# Simulation-based estimation of peer effects*

Brian Krauth
Simon Fraser University†

July 30, 2004

**Abstract**

The influence of peer behavior on an individual's choices has received renewed interest in recent years. However, accurate measures of this influence are difficult to obtain. Standard reduced-form methods lead to upwardly biased estimates due to simultaneity, common shocks, and nonrandom peer group selection. This paper describes a structural econometric model of peer effects in binary choice, as well as a simulated maximum likelihood estimator for its parameters. The model is nonparametrically identified under plausible restrictions, and can place informative bounds on parameter values under much weaker restrictions. Monte Carlo results indicate that this estimator performs better than a reduced form approach in a wide variety of settings. A brief application to youth smoking demonstrates the method and suggests that previous studies dramatically overstate peer influence.

## 1 Introduction

Conventional wisdom holds that the behavior of individuals, especially young people, is strongly influenced by the behavior of those around them. In recent years, economists have shown renewed interest in the study of peer effects, neighborhood effects, and other non-market social influences that have taken on the general name of "social interaction effects." Theoretical treatments by researchers including Akerlof and Kranton

*I have received helpful commentary from Kim-Sau Chung, Bernard Salanié and Adriaan Soetevent, as well as seminar participants at the University of Calgary, the 2001 Society for Computational Economics conference, the 2002 European Economic Association conference, the 2002 Econometric Society European Summer meetings, and the 2002 Canadian Econometrics Study Group. All errors are mine.

†8888 University Dr., Burnaby BC V5A 1S6, Canada. Email: bkrauth@sfu.ca.

(2000), Becker and Murphy (2001), and Brock and Durlauf (2001a) have made substantial progress in identifying the implications of social interaction effects for aggregate behavior.

In comparison, the state of empirical knowledge on social interaction effects has advanced much more slowly. While there is a long history of empirical work on social interactions, an influential article by Manski (1993) identifies serious methodological problems in the bulk of the literature. He notes that the simple reduced-form relationship between an individual's choices and those in his or her social group is a result of three distinct effects. A person's choices can be directly influenced by either the choices (what Manski calls "endogenous social effects" or simply "endogenous effects") or characteristics (what Manski calls "contextual effects") of those in his or her social group. In addition, there may be what Manski calls "correlated effects", in which individuals in a social group exhibit similar behavior because of common unobserved factors. Correlated effects can arise through simultaneity, nonrandom group selection, or common shocks. While endogenous effects, contextual effects, and correlated effects have very different policy implications, Manski demonstrates that standard methods are unable to distinguish between them. In response to this critique, several more recent studies have developed new methods of addressing at least some of these problems using experimental data, instrumental variables, and other identification strategies. While these newer methods are a significant improvement over a reduced form analysis, each has significant limitations as well.

This paper proposes a structural approach to the estimation of endogenous social effects that does not require experimental data or an instrumental variable, can be applied to commonly available survey data, and is nonparametrically identified under reasonable and transparent identifying restrictions. In addition, the estimation method provides several avenues for placing informative bounds on parameter values under weaker restrictions than those needed for point identification. The econometric model is based on Brock and Durlauf's (2001a) treatment of binary choice with endogenous social effects, but adds several features to account for correlated effects as well. Selection and common shocks are addressed by allowing both observable and unobservable characteristics to be correlated across peer group members. Simultaneity is addressed by treating peer choice as an endogenous variable. This equilibrium-based structural approach produces a log likelihood function which involves a series of high-dimensional integrals, so the model is estimated by simulated maximum likelihood.

Nonparametric identification in the model is achieved through a restriction on the correlated effects. In the baseline version of the model it is assumed is that the within-group correlation in unobservable variables is equal to the correlation in observable variables. This "equal correlation" restriction, analogous to one introduced by Altonji, Elder, and Taber (2000) to model selection issues in their analysis of Catholic school effects, will hold on average if the observable variables represent a random subset of the relevant variables. In addition to providing point estimates under the equal correlation restriction, techniques developed in this paper can also be used to find informative bounds on parameter values under much weaker restrictions on the correlated effect.

The model can be estimated from either individual-based random samples or group-based samples. Monte Carlo results indicate that the estimator performs well in moderately sized samples of either type, and is not highly sensitive to several potential forms of misspecification. As a result, the estimator developed here can have wide empirical applicability in the estimation of social interaction effects. A brief empirical example, on close friend influences in youth smoking, demonstrates one of these applications. Although a full analysis of social interaction effects in youth smoking is beyond the scope of this paper, the results suggest that friends are substantially less influential than would be implied by a reduced form analysis and that the application of structural estimation to the question of peer influence in youth smoking merits further investigation.

## 1.1 Related literature

The contemporary empirical literature dealing with social influences on individual choice starts from Manski's (1993) critique of what had until that point been the dominant mode of empirical analysis in that literature. In these early studies, social influences were measured using simple reduced form methods and standard survey data. A typical regression would include the respondent's choice as the dependent variable, and the respondent's characteristics as well as the average choice within the respondent's social group as explanatory variables. The coefficient on the reference group average choice would then be interpreted as measuring the endogenous social effect. Alternatively, the reference group average of one or more background characteristics would be used as the explanatory variable rather than the group average choice, in which case the coefficient would be interpreted as a contextual effect. Manski's critique of this approach is that the actual parameters of interest are not identified: the reduced form coefficient can be interpreted as an endogenous, contextual, or correlated effect, or some combination.

Moffitt (2001) explains the importance of distinguishing between these three effects. First, both endogenous and contextual effects imply that groups matter, i.e., an individual's social group memberships influence his or her choices. Second, endogenous effects imply a "social multiplier," i.e., the aggregate effect of a policy intervention will be larger than the individual-level direct effect. If strong enough, endogenous effects may also imply multiple group-level equilibria. If groups matter or if there are large social multipliers, then evaluation of various social policies should consider their indirect effects via the social network in addition to their direct effects. In addition, policies such as housing mobility programs (Katz, Kling and Liebman 2001) depend on the existence of strong endogenous or contextual effects to be effective. Contextual effects do not imply a social multiplier, and correlated effects imply neither that social groups matter nor that there is a social multiplier. Manski's critique thus implies that results from reduced form studies of social interaction effects have no useful policy implications.

