *Computing person and firm effects using linked longitudinal employer-employee data*, mimeo (2002), 1–17.
- [2] Stephen E. Fienberg, Michael M. Meyer, and Stanley S. Wasserman, *Statistical analysis of multiple sociometric relationships*, Journal of the American Statistical Association **80** (1985), no. 389, 51–67.
- [3] O. Frank and D. Strauss, *Markov graphs*, Journal of the American Statistical Association **81** (1986), 832–842.
- [4] Hannon and Tuma, *Social dynamics*, New York:Academic Press, New York, 1984.
- [5] Paul W. Holland and Samuel Leinhardt, *An exponential family of probability distributions for directed graphs*, Journal of the American Statistical Association **76** (1981), no. 373, 33–50.
- [6] Johan (2004). Koskinen, *Bayesian inference for longitudinal social networks*, Research Report 2004:4 Department of Statistics, Stockholm University (2004).
- [7] Tom A. B. Snijders, *The statistical evaluation of social network dynamics*, Sociological Methodology **30** (2001), 361–395.
- [8] S. Wasserman and P. Pattison, *Logit models and logistic regressions for social networks: An introduction to markov graphs and p^** , Psychometrika **61** (1996), 401–25.