Smoking Behavior and Friendship Formation: The Importance of Time Heterogeneity in Studying Social Network Dynamics

Joshua A. Lospinoso
Department of Statistics
University of Oxford
Network Science Center
United States Military Academy
Email: http://www.stats.ox.ac.uk/~lospinos

Danielle J. Satchell
School of Osteopathic Medicine
University of Medicine and Dentistry of New Jersey
Network Science Center
United States Military Academy
Email: djsatchell@gmail.com

***Abstract*—This study illustrates the importance of assessing and accounting for time heterogeneity in longitudinal social network analysis. We apply the time heterogeneity model selection procedure of [1] to a dataset collected on social tie formation for university freshman in the Netherlands by [2]. Within the context of analyzing selection effects for smoking homophily to understand the implications of tobacco policy at a university, we show that failing to account for time heterogeneity yields quite different results substantively from the model arrived at using [1]. While the results are limited by the small scope of the dataset, the paper motivates the testing of time heterogeneity within longitudinal studies of social network behavior and further study of tobacco policy within university settings.**

I. Introduction

The study of social relationships and health has interested researchers for some time. [3]–[5] study health outcomes as embedded in social support mechanisms and found a significant role for social relationships on mortality rates. In other contexts, researchers have found that social relationships can also propagate risky behavior. Selection effects–whereby people choose to become friends with similar individuals (see [6])–and influence effects–whereby people are influenced towards risky behavior by their social contacts (see [7]–[9])–could provide environments in which social relationships increase factors of mortality and morbidity. Researchers have studied smoking behaviors embedded in social networks in a number of ways, e.g. through analysis of social positions [10], traditional, longitudinal statistical models [11], and stochastic actor based models [9]. Specific policy implications of tobacco-related workplace health promotion has been studied extensively (e.g. [12]), and have important implications for health outcomes.

Due to an array of difficulties in statistical analysis of social network data, the use of traditional methods may lead to erroneous results. Chief among these difficulties is the independence of observations assumption maintained for many of these traditional methods. Using recently developed tools for the longitudinal study of actor behavior embedded in social networks (see [13], [14]), researchers have made strides towards overcoming many of these difficulties.

This study revisits a small dataset collected by Van de Bunt [2], [15] which can be used to explore the effects of designated smoking areas on relationship formation among university students, and illustrates why the use of tools to assess and account for time heterogeneity developed by [16] and applied by [1], [17] represent a crucial part of the modeling process.

With methods to better detect and account for time heterogeneity in parameterizations of coevolution models, researchers can be more confident in the results of inference drawn from survey data used to study selection and influence effects on smoking behavior. These inferences can be used to guide healthcare policy and prevention programs that target the sorts of risk factors observed in the data.

II. Data

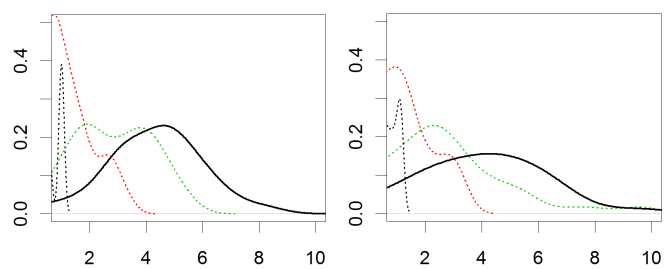
The data from this study comes from [2], [15], who collected data for university freshmen over seven time periods.[1] Of a body of 49 students, 17 either dropped out of the program or did not respond more than three times, yielding a dataset of 7 time periods for 32 students. There are four distinct programs which the students were assigned to: a two year, a three year, and a four year program. The sex and smoking behavior (yes or no) were also recorded at the onset of the study. The sociometric data was collected at each observation, and permitted the students to assign one of six categories listed in Table I to their relationship with each of the other students in the study. In the setting of the university, the smokers would have to separate themselves from the rest of the group to visit the smoking area. [2] remarks that they did so often. This smoking area was designated far away from where students and faculty would gather for coffee breaks between lectures. Students of the three programs had overlapping curricula for

This research was funded by U.S. Army Project Number 611102B74F and MIPR Number 9FDATXR048.

[1]The first four time points are three weeks apart, and the last three time points are six weeks apart

Label	Description
Best friendship	Persons whom you would call your real friends.
Friendship	Persons with whom you have a good relationship, but whom you do not (yet) consider a 'real' friend.
Friendly relationship	Persons with whom you regularly have pleasant contact during classes. The contact could grow into a friendship.
Neutral relationship	Persons with whom you have not much in common. In case of an accidental meeting the contact is good. The chance of it growing into a friendship is not large.
Unknown person	Persons whom you do not know.
Troubled relationship	Persons with whom you can't get on very well, and with whom you definitely do not want to start a relationship. There is a certain risk of getting into a conflict.

TABLE I: Labels and descriptions for the response categories of the Van de Bunt 1999 student survey dataset.



(a) In Degree Distribution (b) Out Degree Distribution

Fig. 1: Kernel density plots for average degree and outdegree across all observations. Best friendship is given by the black dotted line, friendship by the red dotted line, and friendly friendship by the green dotted line. The bold, solid, black line corresponds to the sum of all three friendly relationships.

	$1 \to 2$	$2 \to 3$	$3 \to 4$	$4 \to 5$	$5 \to 6$
$1 \to 1$	87	94	98	140	130
$1 \to 0$	43	52	77	90	38
$0 \to 1$	23	36	48	35	100
$0 \to 0$	871	842	801	759	756
J	.57	.52	.44	.53	.49

TABLE II: Network turnover frequency corresponding to the period indicated by column and the tie outcomes indicated by row. The Jaccard index J is reported for each period on the bottom row. For example, the cell corresponding to row $1 \to 0$ and column $1 \to 2$ has value 43, meaning that 43 relationships were present in the first observation but deleted during the second.

the first few months, but diverged afterwards (especially for the two year program).

Figure 1 displays kernel density plots on the in degree and outdegree distributions of the survey responses for the friendly relationships (i.e. best friendship, friendship, and friendly friendship). Experience with such datasets suggests that meaningful relationship definitions (i.e. friendship) should result in degree distributions with most probability mass between three and eight; the density plots indicate a compliance with such a criteria. As in the original study, we dichotomize friendly relationship, friendship, and best friendship into a present tie, or a 1 on the digraph, and all other ties to a 0.