In response to these issues, empirical researchers have pursued a number of identification strategies. One stream of the literature (Kremer and Levy 2001, Sacerdote 2001, Katz et al. 2001) focuses on special cases where individuals are randomly assigned to reference groups, so any correlated effects due to selection are avoided. Lagged peer variables are used to avoid correlated effects due to common shocks, and the coefficient on one or more peer characteristics is interpreted as a contextual effect. Another stream (Evans, Oates and Schwab 1992, Gaviria and Raphael 2001, Hoxby 2000, Ioannides and Zabel 2002) looks for credibly exogenous sources of variation in peer characteristics, and uses this exogenous variation as an instrumental variable for peer choices. If the variation in group characteristics is truly exogenous, and there is no contextual effect, this method consistently estimates the endogenous effect. A third stream (Glaeser, Sacerdote and Scheinkman 1996, Glaeser, Sacerdote and Scheinkman 2002, Topa 2001) uses structural models to infer the magnitude of endogenous effects from aggregate statistics such as social multipliers or the variance in behavior across aggregates.

Each of these approaches has both advantages and limitations. Identification strategies based on random assignment avoid correlated effects due to selection, but are only applicable to a few special cases, including first-year college roommates and government-assisted housing, where a central authority conducts the group assignment. Identification using instrumental variables requires that the IV be both credibly exogenous and relevant to the outcome. These standard requirements are particularly difficult to meet in this case because they imply the IV must be a group-level variable which affects everyone in the group except the respondent. As Brock and Durlauf (2001b) note, this means that the IV must be the group average of some individual-level variable that does not produce contextual effects. Even in cases where such a restriction is plausible, it is likely to be sufficiently controversial that it would be desirable to corroborate the results under alternative identification schemes. Identification using aggregate/structural approaches often suffers from the problem that it is difficult to evaluate how sensitive the results are to the strong functional form assumptions made. All three approaches are in most cases[1] able to distinguish between endogenous and contextual effects only by assuming that one or the other is absent.

The key advantage of the approach presented here is wider applicability than alter-

[1]One exception is Ioannides and Zabel (2002); see Section 2.2.

native methods due to significantly less demanding data requirements. Neither random assignment nor exogenous variation in characteristics are required, and the model can be estimated from standard survey data with either an individual-based or group-based sample design. A related advantage is that these lower data requirements greatly facilitate the estimation of endogenous effects within small and informal groups such as close friends. All of the studies referenced above consider formally defined groups such as classrooms, schools, census tracts, college dorm rooms, etc., in part because their identification strategies are difficult to apply to informal groups. Close friends may be the most influential peers, so a full empirical understanding of social influences in behavior requires the development of tools which can be used with both formal and informal groups. This advantage comes with some associated limitations. First, as with most of the literature, endogenous effects are identified only under the assumption that there are no contextual effects. In addition, the approach described here faces two issues common to structural models: some assumptions are strong and not necessarily testable, and the computational cost and complexity of the estimator is far greater than for OLS or IV methods. Unlike most structural approaches in the literature, however, the approach developed here provides a number of techniques for analyzing the sensitivity of results to critical assumptions.

## 2 The model

The econometric model is based on the standard model of binary choice with social interactions formalized by Brock and Durlauf (2001a), with two substantive differences: the size of the peer group is finite and there may be correlated effects. Both of these features are necessary for many empirical applications.

## 2.1 Preferences and choices

Individuals in the model are organized into a number of non-overlapping peer groups. Groups are indexed by $g$ and individuals are indexed within each group by $i$, so that the pair $(g, i)$ identifies an individual. The size of group $g$ is exogenous and given by $n_g$. Each individual makes a binary choice $y_{gi} \in \{0, 1\}$, and has a utility function $u_{gi}(y_{gi}; \mathbf{y}_g)$ such that:

$$u_{gi}(1; \mathbf{y}_g) - u_{gi}(0; \mathbf{y}_g) = \beta \mathbf{x}_{gi} + \gamma \bar{y}_{gi} + \epsilon_{gi} \tag{1}$$

where $\mathbf{x}_{gi} \equiv \left(1, x_{gi}^1, x_{gi}^2, \ldots, x_{gi}^k\right)'$ is vector of exogenous characteristics which are observable in the data, $\mathbf{y}_g \equiv (y_{g1}, \ldots, y_{gn_g})'$ is the vector of choices made by the members of the group, $\bar{y}_{gi}$ is the average choice made by the other group members ($\bar{y}_{gi} \equiv \frac{1}{n_g - 1} \sum_{j \neq i} y_{gj}$), and $\epsilon_{gi}$ is an exogenous term which is not observed in the data. The parameter $\gamma \geq 0$ is the endogenous social effect; if $\gamma > 0$ an individual's incentive to choose $y_{gi} = 1$ is increasing in the fraction of his or her peers that do so. As with much of the literature, this model assumes that there is no contextual effect. Section 5.2 discusses the implications of relaxing this assumption.

## 2.2 Correlated effects

The three primary sources of correlated effects are simultaneity, nonrandom group selection, and common shocks. The model accounts for correlated effects due to simultaneity by treating peer behavior as an endogenous variable. Correlated effects due to nonrandom group selection and common shocks are introduced into the model[2] by allowing for

[2] An alternative approach for accounting for selection into groups in a structural model would be to formally model the selection process itself. This approach has proved useful in a number of applications such as the selection of workers to firms (Heckman and Sedlacek 1985) and families to neighborhoods (Epple and Sieg 1999), and has been used in the social interactions literature by Ioannides and Zabel (2002). These authors, following a suggestion by Brock and Durlauf (2001b), estimate both endogenous and contextual neighborhood effects in housing demand by estimating neighborhood selection equations and constructing exogenous instruments for neighbors' housing demand using the neighbors selection correction term. While this approach has promise in a number of applications, especially in estimating social interaction effects at the neighborhood or school level, the data requirements for estimating the selection equation are substantial. Identification of the selection equation is facilitated by the presence of fixed and measurable neighborhood/school characteristics, a condition that is

$\epsilon_{gi}$ to be correlated across members of a given peer group. In particular, (using the case $n_g = 3$ as an example) the joint distribution of characteristics across group members is assumed to take the form:

$$\begin{bmatrix} \beta\mathbf{x}_{g1} \\ \beta\mathbf{x}_{g2} \\ \beta\mathbf{x}_{g3} \\ \epsilon_{g1} \\ \epsilon_{g2} \\ \epsilon_{g3} \end{bmatrix} \sim N\left( \begin{bmatrix} \mu \\ \mu \\ \mu \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & \rho_x\sigma^2 & \rho_x\sigma^2 & 0 & 0 & 0 \\ \rho_x\sigma^2 & \sigma^2 & \rho_x\sigma^2 & 0 & 0 & 0 \\ \rho_x\sigma^2 & \rho_x\sigma^2 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \rho_\epsilon & \rho_\epsilon \\ 0 & 0 & 0 & \rho_\epsilon & 1 & \rho_\epsilon \\ 0 & 0 & 0 & \rho_\epsilon & \rho_\epsilon & 1 \end{bmatrix} \right) \quad (2)$$

with the distribution being defined similarly for other values of $n_g$, and where $\rho_x \in \left(-\frac{1}{n_g-1}, 1\right)$ and $\rho_\epsilon \in \left(-\frac{1}{n_g-1}, 1\right)$ to ensure that the covariance matrix is positive definite.