Another important feature of network data is the amount of tie turnover–that is, the amount of links or ties which persist from observation to observation versus those that change. Table II shows the frequency of the four possible outcomes for link transition. The Jaccard similarity coefficient J is also reported, which is a good indication of how much network turnover occurs from period to period.[2] The number of persistent ties generally increases from period to period, and tie creation generally increases substantially over the life of the study. With Jaccard indices close to .5, we should have reasonable power to estimate statistical parameters (see [13]).

This data is available for download from the Siena website at http://stat.gamma.rug.nl/siena_datasets.htm.

III. Stochastic Actor Oriented Models (SAOM)

A social network composed of n actors is modeled as a directed graph (digraph), represented by an adjacency matrix $(x_{ij})_{n\times n}$, where $x_{ij} = 1$ if actor i is tied to actor j, $x_{ij} = 0$ if i is not tied to j, and $x_{ii} = 0$ for all i (self ties are not permitted). It is assumed here that the social network evolves in continuous time over an interval $\mathcal{T} \subset \mathbb{R}^1$ according to a Markov process. Accordingly, the digraph $\mathbf{x}(t)$ models the state of social relationships at time $t \in \mathcal{T}$. Changes to the network called *updates*, occur at discrete time points defining the set $\mathcal{L} \subset \mathcal{T}$. Elements of the set are denoted L_a with consecutive natural number indices a so that $L_1 < L_2 < \ldots < L_{|\mathcal{L}|}$, where the notation $|.|$ is used to denote the number of elements in a set. The network is observed at discrete time

[2]The Jaccard index is given by

$$J = \frac{\#\{1 \to 1\}}{\#\{1 \to 1\} + \#\{1 \to 0\} + \#\{0 \to 1\}}$$

or the number of ties which persist from observation to observation divided by the sum of persistent ties and ties which are created and destroyed. The range of this measure is from 0 to 1, where higher numbers indicate less network *turnover*.

points called *observations* defining the set $\mathcal{M}$ with elements M_a indexed similarly with consecutive natural number indices a so that $M_1 < M_2 < ... < M_{|\mathcal{M}|}$. Define a set of *periods* with elements $W_a \in \mathcal{W}$ each representing the continuous time interval between two consecutive observations M_a and M_{a+1}:

$$W_a = [M_a, M_{a+1}] = \{t \in \mathcal{T} : M_a \leq t \leq M_{a+1}\}.$$

By definition, $|\mathcal{W}| = |\mathcal{M}| - 1$. When $\mathcal{L} = \mathcal{M}$, we have *full information* on the network updates over the interval $\mathcal{T}$. We use upper case to denote random variables (e.g. $\mathbf{X}, M, L$).

There is a variety of models proposed in the literature for longitudinally observed social networks. In this paper, we consider the approach proposed by [13], the stochastic actor oriented model (SAOM). Here, the stochastic process $\{\mathbf{X}(t) : t \in \mathcal{T}\}$ with digraphs as outcomes is modeled as a Markov process so that for any time $t_a \in \mathcal{T}$, the conditional distribution for the future $\{\mathbf{X}(t) : t > t_a\}$ given the past $\{\mathbf{X}(t) : t \leq t_a\}$ depends only on $\mathbf{X}(t_a)$.

From the general theory of continuous-time Markov chains [18] follows the existence of the *intensity matrix* that describes the rate at which $\mathbf{X}(t) = \mathbf{x}$ tends to transition into $\widetilde{\mathbf{X}}(t+dt) = \widetilde{\mathbf{x}}$ as $dt \to 0$:

$$q(\mathbf{x}, \widetilde{\mathbf{x}}) = \lim_{dt \downarrow 0} \frac{P\{\mathbf{X}(t+dt) = \widetilde{\mathbf{x}} \mid \mathbf{X}(t) = \mathbf{x}\}}{dt} \qquad (\widetilde{\mathbf{x}} \neq \mathbf{x}), \tag{1}$$

where $\widetilde{\mathbf{x}} \in \mathcal{X}$. The SAOM supposes that a digraph update consists of exactly one tie variable change. Such a change is referred to as a *ministep*. This property can be expressed as

$$q(\mathbf{x}, \widetilde{\mathbf{x}}) > 0 \;\Rightarrow\; \sum_{i,j} \text{abs}(\mathbf{x}_{ij} - \widetilde{\mathbf{x}}_{ij}) = 1 \tag{2}$$

where abs(.) denotes the absolute value. Therefore we can use the notation

$$q_{ij}(\mathbf{x}) = q(\mathbf{x}, \widetilde{\mathbf{x}}) \text{ where } x_{ij} \neq \widetilde{x}_{ij} \tag{3}$$

SAOMs consider two principal concepts in constructing the intensity matrix: how often actors update their tie variables and what motivates their choice of which tie variable to update. This is expressed by the formulation

$$q_{ij}(\mathbf{x}) = \lambda_i(\mathbf{x}) p_{ij}(\mathbf{x}). \tag{4}$$

The interpretation is that actor i gets *opportunities* to make an update in her/his outgoing tie at a rate of $\lambda_i(\mathbf{x})$ (which might, but does not need to, depend on the current network); if such an opportunity occurs, the probability that i selects x_{ij} as the tie variable to change is given by $p_{ij}(\mathbf{x})$. The actors are not required to make a change when an opportunity occurs, which is reflected by the requirement $\sum_j p_{ij}(\mathbf{x}) \leq 1$, without the need for this to be equal to 1. The probabilities $p_{ij}(\mathbf{x})$ are dependent on the so-called *evaluation function*, as described below.

A. Rate Function

The rate function describes the rate at which an actor i updates tie variables. Waiting times between opportunities for actor i to make an update to the digraph are exponentially distributed with rate parameter $\lambda_i(\mathbf{x})$, and it follows that waiting times between any two opportunities for updates across all actors are exponentially distributed with rate parameter

$$\lambda_+(\mathbf{x}) = \sum_i \lambda_i(\mathbf{x}). \tag{5}$$

It is possible to specify any number of functional forms for $\lambda_i(\mathbf{x})$, to include combinations of actor-level covariates and structural properties of the current state of the network $\mathbf{x}$; however, in many applications, rate functions are modeled as constant terms.

B. Evaluation Function

Once an actor i is selected for an update, the actor must select a tie variable x_{ij} to change. Define $\mathbf{x}(i \rightsquigarrow j) \in \mathcal{X}$ as the digraph resulting from actor i modifying his tie variable with j during a given time period t so that $x_{ij}(i \rightsquigarrow j) = 1 - x_{ij}$, and formally define $\mathbf{x}(i \rightsquigarrow i) = \mathbf{x}$.

The SAOM assumes that the probabilities $p_{ij}(\mathbf{x})$ depend on the *evaluation function* that gives an evaluation of the attraction toward each possible next state of the network, denoted here by $f_{ij}(\mathbf{x})$. This attraction is conveniently modeled as a linear combination of the relevant features of each potential change $i \rightsquigarrow j$:

$$f_{ij}(\mathbf{x}) = \beta^T s_i(\mathbf{x}(i \rightsquigarrow j)) \tag{6}$$

where s_i is a vector-valued function containing structural features of the digraph as seen from the point of view of actor i, and β is a statistical parameter. [13], following the the econometric literature on discrete choice (see, e.g., [19], [20]), models the choice of $i \rightsquigarrow j$ as a myopic, stochastic optimization of a conditional logit. This amounts to choosing the greatest value $f_{ij}(\mathbf{x}) + \epsilon_{ij}$, where ϵ_{ij} is a Gumbel distributed error term. This leads to the conditional choice probabilities $p_{ij}(\mathbf{x})$ that actor i chooses to change tie variable $i \rightsquigarrow j$ given the current digraph $\mathbf{x}$:

$$p_{ij}(\mathbf{x}) = \frac{\exp f_{ij}(\mathbf{x})}{\sum_{k=1}^{n} \exp f_{ik}(\mathbf{x})}. \tag{7}$$

In accordance with the formal definition $\mathbf{x}(i \rightsquigarrow i) = \mathbf{x}$, the choice $j = i$ is interpreted as keeping the current digraph as it is, without making a change.

For a thorough menu of what kinds of statistics s_i are appropriate for actor oriented models, see [14], [21]. We will present here those fundamental structural statistics which are used in this study: outdegree (density), reciprocity, and transitive triplet, three cycle, betweenness, and in-degree popularity effects. Figure 2 illustrates the effects graphically, and gives mathematical definitions.

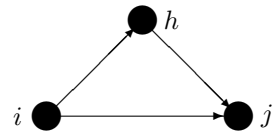
1) *Outdegree (density) effect*, defined by the number of outgoing ties for some given ego i (see Figure 2a).

$s_{i1}(\mathbf{x}) = \sum_j x_{ij} = 1$

(a) Outdegree

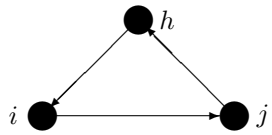
$s_{i2}(\mathbf{x}) = \sum_j x_{ij}x_{ji} = 1$

(b) Reciprocity



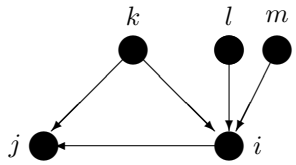
$s_{i3}(\mathbf{x}) = \sum_{j,h} x_{ij}x_{ih}x_{hj} = 1$

(c) Transitive Triplet



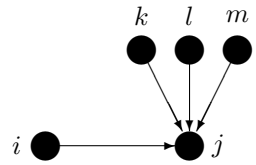
$s_{i4}(\mathbf{x}) = \sum_{j,h} x_{ij}x_{jh}x_{hi} = 1$

(d) Three cycle



$s_{i5}(\mathbf{x}) = \sum_{j,h} x_{hi}x_{ij}(1 - x_{hj}) = 2$

(e) Betweenness



$s_{i6}(\mathbf{x}) = \sum_j x_{ij} \sum_h x_{jh} = 4$

(f) In-degree popularity

Fig. 2: Graphical representation of structural effects

2) *Reciprocity effect*, defined by the number of outgoing ties that are matched (or *reciprocated*) by a corresponding incoming tie (see Figure 2b).
3) *Transitive Triplet*, defined by the number of patterns matching Figure 2c. With a positive coefficient, this effect represents a tendency towards network closure.
4) *Three cycles* [21] states that this effect may be regarded as a generalized reciprocity. In conjunction with positive estimates for network closure (e.g. transitive triplet), a negative three cycles coefficient indicates a tendency towards local heirarchy (see Figure 2d).
5) *Betweenness* represents the tendency of actors to want to position themselves between actors who are not tied to each other (see Figure 2e).
6) *In-degree Popularity* represents the tendency of actors to want to form relationships with high in-degree (i.e. popular) alters (see Figure 2f).

These effects features are regarded as representing fundamental aspects of social network dynamics, but more features may also be formulated. One possibility is to include exogenous actor-level covariates $\mathbf{c} \in \mathbb{R}^n$. Throughout this study, we will make use of sex, smoking behavior, and program membership to help explain variation in tie choices. Figure III illustrates graphically how the four possible combinations of ego/alter ties are modeled. For the multi-valued covariate program membership, we use only the *same covariate effect* (i.e. only situations in Tables IIIc and IIId are differentiated).

See also [7], [13], [14], [22] for thorough developments of possible statistics and extensions to models which consider the coevolution of network and endogenous behavioral characteristics.

C. Time Heterogeneous Parameters

Using the score type test of [16], [1] proposed a test for time heterogeneous parameters in SAOMs. Their formulation considers that the parameters β in (6) are permitted to vary over time. Formally, consider a SAOM formulated as in (6) with some set of effects $\mathcal{K} = \{K_k : k \in \mathbb{N}_1\}$ included. We initially assume that β does not vary over time, yielding a *restricted model*. Our data contains $|\mathcal{M}| \ll |\mathcal{L}|$ observations, so we estimate the restricted model by the method of moments mentioned in Section III-D. For reasons introduced in Section **??**, we wish to test whether the *restricted model* is misspecified with respect to time heterogeneity. An *unrestricted model* which allows for time heterogeneity in all of the effects is considered as a modification of (6):

$$f_{ij}^{(a)}(\mathbf{x}) = \sum_{K_k \in \mathcal{K}} \left(\beta_k + \delta_k^{(a)}\right) s_{ik}(\mathbf{x}(i \rightsquigarrow j)) \tag{8}$$

where $\delta_k^{(a)}$ is called the time dummy interacted effect parameter for effect k and period a. Define also the vectors $\delta_k = \left(\delta_k^{(2)}, ..., \delta_k^{(|\mathcal{W}|)}\right)$ and $\delta = \left(\delta_1, ..., \delta_{|\mathcal{K}|}\right)$.