In addition to the functional form restriction of joint normality, equation (2) places several substantive restrictions on the model. Two of these restrictions are innocuous: that the distribution is symmetric (since the ordering of group members is arbitrary), and that the utility function has been normalized so that $\epsilon_{gi}$ has mean zero and unit variance. A more substantive restriction is that, as in the standard probit model, the observable and unobservable terms are uncorrelated, i.e., $cov(\epsilon_{gi}, \beta\mathbf{x}_{gi}) = 0$. Potentially more controversially but in the same spirit, it is also assumed that there is no correlation between one group member's observables and the unobservables of the other group members, i.e., $cov(\epsilon_{gi}, \beta\mathbf{x}_{gj}) = 0$ for $i \neq j$. Section 5.2 briefly discusses the implications of nonzero correlation between observables and unobservables.

## 2.3 Equilibrium

Given the preferences of each agent, let $Y_g$ be the set of pure strategy Nash equilibria of the normal form game defined by players $i \in \{1, \ldots, n_g\}$, strategy space $\{0,1\}^{n_g}$ and

not always met in a given application. The approach developed in this paper is complementary to approaches built on more formal models of group selection, and has less demanding data requirements.

payoff functions that satisfy equation (1):

$$Y_g \equiv \{\mathbf{y} \in \{0,1\}^{n_g} : y_i = 1 \Leftrightarrow u_{gi}(1;\mathbf{y}) - u_{gi}(0;\mathbf{y}) > 0 \quad \forall i \in \{1, 2, \ldots, n_g\}\} \tag{3}$$

The properties of Nash equilibria for this model follow closely from Milgrom and Roberts' (1990) results on supermodular games. In particular, for any preference profile, the set of pure strategy Nash equilibria has a minimal and maximal value, and all rationalizable strategy profiles, all (pure or mixed) Nash equilibria, and all correlated equilibria lie in the interval $[\min(Y_g), \max(Y_g)]$. Equilibrium is unique for almost all preference profiles if there is no endogenous effect ($\gamma = 0$), and is nonunique for a positive (probability) measure of preference profiles if there is an endogenous effect ($\gamma > 0$) (Krauth 2001). Multiplicity of equilibria complicates estimation, as uniqueness of the likelihood function for equilibrium behavior requires the imposition of an equilibrium selection rule.

A selection rule is simply a function $\text{sel}(\mathbf{y}, Y)$ which assigns a probability to each pure strategy Nash equilibrium:

$$\text{sel}(\mathbf{y}, Y) \equiv \Pr(\mathbf{y}_g = \mathbf{y} | Y_g = Y) \tag{4}$$

In order to describe a well defined probability distribution, $s$ must obey the constraints $\text{sel}(\mathbf{y}, Y) \geq 0$ and $\sum_{\mathbf{y}} \text{sel}(\mathbf{y}, Y) = 1$. In addition, the requirement that only pure strategy Nash equilibria are selected implies the constraint $(\mathbf{y} \notin Y) \Rightarrow (\text{sel}(\mathbf{y}, Y) = 0)$. Imposing a selection rule pins down a unique likelihood function and, if correctly specified, produces consistent point estimates of parameters. This paper will consider three specific selection rules:

$$\text{Low-activity equilibrium:} \qquad \text{sel}(\min(Y), Y) = 1 \tag{5}$$

$$\text{High-activity equilibrium:} \qquad \text{sel}(\max(Y), Y) = 1 \tag{6}$$

$$\text{Random equilibrium:} \qquad \mathrm{sel}(\mathbf{y}, Y) = \begin{cases} \frac{1}{\#Y} & \text{if } \mathbf{y} \in Y \\ 0 & \text{if } \mathbf{y} \notin Y \end{cases} \tag{7}$$

More complex rules are also possible, and may be appropriate to a particular application.

Although a selection rule must be imposed to achieve point identification, selection rule free estimation methods are also available which identify bounds on the model parameters. Under some conditions, these bounds are informative. Section 4.2 describes selection rule free estimation of the model.

## 2.4 Identifying restrictions on the correlated effects

Although the model as specified is formally identified, there is no obvious nonparametric means of distinguishing between correlated effects ($\rho_\epsilon$) and endogenous effects ($\gamma$). This section outlines some plausible restrictions which facilitate nonparametric identification. Section 3.3 provides a heuristic argument for nonparametric identification under such restrictions.

### 2.4.1 Baseline restriction: Equal correlation

One plausible approach, which is used as the baseline identifying assumption in this paper, is to use information in the data on $\rho_x$ to provide a reasonable guess for $\rho_\epsilon$. In a sense, this is already done informally: the reason why applied researchers are particularly concerned about positive between-peer correlation in unobservables (which, after all, is just another species of omitted variables bias) is that there is often ample evidence of positive correlation in observable characteristics among peers. In general, a restriction using information on $\rho_x$ to restrict $\rho_\epsilon$ would take the form $\rho_\epsilon = f(\rho_x)$ for some known function $f$.

In particular, the baseline restriction for this paper is that the two correlation coefficients are *equal*:

$$\rho_\epsilon = \rho_x \equiv \rho \tag{8}$$

The equal correlation assumption and its justification are in the spirit of Altonji, Elder, and Taber's (2000, AET) work on the effects of Catholic schools. The literature on Catholic school effects faces similar issues to the social interaction effects literature: although students in Catholic schools experience better outcomes on average than observationally similar students in public schools, it is difficult to distinguish the effect of Catholic school from the effect of unobserved characteristics which lead to both better outcomes and increased probability of selecting a Catholic school. Their partial solution to this identification problem is to estimate a selection equation and outcome equation, then use the correlation between the fitted values in the two equations as a proxy for the correlation between the unobserved terms.

The current setting is somewhat different from that faced by AET, but much of their intuition and argument can be adapted. Suppose that the incremental utility previously defined in equation (1) can alternatively be written as a linear function of a large set of relevant variables:

$$u_{gi}(1;\mathbf{y}_g) - u_{gi}(0;\mathbf{y}_g) = \Gamma Z_{gi} + \gamma \bar{y}_{gi} \tag{9}$$

where $Z_{gi}$ is the complete vector of individual $i$'s relevant characteristics and $\Gamma$ is a vector of coefficients. Now suppose that we randomly divide this large set of relevant variables into an "observed" subset $Z_{gi}^x$ and an "unobserved" subset $Z_{gi}^\epsilon$. Let $\Gamma^x$ and $\Gamma^\epsilon$ be the corresponding subvectors of $\Gamma$. Now suppose that we calculate the within-peer-group correlations $\rho_x \equiv corr(\Gamma^x Z_{gi}^x, \Gamma^x Z_{gj}^x)$ and $\rho_\epsilon \equiv corr(\Gamma^\epsilon Z_{gi}^\epsilon, \Gamma^\epsilon Z_{gj}^\epsilon)$. Since the partition is random

$$E(\rho_x) = E(\rho_\epsilon) \tag{10}$$

where the expectations are taken across the distribution of possible random partitions.