$\mathbf{x}$	Ego $s_{i11}\{\mathbf{x}\}$ $\sum_j x_{ij}v_i$	Same $s_{i12}\{\mathbf{x}\}$ $\sum_{ij} x_{ij}\mathbf{1}(v_i = v_j)$	Ego x Alter $s_{i13}\{\mathbf{x}\}$ $\sum_{ij} x_{ij}v_iv_j$	$\Delta f_{ij}\{\mathbf{x}\}$
i ○ → ● j (a) $x_{ij} : v^{01}$				0
i ● → ○ j (b) $x_{ij} : v^{10}$	1			β_{11}
i ○ → ○ j (c) $x_{ij} : v^{00}$		1		β_{12}
i ● → ● j (d) $x_{ij} : v^{11}$	1	1	1	$\beta_{11} + \beta_{12} + \beta_{13}$

● corresponds to $v_i = 1$

$\Delta f_{ij}\{\mathbf{x}\}$ corresponds to the change statistic for the proposed tie indicated in the column $\mathbf{x}$.

TABLE III: Graphical representation of ego-alter covariate (selection) effects. The table illustrates how all of the possible selections of smoker/non-smoker alters by smoker/non-smoker egos are explicitly enumerated with the parameters of the model. A cell with value 1 indicates that the statistic is increased by the indicated tie. The inferences drawn in the results section will add the parameters together with the formula indicated in the right hand column.

Equation (8) applies for updates occurring during the period W_a. By convention, $\delta_k^{(1)} = 0$ for all k such that $K_k \in \mathcal{K}$ so that the first period is called the *base period*; therefore, the vector of time dummy interacted effect parameters δ has length $(|\mathcal{W}|-1)|\mathcal{K}|$.[3] The test of [1] is the following omnibus test:

$$\begin{aligned} H_0 &: \delta = \mathbf{0} \\ H_1 &: \delta \neq \mathbf{0}. \end{aligned} \tag{9}$$

This approach departs from [2] in at least one important way. [2] estimates a model for each intervening period between observations separately. This is an equivalent approach to providing time dummies (**??**) for all effects and all periods. This approach estimates a model jointly across all observations, but only includes time dummies for those effects which contain evidence for time heterogeneity.

D. Estimation

A key feature of Equation (7) is the convenient form of its *log odds ratios* between any two potential next networks $\mathbf{x}(i \rightsquigarrow j)$ and $\mathbf{x}(i \rightsquigarrow k)$:

$$\begin{aligned} \log\left[\frac{p_{ij}(\mathbf{x})}{p_{ik}(\mathbf{x})}\right] &= \log\left[\frac{\dfrac{\exp f_{ij}(\mathbf{x})}{\sum_h \exp f_{ih}(\mathbf{x})}}{\dfrac{\exp f_{ik}(\mathbf{x})}{\sum_h \exp f_{ih}(\mathbf{x})}}\right] \\ &= \log\left[\frac{\exp f_{ij}(\mathbf{x})}{\exp f_{ik}(\mathbf{x})}\right] \\ &= f_{ij}(\mathbf{x}) - f_{ik}(\mathbf{x}) \end{aligned} \tag{10}$$

This is the characteristically simple property that makes estimation in classical discrete choices quite straightforward (see, e.g., [23]), and is also the basis for the SAOM. If the network updates $\mathcal{L}$ are all observed so that we have full information (i.e. $\mathcal{L} = \mathcal{M}$), and in the usual situation that the rate parameters are independent of the parameters of the objective function, a convenient partial likelihood is available for the statistical parameters of the objective function. We use the notation $\mathbf{x}^{(a)} = \mathbf{x}(m_a) = \mathbf{x}(l_a)$ for the a-th network observed in the dataset.

Define a vector of binary variables $\mathbf{d}$ such that

$$d_{ij}^{(a)} = \begin{cases} 1 & \text{if} \quad \mathbf{x}^{(a+1)} = \mathbf{x}^{(a)}(i \rightsquigarrow j) \\ 0 & \text{if} \quad \mathbf{x}^{(a+1)} \neq \mathbf{x}^{(a)}(i \rightsquigarrow j) \end{cases}, \tag{11}$$

which denotes whether actor i selected variable $x_{ij}^{(a)}$ to update. Note that $\sum_{j=1}^n d_{ij}(a) = 1$ for all a. The partial log

[3] Because $\delta_k^{(1)}$ is fixed, it is implicitly omitted from δ_k and δ throughout the notation.

likelihood for the objective function parameters is

$$l(\beta) = \sum_{ija} d_{ij}^{(a)} \log p_{ij}(\mathbf{x}^{(a)})$$
$$= \sum_{ija} d_{ij}^{(a)} \log \left(\frac{\exp f_{ij}(\mathbf{x}^{(a)})}{\sum_k \exp f_{ik}(\mathbf{x}^{(a)})} \right) \quad (12)$$

Under regularity conditions, a solution $\hat{\beta}_{\text{ML}}$ to

$$\nabla l(\beta) = 0$$

where

$$\nabla = (\frac{\partial}{\partial \beta_1}, \frac{\partial}{\partial \beta_2}, ...)$$

is a maximum likelihood estimate for β.

Unfortunately, complete information on each network update is extremely rare, and a likelihood function as in (12) is not readily available. Even when tie creation is observed in continuous time, it is not often that observations concerning termination of ties are also available. It is far more often the case that the data observed will be in the form of panel data, where typically $|\mathcal{M}| \ll |\mathcal{L}|$. Accordingly, the method of moments as proposed by [13] can be used as an alternative method for obtaining reasonable estimates. Consider the estimating function

$$g_n(\theta; z_n) = \sum_{m_a \in \mathcal{M}} \Big(E_\theta \big\{ u\big(\mathbf{X}^{(a)}\big) \mid \mathbf{X}^{(a-1)} = \mathbf{x}^{(a-1)} \big\} - u\big(\mathbf{x}^{(a)}\big) \Big), \quad (13)$$

which is simply the sum of deviations between the expected value of the statistics for the simulated networks and the observed networks; z_n simply means all of the available data, and θ is a vector of parameters for the objective and rate functions described earlier. $u(\mathbf{x})$ is a function that corresponds to appropriately chosen statistics calculated from the digraph for the parameters θ (based on the statistics in Section III-B). The method of moments involves finding the *moment estimate* $\hat{\theta}$ solving the moment equation

$$g_n(\theta; z_n) = 0 \quad (14)$$

The specific details of fitting moment estimates are rather involved, and entail simulating networks $\mathbf{X}^{(a)}$ many times to achieve a reliable result for the expectation in (14).[4] This simulation is very straightforward. Take an initial network $\mathbf{x}(l_1)$ and proceed as follows for each update $l_a \in \mathcal{L}$:

1) Set $l_a = l_{a-1} + \text{Expon}(\lambda_+)$
2) Select actor i with probability $\dfrac{\lambda_i(l_{a-1})}{\lambda_+(l_{a-1})}$.
3) Select actor j with probability $p_{ij}(l_{a-1})$.
4) If $i \neq j$, set $x_{ij}(l_a) = 1 - x_{ij}(l_{a-1})$.
5) Repeat until some specified conditions (e.g. number of updates $|\mathcal{L}|$ or some holding time $(l_{|\mathcal{L}|} - l_a)$ is exceeded) are satisfied.

[4]That this can take a considerable amount of time per estimation motivates the use of the score-type test of [1]

See [13] for guidelines on the selection of appropriate statistics for $u(\mathbf{X}(t))$ and information on how to estimate the root of $g_n(z_n, \theta)$, and [24] for the estimation of the derivative matrix, covariance matrix, and standard errors.

E. Modeling approach

We aim to analyze the data of [2], [15] with a time heterogeneous model specification, as in Equation (8). Using the forward model-selection technique of [1] and elaborated in [17], we will select a model with time dummy interacted parameters until the test in (9) fails to reject H_0 (this model will be called Model B.3, see below). We will then consider subsets of these parameters as alternate model specifications and estimate them using method of moments:

- **Model A.1**: Structural effects only. This model will be used primarily to assess the stability of the structural effects as we add more parameters.
- **Model A.2**: Structural effects and covariate effects. Parameter estimates for this model will yield a basis for comparison, so that we may compare the interpretation that would result from ignoring potential time heterogeneity.
- **Model B.1**: Structural effects only, but using the structural effect time heterogeneity parameters included in Model B.3. As before, this model will be used primarily to assess the stability of the structural effects as we add more parameters.
- **Model B.2**: Structural effects and covariate effects, but using the structural effect time heterogeneity parameters included in Model B.3. No covariate time heterogeneity parameters are used. This will help us to assess whether any time heterogeneity ultimately found in them under Model B.3 can be explained by time heterogeneity in the structural effects.
- **Model B.3**: This is the model which, after using the forward model selection of [1], fails to reject H_0 for (9), indicating that much of the time heterogeneity in the chosen effects has been accounted for.

All of the model terms will include rate/covariate interaction terms presented in **??** for smoking behavior, sex, and program membership. After estimating these models, we use Equation (7) to calculate conditional probabilities for smoker and non-smoker alter selection so that we may more easily interpret Model A and Model B.

IV. Results

The results of the five models presented in Section III-E are presented in Table IV. Note that the list of effects included in the first column is not exhaustive for time heterogeneity terms; they represent the final model arrived at after using the model selection procedure of [1], [17], which incrementally adds time dummy interected parameters until the joint test fails to reject H_0 given in (9). Perhaps surprisingly, the program/rate interaction terms are near-zero with large standard errors across all models and are therefore not presented.

Across all estimations, convergence is quite good, as evidenced by very low t^* test statistics (see [14] for more information).

A. Model A: Time homogeneous parameters

Estimates for structural effects are stable across Model A.1 and Model A.2 (with the exception of outdegree), which is encouraging. We would expect outdegree to be systematically different on the basis of Table III, since inclusion of the other covariate effects turns the outdegree effect into a sort of *base case*. Interestingly, the relatively large standard error and small magnitude of the outdegree estimate indicates that we have little evidence to reject $H_0 : \beta_1 = 0$. As the network is rather small, and there are three effects with large (negative) and significant estimates, this result should not be of much concern. The structural effects results indicate reciprocity and a tendency towards closure as we expect in many social settings. A negative three cycles effect indicates some tendencies towards local heirarchy. The small and negative betweenness coefficient could proxy for tendencies towards network closure, or it could be an artifact of a small network.

We see, as might be expected, that membership in the same program causes a slight increase in probability of friendship formation.[5]

Using the result of Table III, it is possible to show that the probability of a male selecting a male is .24 greater than selecting a female, *ceteris paribus*. Women select women versus men with probility .12 less than selecting a male, indicating a universally greater tie-formation attractiveness for males. These results can be calculated directly from (7) when considering an ego i, conditioned on the sex of i and on forcing a tie formation for i, with two opportunities $i \rightsquigarrow j$ and $i \rightsquigarrow k$ where j and k have opposite sexes.

On the basis of Model A's estimation results, we might conclude that smoking has no significant effect on friendship formation, due to the large standard errors associated with the parameter estimates for the smoking covariate effects. Interpretation of the parameters in a similar manner as performed for sex yields the result that smokers choose smoking alters with probability .55, and non-smokers choose smokers with probability .44. This indicates a slight tendency towards universally more tie-formation related attractiveness of smoking alters, but again, due to the size of the standard errors, parameter interpretation of Model A would likely fail to uncover a meaningful relationship between smoking behavior and social relationship formation.

We note that the time heterogeneity test of [1] for Models A.1 and A.2 rejects the hypothesis of time homogeneous parameters, which helps to motivate the exploration of Model B.

[5]Using Equation (7), we expect roughly a .13 increase in probability for x_{ij} if i, j are in the same program. This result comes from considering a notional case of a some selected alter considering the ties $i \rightsquigarrow j$ and $i \rightsquigarrow k$ exclusively. Using Equation (7) and the estimates of Table IV, it is easy to show that if i, j are in the same program and i, k are not, $p_{ij} = .627$.

B. Model B: Time heterogeneous parameters

We have a similar result in insignificant outdegree estimates but again due to the small size of the network an presence of large (negative), significant estimates for other effects, there is little reason for concern. The structural effects largely agree with the results of Model A in both magnitude and in direction, so we still detect tendencies towards network closure, reciprocity, and local heirarchy. Where the models begin to differ is in the standard errors of the covariate effects, and in the magnitudes of the smoking effects. Where Model A gives us little reason to think that smoking is an important factor for explaining network formation, the largest estimates (in magnitude) for Model B are actually smoker selection effects. In order to fix ideas on how these two models differ, we construct a so-called *ego-alter selection table*, where we suppose again that some ego i has a fixed smoking behavior, and that i is forced to form a tie with one of two alters, j or k. These two alters have opposite smoking behaviors, and we consider the probability that i selects either on the basis of the parameter estimates.

C. Ego/alter selection results for Model A and Model B

Table V presents the ego/alter selection table. Using the coefficients indicated by Table III and the estimates from Table IV, we estimate $f_{ij}(\mathbf{x})$ for each of the four possible combinations of smoker/non-smoker ego/alter. The estimates reported are calculated from column two for each of the models, following Table III. The probabilities reported in parenthesis refer to the proportion of times that some ego i with a given smoking behavior v_i would form a tie with alter j with a given smoking behavior v_j instead of another alter k with smoking behavior $1 - v_j$ in a notional, empty network containing actors i, j, k only. These probabilities are available directly from Equation (7).