The use of this argument to justify assuming $\rho_x = \rho_\epsilon$ depends on two key elements. First, the observed variables must be a random subset of the relevant variables, in the sense that all relevant variables are observed with equal probability. Alternatively, one might guess that characteristics more likely to be observed by an econometrician are also more likely to be observed by those forming peer groups, in which case $\rho_x$ is actually an upper bound for $\rho_\epsilon$. In that case, a lower bound can be placed on $\gamma$ using the methodology described in Section 4.1. Second, equation (10) only implies that the *expected* correlation is the same for the observable and unobservable components, when expectations are taken over the set of random subsets of variables. Equation (8) asserts equality in the *realized* correlation for the particular set of variables observed in a given data set. All else being equal, the appropriateness of this assumption will be greater in cases where the number of explanatory variables is greater.

### 2.4.2 Alternative point restrictions

Alternatively, the model parameters are nonparametrically identified under any other point restriction on either $\rho_\epsilon$ or $\gamma$, including:

$$\rho_\epsilon = \rho^* \quad \text{for some known } \rho^* \tag{11}$$

$$\gamma = \gamma^* \quad \text{for some known } \gamma \tag{12}$$

$$\rho_\epsilon = f(\rho_x) \quad \text{for some known function } f(.) \tag{13}$$

For example, in settings where the peer group is assigned through a random mechanism (Sacerdote 2001) it may be reasonable to assume that $\rho_\epsilon = 0$. Alternatively, a researcher may estimate $\rho_\epsilon$ under the restriction $\gamma = 0$. The result will indicate how large the correlated effect must be to explain the data in the absence of an endogenous effect; if the implied correlation in unobservables is implausibly large (particularly in light of the estimated correlation in observables $\rho_x$), this result could be interpreted as evidence in favor of an endogenous effect.

### 2.4.3 Interval restrictions

In most applications, the values of $\rho_\epsilon$ and $\gamma$ are not known in advance. However, one may have reasonable confidence in a particular upper or lower bound on $\rho_\epsilon$. For example, one may be confident that peer group members are at least as similar as would occur under random group assignment (which implies $\rho_\epsilon \geq 0$) or one may be confident that the correlation in observable characteristics provides an upper bound on the correlation in unobservables (which implies $\rho_\epsilon \leq \rho_x$). Section 4.1 describes a method for constructing consistent bounds on $\gamma$ given bounds imposed on $\rho_\epsilon$. In some cases the resulting bounds on $\gamma$ can be informative even for very conservative restrictions on $\rho_\epsilon$.

# 3 Estimation

The econometric model is defined by the utility function (1), the joint probability distribution of the exogenous variables (2), the definition of equilibrium (3), and an equilibrium selection rule, of which (5)-(7) are examples. In practice, the additional restriction (8), or some alternative restriction on $\rho_\epsilon$, will also be imposed.

The model can be estimated using data with either an individual-based or group-based sampling design, though many details of the estimation method vary with the type of data. An individual-based sample (Bertrand, Luttmer and Mullainathan 2000, Evans et al. 1992, Katz et al. 2001) is just a standard random sample of individuals, with data on peer behavior either reported directly by the respondents or derived from some separate aggregate[3] data source such as Census tract data. Group-based samples have been used frequently in the social interactions literature (Gaviria and Raphael 2001, Hoxby 2000, Kremer and Levy 2001, Sacerdote 2001), and are constructed by sampling a number of individuals within each of a set of randomly sampled groups. Data on peer behavior is derived from the peers' own self-reports, provided that the group identification of

[3] One practical constraint on the estimator developed in this section is the size of peer groups. In principle, the estimator described here can be used with peer groups of any size, and that size can vary across the groups within a data set. In practice, the computational cost of the estimator increases with the size of the peer group, which may make simulation-based estimation of this model impractical for extremely large peer groups (Census tracts, cities, etc.). Fortunately, the model can still be estimated, and the estimation method is actually much simpler. Krauth (2004) shows that as the group size increases, a close approximation of the likelihood function can be calculated directly without use of simulation. Standard maximum likelihood methods can then be applied.

each individual is provided in the data. In addition to providing information on the binary choice of each respondent and the average choice in each peer group, the data set must include at least some of each respondent's relevant background characteristics. In addition to these two main cases, the model and estimation methods described here could also be adapted in a conceptually straightforward manner to handle richer data with detailed social network information or multiple levels of reference groups, subject to a few application-driven modeling decisions on the covariance structure of the exogenous variables.

## 3.1 ML and SML estimation

Consider a data set of $N$ individuals (if an individual-based sample) or $N$ groups (if a group-based sample), and index observations by $g = 1, \ldots, N$. In a group-based sample, both $\mathbf{x}_{gi}$ and $y_{gi}$ are observed for all group members. In an individual-based sample, $\mathbf{x}_{gi}$ and $y_{gi}$ are observed for the respondent only, as well as the proportion or number of other group members for whom $y_{gi} = 1$.

Let $\mathbf{y}_g$ be defined as in Section 2. Note that the index of an individual within the group is arbitrary. For ease of notation, let the respondent[4] in an individual-based sample be indexed as group member #1. Let $\mathbf{X}_g$ indicate the vector or matrix of observed explanatory variables:

$$\mathbf{X}_g \equiv \begin{cases} \mathbf{x}_{g1} & \text{if an individual-based sample} \\ (\mathbf{x}_{g1}, \mathbf{x}_{g2}, \ldots, \mathbf{x}_{gn_g})' & \text{if a group-based sample} \end{cases}$$

The data set is thus $\{(\mathbf{X}_g, \mathbf{y}_g)\}_{g=1}^{N}$.