Models A.1 and B.1 are not of much interest for smoker selection interpretation, as they do not include any terms for smoking covariates (accordingly, the probabilites are equal for smokers and non-smokers). Model A.2 suggests that actors have a rather small preference on the basis of smoking behaviors, a result discussed in the last section. The selection probabilities differ by a rather small amount (roughly .05 in either case), and both smokers and non-smokers appear to have a slight selection preference for smoking alters. In other words, the data supports the notion that smokers are (slightly) universally more popular, rather than that smokers select smokers and non-smokers select non-smokers. We note, however, that these results would have to be presented with caution, due to the size of the corresponding standard error (which is roughly twice the size of $\hat{\beta}_{13}$, e.g.).

Turning to Model B, it is apparent that the inclusion of time heterogeneous parameters lends a more substantive interpretation. An immediate difference between Model A and Model B is that the standard errors are smaller in the estimates of smoking covariate effects when time heterogeneous parameters are specified, so we can draw inference with less uncertainty. Further, the difference in magnitudes

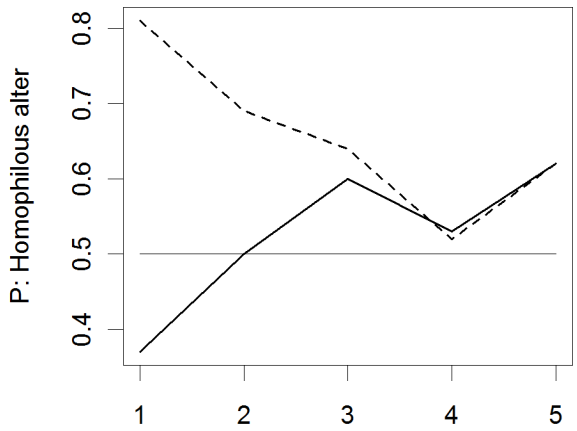


Fig. 3: Selection probabilities given by estimates for Model B.3 shown in Table V. The dotted line represents the probability that an ego who smokes selects another smoking ego for tie formation as opposed to a non-smoker. The solid line represents the probability that a non-smoker selects another non-smoker for tie formation as opposed to a smoker. A horizontal reference line is given at $p = .5$ to show estimates which indicate homophilous selection behaviors.

of the smoking covariate effect parameter estimates between the two models is substantial. Model B.2, which controls for heterogeneity in structural effects but not selection effects, indicates that smokers are universally more popular as alter choices regardless of the ego smoking behavior, and that there is an additional homophily effect of roughly .07 for smokers to select other smokers. After permitting smoking selection effects to vary over time in Model B.3, we a more pronounced version of the same result as in Model B.2 for periods 1 and 2, and the effect diminishes for periods 3, 4, and 5. After Period 2, smokers are no longer universally more attractive, and we see a homophily effect form on the basis of smoking behaivor. Figure 3 illustrates the selection probabilities for smoker/smoker, nonsmoker/nonsmoker selection given by the results of Model B.3. [1] finds that omission of time heterogeneity causes estimates which in some sense *average* over the heterogeneity, and it seems that perhaps we have a similar result here, where the effect is in some sense covered up in time homogeneous models.

V. Discussion

With a time homoegenous model specification, results indicated no sound support for the notion that smoking behavior had a substantive impact on the formation of social relationships. Using a model with time heterogeneity terms, we were able to uncover smoking homophily effects after period 1 had passed. The findings of this study could have important implications in two respects: First, the use of time heterogeneous specifications can be potentially very important towards interpretation of parameters of interest in an array of applied studies. As alluded to in [1], time heterogeneity in parameters of interest can be intrinsically interesting. This study illustrates a case where it is not only interesting, but failing to account for behavioral changes over time can lead to misspecified models which cover up important social dynamics. Second, while the size and scope of this small study is rather limited, and the results may not be generally applicable, more research is indicated in the area of policy regarding the establishment of smoking areas. The researchers note that these smoking areas separated smokers and non-smokers during coffee breaks between lectures.

The original study of [2] was unable to uncover a coherent picture of how smoking affected friendship formation. In contrast, the forward model selection and time heterogeneity testing employed by this paper was able to discover a coherent picture of smoking effects which decline over time, but are important in the initial vetting of potential friendships.

We cannot ascertain influence effects from this particular dataset, but future research might study how smoking areas reinforce smoking behaviors.[6] It is plausible that smoking areas increase smoking homophily selection effects–a notion supported by the results of this study–and also plausible that an ego's friendly contacts' smoking behavior influences the ego's smoking behavior.

As stated before, the results of this study may not be generally applicable for a number of reasons. The dataset is small, and a number of students failed to respond, which may have caused some serious biases in the results. As noted by [2], the operationalization of friendship was not optimal, and perhaps a social activity scale would have been better. A number of covariates have been left out which may have had some important explanatory power, e.g. performance at the school or extracurricular activities.

What we can say with some certainty is that time heterogeneity is a potentially important feature of actor behavior, and that failing to account for it in model selection can cause results which vary widely from a time heterogeneous specification. While positive statements concerning the context in which the study was conducted in, i.e. regarding policies on smoking areas, are tenuous, the results do motivate further research into the role of policies regarding the establishment of smoking areas on a broader scale.

VI. Future Work

This paper represents a contribution to the methodology of studying risky behavior and social networks. The dataset used was selected because of its position as one of the first datasets in the literature to be analyzed with a stochastic actor oriented model. The results of the study indicate that time heterogeneity can play an important role in what inference is drawn from a study at hand.

Nonetheless, the context for this methods study could be enlarged into a wide-scale study of smoking behavior and social network evolution. [26] has conducted a trial study which could be enlarged in such a manner. Surveys are given

[6]The smoking behavior was collected at only one time period. In other studies, e.g. [25], such information is collected longitudinally.

to school children periodically which ask for relationships of friendship and for personal assessments of risky behaviors. Many classrooms and schools are observed over the period of perhaps two or three years. A multilevel approach is then used to isolate classroom effects from the main effects, and inference is drawn on the latter. Using the evidence from this paper, attention could be paid to time heterogeneity across these two years when drawing inference.

Determining the risk factors for adolescent smoking is among the highest priority for researchers because of the well established adverse health effects from smoking. By collecting data over time and using the stochastic actor oriented framework, previously elusive peer effects can be disentangled from selection effects. If these risk factors can be identified, intervention programs can be made more effective.