Let $\theta_0 = (\beta, \gamma, \rho_x, \rho_\epsilon, \mu, \sigma)$ be the true parameter vector, let $\Theta \subset R^{k+6}$ be the feasible

[4]With a true random sample and peer groups of fixed size, the probability that two randomly sampled individuals will be in the same peer group goes to zero as the population size goes to infinity. In practice, there may be applications where deviations from true random sampling leads to multiple observations within a given group. This case can be treated as just a special case of the group-based sample with exogenously censored elements of $\mathbf{x}_g$.

parameter space[5] for $\theta_0$, and let $\theta$ be an arbitrary element of $\Theta$. The model defined in Section 2 implies that, for any $\theta$, it is possible to calculate the conditional probability of observing a particular outcome. The maximum likelihood estimator of $\theta_0$ is then given by:

$$\begin{aligned} \theta^{ML} &\equiv \arg\max_{\theta \in \Theta} \sum_{g=1}^{N} \ln \Pr(\mathbf{y}_g, \mathbf{X}_g; \theta) \\ &= \arg\max_{\theta \in \Theta} \sum_{g=1}^{N} (\ln \Pr(\mathbf{y}_g | \mathbf{X}_g; \theta) + \ln \Pr(\mathbf{X}_g; \theta)) \end{aligned}$$

While $\Pr(\mathbf{X}_g; \theta)$ can be calculated analytically from the multivariate normal PDF, direct calculation of $\Pr(\mathbf{y}_g, \mathbf{X}_g; \theta)$ is infeasible as it requires the evaluation of a complex multidimensional integral. As a result, simulation must be used to estimate the likelihood function.

Let $\{P_g^s(\theta)\}_{s=1}^S$ be an unbiased simulator for $\Pr(\mathbf{y}_g | \mathbf{X}_g; \theta)$, i.e., a sequence of $S$ independent random variables such that $E(P_g^s(\theta)) = \Pr(\mathbf{y}_g | \mathbf{X}_g; \theta)$. The simulated maximum likelihood (SML) estimator of $\theta_0$ is then defined as:

$$\theta^{SML} \equiv \arg\max_{\theta \in \Theta} \sum_{g=1}^{N} \left( \ln \left( \frac{1}{S} \sum_{s=1}^{S} P_g^s(\theta) \right) + \ln \Pr(\mathbf{X}_g; \theta) \right)$$

**Proposition 1 (Consistency of SML estimator)** *If $\theta_0$ is identified, then $\theta^{SML}$ is a consistent estimator of $\theta_0$*

Proof: If $\theta_0$ is identified then $\theta^{ML}$ is a consistent estimator of $\theta_0$. By the weak law of large numbers, $\frac{1}{S} \sum_{s=1}^{S} P_g^s(\theta) \xrightarrow{p} E(P_g^s(\theta)) = \Pr(\mathbf{y}_g | \mathbf{X}_g; \theta)$. By Proposition 3.1 in Gouriéroux and Monfort (1996), this implies that $\theta^{SML}$ is a consistent estimator of $\theta_0$. □

[5]Note that $\mu$ and $\sigma$ are not free parameters, as $\mu = E(\beta \mathbf{x})$ and $\sigma^2 = Var(\beta \mathbf{x})$. As a result, the implementation of this estimator saves significant computational time by using a two step (LIML) estimator in which $\mu$ is replaced by $\beta \bar{x}$ and $\sigma$ is replaced by $\beta \Sigma_x \beta'$ where $\Sigma_x$ is just the standard variance/covariance matrix of the $x$'s.

## 3.2 Simulation methodology

This section outlines two methods for generating $P_g^s(\theta)$, both of which are based on the Geweke, Hajivassiliou, and Keane (GHK) algorithm. The GHK simulator is described by Hajivassiliou, McFadden, and Ruud (1996), and was found to be the most robust and accurate among numerous techniques compared by those researchers.

Let $\mathbf{c} \sim N(M, \Sigma)$ be a random vector of length $n$, let $\chi$ be any subset of $R^n$ which can be expressed as the union of a finite collection of disjoint rectangles[6], and let $\xi$ be a vector of $n$ independent $U(0,1)$ random variables. The GHK simulator is a function $GHK$ such that $E(GHK(M, \Sigma, \chi, \xi)) = \Pr(\mathbf{c} \in \chi)$. The typical application uses a computer to generate an independent pseudorandom sequence $\{\xi^s\}_{s=1}^S$, then uses $\frac{1}{S}\sum_{s=1}^S GHK(M, \Sigma, \chi, \xi^s)$. to estimate $\Pr(\mathbf{c} \in \chi)$. In the course of calculating $GHK(.)$, the GHK algorithm also calculates a random vector $ghk(M, \Sigma, \chi, \xi) \sim \mathbf{c}|\mathbf{c} \in \chi$. Proposition 2 below derives two methods for using these two functions to estimate $\Pr(\mathbf{y}_g|\mathbf{X}_g; \theta)$.

**Proposition 2 (Simulators for** $\Pr(\mathbf{y}_g)$**)** *Let* $\mathbf{c}_g \equiv (c_{g1}, c_{g2}, \ldots, c_{gn_g})'$ *where:*

$$c_{gi} \equiv \begin{cases} -\frac{\beta \mathbf{x}_{gi} + \epsilon_{gi}}{\gamma} & \text{if } \gamma > 0 \\ \beta \mathbf{x}_{gi} + \epsilon_{gi} & \text{if } \gamma = 0 \end{cases}$$

*Let* $Y(\mathbf{c})$ *be the set of pure strategy Nash equilibria when* $\mathbf{c}_g = \mathbf{c}$. *Then:*

1. *The set* $c(\mathbf{y}) \equiv \{\mathbf{c} : \mathbf{y} \in Y(\mathbf{c})\}$ *is a rectangle.*

2. *The set* $C(Y) \equiv \{\mathbf{c} : Y = Y(\mathbf{c})\}$ *can be expressed as the union of a finite collection of disjoint rectangles.*

3. *The conditional distribution* $\mathbf{c}_g|\mathbf{X}_g$ *is multivariate normal with mean* $M_g$ *and co-*

[6] A rectangle is a set which can be defined as $\chi = \{\mathbf{X} \in R^n : a_\chi \leq \mathbf{X} \leq b_\chi\}$ for some pair of vectors $a_\chi, b_\chi \in (R \cup \{-\infty, \infty\})^n$

*variance matrix $\Sigma_g$, where:*

$$M_g \equiv - \begin{bmatrix} \frac{\beta x_{g1}}{\gamma} \\ \frac{\mu+\rho_x(\beta x_{g1}-\mu)}{\gamma} \\ \frac{\mu+\rho_x(\beta x_{g1}-\mu)}{\gamma} \end{bmatrix}, \quad \Sigma_g \equiv \begin{bmatrix} \frac{1}{\gamma^2} & \frac{\rho_\epsilon}{\gamma^2} & \frac{\rho_\epsilon}{\gamma^2} \\ \frac{\rho_\epsilon}{\gamma^2} & \frac{1+(1-\rho_x^2)\sigma^2}{\gamma^2} & \frac{\rho_\epsilon+(\rho_x-\rho_x^2)\sigma^2}{\gamma^2} \\ \frac{\rho_\epsilon}{\gamma^2} & \frac{\rho_\epsilon+(\rho_x-\rho_x^2)\sigma^2}{\gamma^2} & \frac{1+(1-\rho_x^2)\sigma^2}{\gamma^2} \end{bmatrix} \tag{14}$$