REFERENCES

[1] J. Lospinoso, M. Schweinberger, T. Snijders, and R. Ripley, "Assessing and accounting for time heterogeneity in stochastic actor oriented models," *Advances in Data Analysis and Computation*, vol. Special Issue on Social Networks (Under Review), 2010.

[2] G. G. Van De Bunt, M. A. J. Van Duijn, and T. Snijders, "Friendship networks through time: An actor-oriented dynamic statistical network model," *Comput. Math. Organ. Theory*, vol. 5, no. 2, pp. 167–192, 1999.

[3] L. Berkman and L. Syme, "Social networks, host resistance, and mortality: A nine year follow up study of alameda county residents," *American Journal of Epidemiology*, vol. 109, no. 2, pp. 186–204, 1979.

[4] L. F. Berkman, "Assessing the physical health effects of social networks and social support," *Annual Review of Public Health*, vol. 5, no. 1, pp. 413–432, 1984.

[5] S. Cohen, "Social relationships and health," *American Psychologist*, vol. 69, no. 8, pp. 676–684, 2004.

[6] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual Review of Sociology*, vol. 27, no. 1, pp. 415–444, 2001.

[7] T. Snijders, C. Steglich, and M. Schweinberger, "Modeling the co-evolution of networks and behavior," in *Longitudinal models in the behavioral and related sciences*, K. van Montfort, H. Oud, and A. Satorra, Eds. Lawrence Erlbaum, 2007, pp. 41–71.

[8] W. Burk, C. Steglich, and T. Snijders, "Beyond dyadic interdependence: Actor-oriented models for co-evolving social networks and individual behaviors," *International Journal of Behavioral Development*, vol. 31, pp. 397–404, 2007.

[9] C. Steglich, T. Snijders, and P. West, "Applying siena: An illustrative analysis of the co-evolution of adolescents' friendship networks, taste in music, and alcohol consumption," *Methodology*, vol. 2, pp. 48–56, 2006.

[10] S. T. Ennett and K. E. Bauman, "Peer group structure and adolescent cigarette smoking: A social network analysis," *Journal of Health and Social Behavior*, vol. 34, no. 3, pp. 226–236, 1993.

[11] N. A. Christakis and J. H. Fowler, "The collective dynamics of smoking in a large social network," *New England Journal of Medicine*, vol. 358, no. 21, pp. 2249–2258, 2008.

[12] B. Houle and M. Siegel, "Smoker-free workplace policies: Developing a model of public health consequences of workplace policies barring employment to smokers," *Tobacco Control*, vol. 18, no. 1, pp. 64–69, 2009.

[13] T. Snijders, "The statistical evaluation of social network dynamics," in *Sociological Methodology*, M. Sobel and M. Becker, Eds. Boston and London: Basil Blackwell, 2001, pp. 361–395.

[14] T. Snijders, C. Steglich, and C. van de Bunt, "Introduction to actor-based models for network dynamics," *Social Networks*, vol. 32, pp. 44–60, 2010.

[15] G. G. V. de Bunt, "Friends by choice. an actor-oriented statistical network model for friendship networks through time," Ph.D. dissertation, University of Groningen, 1999.

[16] M. Schweinberger, "Statistical methods for studying the evolution of networks and behavior," Ph.D. dissertation, University of Groningen, 2007.

[17] J. Lospinoso, "Testing and modeling time heterogeneity in longitudinal studies of social networks: A tutorial in rsiena," *Connections*, vol. Under review., 2011.

[18] J. Norris, *Markov Chains*. Cambridge University Press, 1997.

[19] G. Maddala, *Limited-dependent and Qualitative Variables in Econometrics*, 3rd ed. Cambridge University Press, 1983.

[20] D. McFadden, "Conditional logit analysis of qualitative choice behavior," in *Frontiers in Econometrics*, P. Zarembka, Ed. Academic Press, 1973, pp. 105–142.

[21] R. Ripley and T. Snijders, *Manual for RSiena version 4.0*, 2009.

[22] T. Snijders, "Longitudinal methods of network analysis," in *Encyclopedia of Complexity and System Science*, B. Meyers, Ed. Springer, 2009, pp. 5998–6013.

[23] W. Greene, *Econometric Analysis*, 6th ed. Prentice Hall, 2007.

[24] M. Schweinberger and T. Snijders, "Markov models for digraph panel data: Monte carlo-based derivative estimation," *Comput. Stat. Data Anal.*, vol. 51, no. 9, pp. 4465–4483, 2007.

[25] C. E. Steglich, T. A. Snijders, and M. Pearson, "Dynamic networks and behavior: Separating selection from influence," *Sociological Methodology*, vol. To be published., 2010.

[26] L. Mercken, T. Snijders, C. Steglich, E. Vartiainen, and H. de Vries, "Dynamics of adolescent friendship networks and smoking behavior," *Social Networks*, vol. 32, pp. 72–81, 2010.