*for an individual based sample, and:*

$$M_g \equiv - \begin{bmatrix} \frac{\beta x_{g1}}{\gamma} \\ \frac{\beta x_{g2}}{\gamma} \\ \frac{\beta x_{g3}}{\gamma} \end{bmatrix}, \quad \Sigma_g \equiv \begin{bmatrix} \frac{1}{\gamma^2} & \frac{\rho_\epsilon}{\gamma^2} & \frac{\rho_\epsilon}{\gamma^2} \\ \frac{\rho_\epsilon}{\gamma^2} & \frac{1}{\gamma^2} & \frac{\rho_\epsilon}{\gamma^2} \\ \frac{\rho_\epsilon}{\gamma^2} & \frac{\rho_\epsilon}{\gamma^2} & \frac{1}{\gamma^2} \end{bmatrix} \tag{15}$$

*for a group-based sample.*

4. *Let $\xi^s$ be a vector of $n_g$ independent random variables from the standard uniform distribution. Finally, let the "Full GHK simulator" for observation $g$ be given by:*

$$P_g^s(\theta) \equiv \sum_Y sel(\mathbf{y}_g, Y) GHK(M_g, \Sigma_g, C(Y), \xi^s) \tag{16}$$

*or let the "GHK-frequency hybrid simulator" be given by:*

$$P_g^s(\theta) \equiv sel(\mathbf{y}_g, Y(ghk(M_g, \Sigma_g, c(\mathbf{y}_g), \xi^s)) GHK(M_g, \Sigma_g, c(\mathbf{y}_g), \xi^s) \tag{17}$$

*Both of these simulators are consistent, i.e., $\frac{1}{S}\sum_{s=1}^{S} P_g^s(\theta) \xrightarrow{p} \Pr(\mathbf{y}_g|\mathbf{X}_g;\theta)$.*

Proof: A constructive proof for each part of the proposition is provided in the appendix.

Proposition 2 defines two alternative methods for estimating $\Pr(\mathbf{y}_g|\mathbf{X}_g;\theta)$. The full GHK simulator calculates the set $C(Y)$ in terms of a finite union of disjoint rectangles, uses the GHK simulator to estimate the probability of each rectangle, and then adds up (with weights given by $\text{sel}(\mathbf{y}_g, Y)$) across all $Y$. It inherits the GHK simulator's useful properties of low variance, as well as continuity and differentiability. Its primary drawback is that the number of rectangles in $C(Y)$, and thus the computational cost, grows rapidly in $n_g$. The GHK-frequency hybrid uses the GHK simulator on the single rectangle $c(\mathbf{y}_g)$ to accurately estimate the probability that $\mathbf{y}_g$ is a Nash equilibrium. As a side effect, the GHK simulator generates a random $\mathbf{c}_g$ such that $\mathbf{y}_g$ is a Nash equilibrium; that random $\mathbf{c}_g$ is used for a simple frequency simulator to estimate the probability that $\mathbf{y}_g$ will actually be observed given that it is a Nash equilibrium. The GHK-frequency hybrid simulator has the advantage over the full GHK simulator that its computational cost does not grow rapidly in $n_g$, because only one rectangle probability is calculated per observation. Its primary disadvantages are that it has a somewhat higher variance than the full GHK simulator and is discontinuous in $\theta$. This discontinuity complicates maximization of the resulting likelihood function, as standard methods do not work well with a discontinuous function (Gouriéroux and Monfort 1996, p. 96). Based on these characteristics, it is recommended that the full GHK simulator be used if peer groups are relatively small ($n_g \leq 8$), with the GHK-frequency hybrid used for data sets with larger groups. For purposes of demonstrating both methods, the Monte Carlo results in Section 5.2 are calculated using the full GHK simulator for individual-based samples and the GHK-frequency hybrid for group-based samples.

The computer code implementing the estimation method is available from the author. The implementation of the GHK simulator is adapted from Vassilis Hajivassiliou's GAUSS code, and optimization is done by either a BFGS-Brent optimization routine written by Bo Honoré and Ekaterini Kyriazidou (for the full GHK simulator) or a simulated annealing routine written by William Goffe (for the GHK-frequency hybrid simulator). As is generally the case in simulation-based estimation, the matrix of pseudorandom numbers used in the simulator is kept constant through the entire estimation procedure. In order to minimize simulation error, the pseudorandom numbers are derived from randomized Halton sequences (Train 2002, Bhat 2003).

## 3.3 Identification

This section provides a heuristic argument that the model is nonparametrically identified. Nonparametric identification is a key issue in applications of structural models such as the one presented here. If the structural parameters are identified strictly as a result of arbitrary functional form assumptions rather than through economically substantive and justifiable restrictions, the model has little empirical usefulness.

The fundamental source of identification in this case is the treatment of both the respondent's choice ($y_{gi}$) and that of his or her peers ($\bar{y}_{gi}$) as endogenous. Consider two conditional expectation functions, $E(y_{gi}|\mathbf{x}_{gi}, \bar{y}_{gi})$ and $E(\bar{y}_{gi}|\mathbf{x}_{gi}, y_{gi})$, which can in principle be estimated from the data. First, the relationship between an individual's characteristics and his or her choice, i.e., $\frac{\partial E(y_{gi}|\mathbf{x}_{gi},\bar{y}_{gi})}{\partial \mathbf{x}_{gi}}$ is increasing in $\beta$, thus identifying that vector of parameters. The relationship between an individual's choice and the average choice of his or her peers, i.e., $\frac{\partial E(y_{gi}|\mathbf{x}_{gi},\bar{y}_{gi})}{\partial \bar{y}_{gi}}$, is increasing in both $\rho_\epsilon$ and $\gamma$. This is exactly the identification problem described by Manski (1993), and is the reason why one or another of the restrictions discussed in Section 2.4 is needed to provide nonparametric identification. Once such a restriction is made, $\gamma$ is identified. With group-based data, the correlation in observables $\rho_x$ is identified directly by the corresponding sample moment.

With individual-based data, the identification of $\rho_x$ is more complex. It is identified from the relationship between an individual's characteristics and the average choice of his or her friends $\beta\frac{\partial E(\bar{y}_{gi}|\mathbf{x}_{gi},y_{gi})}{\partial \mathbf{x}_{gi}}$, which is increasing in $\rho_x$. To gain intuition for this, consider the following numerical example and estimation by a simple indirect inference scheme. The example features a simulated data set of 100,000 with the same parameter settings as in the baseline model of Section 5. To facilitate comparison across different simulations, the matrix of pseudorandom variables used to generate the simulations is fixed across trials. Based on a given Monte Carlo sample, define two reduced form coefficients: let $\hat{\psi}$ be the coefficient on $\bar{y}$ in a naive probit regression of $y$ on $\mathbf{x}$ and $\bar{y}$, and let $\hat{\omega}$ be the coefficient on $y\beta\mathbf{x}$ from an OLS regression of $\bar{y}$ on $(y, y\beta\mathbf{x}, (1-y)\beta\mathbf{x})$. Since these are random variables it is also convenient to define $\psi(\rho,\gamma) \equiv E(\hat{\psi};\rho,\gamma)$ and $\omega(\rho,\gamma) \equiv E(\hat{\omega};\rho,\gamma)$. Now, suppose that we were to attempt to recover $\rho$ and $\gamma$ from the reduced-form coefficients $\hat{\psi}$ and $\hat{\omega}$. Based on the simulation described above, I estimate that $\psi(0.25, 0.5) \approx 1.511$ and $\omega(0.25, 0.5) \approx 0.0132$. Also using the simulated data, but with different parameter values, Figure 1 shows the set of all $(\rho, \gamma)$ pairs such that $\psi(\rho,\gamma) \approx 1.511$ (the downward-sloping curve) and the set of all pairs such that $\omega(\rho,\gamma) \approx 0.0132$ (the upward-sloping curve). These two curves intersect at only one point: (0.25,0.5). Not only does this explanation provide some insight into the usefulness of the common-correlation assumption, it suggests an explanation for another apparent feature of the estimator. The Monte Carlo experiments suggest that the accuracy of $\hat{\gamma}$ is closely related to the explanatory power of the exogenous variables, $var(\beta\mathbf{x})$. In the context of Figure 1, this can be explained by high variance in $\hat{\omega}$ when $var(\beta\mathbf{x})$ is low.

# 4 Extensions

## 4.1 Estimation with alternative restrictions on $\rho_\epsilon$

Although the baseline model's assumption of equal correlation in observables and unobservables is reasonable in some applications, an empirical researcher may wish to estimate under alternative restrictions. For point restrictions of the form (11), (12), or (13), estimation is simply a matter of optimizing the simulated log-likelihood function with the appropriate constraint substituted for the equal-correlation restriction (8).

It is also possible, as suggested in Section 2 to place bounds on $\gamma$ under interval restrictions on $\rho_\epsilon$. The method for doing this is straightforward. Let the function $\hat{\gamma}(\rho^*)$ be defined as the ML estimate of $\gamma$ under the restriction that $\rho_\epsilon = \rho^*$. Two things should be noted about $\hat{\gamma}(.)$. First, $\hat{\gamma}(\rho_\epsilon) \xrightarrow{p} \gamma$, where $\rho_\epsilon$ and $\gamma$ are the true parameter values. Second, because the log-likelihood function is continuous in $\theta$, the Maximum

Theorem implies that $\hat{\gamma}(.)$ is also continuous. This continuity suggests that a researcher can calculate $\hat{\gamma}(.)$ at a finite number of points and interpolate between these points.

Now, suppose that one is willing to place an interval restriction on $\rho$ and would like to find corresponding bounds on $\gamma$. The bounds can be defined as follows:

$$\rho_\epsilon \in [\rho^L, \rho^H] \Rightarrow \hat{\gamma} \in \left[ \min_{\rho \in [\rho^L, \rho^H]} \hat{\gamma}(\rho), \max_{\rho \in [\rho^L, \rho^H]} \hat{\gamma}(\rho) \right] \tag{18}$$

In practice, it is simpler to report or graph $\hat{\gamma}(.)$ and allow the reader to choose a reasonable interval restriction and construct the bounds. The application in Section 6 includes an example.

## 4.2 Estimation without an equilibrium selection rule

It is also possible in some cases to learn something about parameter values without imposing an equilibrium selection rule. There is a small literature (Jovanovic 1989, Tamer 2002b, Tamer 2002a) that explores identification in models with multiple equilibria. The key insight in the literature is that models with multiple equilibria imply interval rather than point restrictions on conditional probabilities. Specifically, for an arbitrary choice vector $\mathbf{y}$, any selection rule must obey the condition:

$$\Pr(\{\mathbf{y}\} = Y_g) \leq \Pr(\mathbf{y}_g = \mathbf{y}) \leq \Pr(\mathbf{y} \in Y_g) \tag{19}$$

In other words, the probability of observing a particular outcome $\mathbf{y}$ will be at least as large as the probability that it is the unique Nash equilibrium and no larger than the probability that it is a Nash equilibrium. As Jovanovic notes, equation (19) defines a family of likelihood functions, which includes the likelihood functions defined by (5)-(7). The set of parameter vectors which maximizes one or more of these likelihood functions will by construction contain the ML estimate based on the unknown correct likelihood function. In other words, although an equilibrium selection rule must usually be imposed to achieve point identification[7] of parameters, consistent bounds on parameters may be constructed without imposing a selection rule.

The approach pursued in this paper is the "likelihood bounds" approach discussed by Tamer (2002a). Let $H_g^s(\theta)$ be a random variable with the property that $E(H_g^s(\theta) = \Pr(\mathbf{y}_g \in Y_g|\mathbf{x}_g;\theta)$, and let $L_g^s(\theta)$ be a random variable with the property that $E(L_g^s(\theta)) = \Pr(\{\mathbf{y}_g\} = Y_g|\mathbf{x}_g;\theta)$. The simulators used here are:

$$H_g^s(\theta) \equiv GHK(M_g, \Sigma_g, c(\mathbf{y}_g), \xi^s) \tag{20}$$

and either (full GHK simulator):

$$L_g^s(\theta) \equiv GHK(M_g, \Sigma_g, C(\{\mathbf{y}_g\}), \xi^s) \tag{21}$$

or (GHK-frequency hybrid):

$$L_g^s(\theta) \equiv I\left[Y(ghk(M_g, \Sigma_g, c(\mathbf{y}_g), \xi^s)) = \{\mathbf{y}_g\}\right] GHK(M_g, \Sigma_g, c(\mathbf{y}_g), \xi^s) \tag{22}$$

where $I[.]$ is the indicator function, and $M_g$,$\Sigma_g$,$\xi^s$, $c(.)$, and $C(.)$ are defined as in Proposition 2.