Effect	Coefficient	Model A.1	Model A.2	Model B.1	Model B.2	Model B.3
Outdegree	$\hat{\beta}_1$	-.18 (.25)	-.30 (.37)	-.24 (.29)	-.10 (.41)	-.01 (.44)
Reciprocity	$\hat{\beta}_2$	1.48 (.15)	1.38 (.14)	1.53 (.16)	1.36 (.16)	1.38 (.16)
$\rightarrow 5$	$\hat{\delta}_2^{(5)}$			-.06 (.27)	-.05 (.27)	-.40 (.30)
Trans. Triplets	$\hat{\beta}_3$	.42 (.03)	.42 (.03)	.54 (.04)	.55 (.04)	.56 (.04)
$\rightarrow 3$	$\hat{\delta}_3^{(3)}$			-.11 (.06)	-.12 (.06)	-.22 (.09)
$\rightarrow 4$	$\hat{\delta}_3^{(4)}$			-.19 (.07)	-.19 (.07)	-.25 (.09)
$\rightarrow 5$	$\hat{\delta}_3^{(5)}$			-.31 (.06)	-.31 (.06)	-.38 (.08)
Three Cycles	$\hat{\beta}_4$	-.45 (.06)	-.43 (.06)	-.55 (.07)	-.52 (.07)	-.53 (.07)
$\rightarrow 2$	$\hat{\delta}_4^{(2)}$			-.47 (.15)	-.45 (.14)	-.53 (.17)
$\rightarrow 4$	$\hat{\delta}_4^{(4)}$			.38 (.11)	.39 (.11)	.37 (.11)
Betweenness	$\hat{\beta}_5$	-.24 (.05)	-.20 (.04)	-.29 (.06)	-.26 (.05)	-.27 (.05)
In-deg. Popul.	$\hat{\beta}_6$	-.14 (.03)	-.13 (.02)	-.15 (.03)	-.14 (.03)	-.14 (.03)
Sex Ego	$\hat{\beta}_7$		-.12 (.15)		-.08 (.18)	-.08 (.18)
Same Sex	$\hat{\beta}_8$		-.47 (.22)		-.40 (.24)	-.40 (.24)
Sex $i \times j$	$\hat{\beta}_9$		1.49 (.46)		1.50 (.53)	1.46 (.53)
Same Program	$\hat{\beta}_{10}$		.52 (.08)		.59 (.10)	.59 (.10)
Smoke Ego	$\hat{\beta}_{11}$		-.02 (.14)		.08 (.16)	.10 (.17)
$\rightarrow 2$	$\hat{\delta}_{11}^{(2)}$					-.64 (.34)
$\rightarrow 5$	$\hat{\delta}_{11}^{(5)}$					-1.20 (.33)
Same Smoke	$\hat{\beta}_{12}$		-.24 (.51)		-.55 (.55)	-.52 (.57)
$\rightarrow 2$	$\hat{\delta}_{12}^{(2)}$					.50 (.43)
$\rightarrow 3$	$\hat{\delta}_{12}^{(3)}$					.93 (.47)
$\rightarrow 4$	$\hat{\delta}_{12}^{(4)}$					.66 (.57)
$\rightarrow 5$	$\hat{\delta}_{12}^{(5)}$					1.03 (.54)
Smoke $i \times j$	$\hat{\beta}_{13}$		.46 (1.03)		1.46 (1.12)	1.94 (1.16)
$\rightarrow 2$	$\hat{\delta}_{13}^{(2)}$					-1.11 (1.15)
$\rightarrow 3$	$\hat{\delta}_{13}^{(3)}$					-1.76 (1.20)
$\rightarrow 4$	$\hat{\delta}_{13}^{(4)}$					-1.98 (1.39)
$\rightarrow 5$	$\hat{\delta}_{13}^{(5)}$					-1.96 (1.30)
Rate 1	$\hat{\lambda}_1^{(1)}$	3.51 (.82)	3.42 (.75)	3.30 (.52)	3.31 (.48)	3.29 (.51)
Rate 2	$\hat{\lambda}_1^{(2)}$	4.91 (.83)	4.82 (.72)	4.62 (.63)	4.58 (.61)	4.61 (.65)
Rate 3	$\hat{\lambda}_{1^{(3)}}$	7.87 (1.23)	7.59 (1.14)	7.76 (1.21)	7.68 (1.00)	7.92 (1.05)
Rate 4	$\hat{\lambda}_1^{(4)}$	6.03 (.64)	6.28 (.57)	6.18 (.61)	6.24 (.65)	6.04 (.66)
Rate 5	$\hat{\lambda}_1^{(5)}$	7.08 (.93)	7.14 (.88)	7.09 (.78)	7.06 (.82)	7.13 (.81)
Smoke x Rate	$\hat{\lambda}_2$	-.14 (.34)	-.20 (.38)	.07 (.38)	.08 (.15)	.08 (.16)
$\rightarrow 2$	$\hat{\delta}_2^{(2)}$			.09 (.31)	.07 (.30)	.05 (.32)
$\rightarrow 3$	$\hat{\delta}_2^{(3)}$			-.62 (.27)	-.58 (.31)	-.53 (.32)
Sex x Rate	$\hat{\lambda}_3$	-.30 (.32)	-.24 (.25)	-.28 (.15)	-.29 (.14)	-.26 (.16)
$H_0 : \delta_+^{(+)} = 0$	p	0	0	.01	.04	.19
Convergence	max $\lvert t^* \rvert$	.11	.07	.06	.06	.06

TABLE IV: Results of estimation for five models. Standard errors reported in parenthesis.

$\mathbf{x}$	$\Delta f_{ij}\{\mathbf{x}\}$	A.1	A.2	B.1	B.2	B.3$_1$	B.3$_2$	B.3$_3$	B.3$_4$	B.3$_5$
i ○ → ● j	$\hat{\beta}_1^*$	$-.18$	$-.30$	$-.24$	$-.10$	$-.17$	$-.17$	$-.17$	$-.17$	$-.17$
		(.5)	(.56)	(.5)	(.63)	(.63)	(.50)	(.40)	(.47)	(.38)
i ○ → ○ j	$\hat{\beta}_1^* + \hat{\beta}_{12}^*$	$-.18$	$-.54$	$-.24$	$-.65$	$-.69$	$-.19$	.24	$-.03$	.34
		(.5)	(.44)	(.5)	(.37)	(.37)	(.50)	(.60)	(.53)	(.62)
i ● → ○ j	$\hat{\beta}_1^* + \hat{\beta}_{11}^*$	$-.18$	$-.32$	$-.24$	$-.02$	$-.27$	$-.91$	$-.27$	$-.27$	-1.47
		(.5)	(.45)	(.5)	(.29)	(.19)	(.31)	(.36)	(.48)	(.38)
i ● → ● j	$\hat{\beta}_1^* + \hat{\beta}_{11}^* + \hat{\beta}_{12}^* + \hat{\beta}_{13}^*$	$-.18$	$-.10$	$-.24$	.89	1.15	$-.10$	.32	$-.17$	$-.98$
		(.5)	(.55)	(.5)	(.71)	(.81)	(.69)	(.64)	(.52)	(.62)

● corresponds to a smoker, Δf_{ij} corresponds to the change statistic for the given selection case.

TABLE V: Results for ego selection on the *ceteris paribus* basis of smoking attributes. $\hat{\beta}_k^* = \hat{\beta}_k + \hat{\delta}_k^{(m)}$, or the base estimate plus the dummy given by the corresponding column. If a coefficient was not included in a particular model, it is fixed to zero and calculated thus. Estimated selection probabilities given ego i's smoking status are given in parenthesis below each estimate (i.e. $p_{ij}(\mathbf{x}) = \dfrac{\exp(f_{ij}(\mathbf{x}))}{\exp(f_{ij}(\mathbf{x})) + \exp(f_{ik}(\mathbf{x}))}$) where k is selecting an actor with the opposite smoking behavior from j; see Equation (7)). Note that the conditional probabilities of selection for ego i given her smoking behavior sum to one.