These estimated probabilities provide bounds on the set of likelihood functions that are consistent with some equilibrium selection rule. Let $\ell^*$ be the maximum of all the lower bounds:

$$\ell^* \equiv \max_{\theta \in \Theta} \sum_{g=1}^{N} \left( \ln \left( \frac{1}{S} \sum_{s=1}^{S} L_g^s(\theta) \right) + \ln \Pr(\mathbf{X}_g;\theta) \right) \tag{23}$$

Regardless of the selection rule, the maximum of the true likelihood function must be at least as great as $\ell^*$. This implies that the true maximum likelihood estimate $\theta^{ML}$ is

[7]Tamer (2002b) shows that one can sometimes redefine outcomes in such a way as to get point restrictions on probabilities in this model without imposing a selection rule. However, his solution only works for the two player case.

contained in the set $V$ defined as:

$$V \equiv \left\{ \theta \in \Theta : \sum_{g=1}^{N} \left( \ln \left( \frac{1}{S} \sum_{s=1}^{S} H_g^s(\theta) \right) + \ln \Pr(\mathbf{X}_g; \theta) \right) \geq \ell^* \right\} \tag{24}$$

As the Monte Carlo results in Section 5.2 will show, $V$ is sometimes but not always small enough to provide informative bounds on parameter values. As Tamer (2002a) notes, $V$ does not provide sharp bounds on the parameter values because there are many functions between the upper and lower bounds that are not valid log-likelihood functions.

Figure 2 shows this method graphically for three simulated data sets with different parameter values. The shaded area is the area of allowable likelihood functions constructed using $H_g^s(\theta)$ and $L_g^s(\theta)$. The horizontal dotted line marks the maximum of the lower bounds $\ell^*$. The range of $\gamma$ values in the set $V$ is defined by the dark line. As the figure shows, the upper and lower bounds are identical for $\gamma = 0$ (because equilibrium is almost always unique), and diverge as $\gamma$ increases. As a result, $V$ provides much narrower bounds on $\hat{\gamma}$ when the true value of $\gamma$ is lower.

## 4.3 Estimation with inconsistent reporting

In individual-based samples where the behavior of peers is reported by the respondent, one issue that may appear is inconsistent reporting. For example, in the application described in Section 6, the percentage of young people who self-report that they smoke cigarettes is substantially lower than the average percentage of their four best friends that they report as smokers. Because both the respondents and their friends are drawn from the same population, this implies that they are either underreporting their own smoking, overreporting their friends' smoking, or both. This can be handled by extending the model to incorporate inaccurate reporting. This section describes a simple example of such an extension; richer models of reporting could be defined as needed for applications.

The model of inconsistent reporting is based on three assumptions: individuals truthfully report the behavior of their peers $\bar{y}_{gi}$, truthfully report their own behavior when

$y_{gi} = 0$, and truthfully report their own behavior with some exogenous probability $p_r$ when $y_{gi} = 1$ . Let $r_{gi}$ be a binary variable indicating the behavior a person would self-report in a survey. We suppose that

$$\Pr(r_{gi} = 1) = \begin{cases} 0 & \text{if } y_{gi} = 0 \\ p_r & \text{if } y_{gi} = 1 \text{ and } i = 1 \\ 1 & \text{if } y_{gi} = 1 \text{ and } i \neq 1 \end{cases} \tag{25}$$

Given this, we simply redefine the observed outcome variable as $\mathbf{r}_g = (r_{g1}, r_{g2}, \ldots, r_{gn_g})'$ and find the parameter values which maximize a log-likelihood based on $\Pr(\mathbf{r}_g, \mathbf{X}_g; \theta, p_r)$ rather than $\Pr(\mathbf{y}_g, \mathbf{X}_g; \theta)$. Although $p_r$ can be estimated as part of the likelihood function, it can also be estimated directly. Since $p_r = E(r_{g1})/E(\bar{r}_{g1})$, a natural estimator for $p_r$ is just the ratio of the corresponding sample averages.

### 4.4 Estimation with aggregate variables

There are also many applications where some of the explanatory variables are aggregate and apply to all group members, for example, prices, state or year fixed effects, school characteristics, etc. These can be incorporated into the model by modifying the utility function to include a vector of aggregate variables $z_g$:

$$u_{gi}(1; \mathbf{y}_g) - u_{gi}(0; \mathbf{y}_g) = \lambda z_g + \beta \mathbf{x}_{gi} + \gamma \bar{y}_{gi} + \epsilon_{gi} \tag{26}$$

and proceeding as usual.

## 5 Monte Carlo experiments

This section applies the estimator developed in Section 3 to a series of simulated data sets. The results provide an insight into the statistical performance of the estimator.

## 5.1 Overview of the experiments

Table 1 reports a selection of the Monte Carlo results. The baseline experiment has observations on 1,000 individuals, organized into 5-member peer groups. There is a single $x$ variable with a $N(0,1)$ marginal distribution. The coefficient on $x$ is $\beta_1 = 1$ and the intercept is $\beta_0 = 0$. Both the actual selection rule and the selection rule assumed in estimating the model correspond to equation (5), the "low-activity" equilibrium. The correlated and endogenous effects take on one of three combinations: an endogenous effect with no correlated effect ($\gamma = 1, \rho = 0$), a correlated effect with no endogenous effect ($\gamma = 0, \rho = 0.25$), and both effects ($\gamma = 0.5, \rho = 0.25$). Simulations are generated with 1,000 observations, and both individual-based and group-based samples are generated. Once generated, the simulated data sets are used to estimate (1) a "naive" probit model, in which peer behavior is treated as an exogenous explanatory variable, from the individual-based sample, (2) the structural (SML) model estimated from the individual-based sample, and (3) the structural (SML) model from the group-based sample. Table 1 reports the sample mean of key parameter estimates over 100 Monte Carlo samples, and in some cases the sample standard deviation.

## 5.2 Results

**Baseline:** Rows 1-3 in Table 1 report the results from the baseline experiment. The naive probit dramatically overestimates the true peer effect in all three cases. Even when $\rho = 0.00$, simultaneity produces an upwards bias of approximately 50%. In contrast, the structural estimator eliminates both types of correlated effect and comes very close to the true parameter values on average. The only sign of nontrivial bias in the structural estimates is for the case of $\gamma = 0$. Because $\gamma = 0$ is on the boundary of the allowed parameter space, estimates of $\gamma$ will have some upwards bias in any finite sample. This issue also complicates hypothesis testing on the null of no peer effect. In all cases, the SML estimator clearly dominates the naive estimator.

**Alternative assumptions on x:** Rows 4-6 provide information on how the properties of the explanatory variables matter. Row 4 shows results from a variation on the baseline experiment in which the explanatory variables have low explanatory power for the outcome, i.e., $\beta_1 = 0.1$ instead of $\beta_1 = 1$. As the table shows, this has minimal effect on the estimates from the group-based sample, but produces a substantial increase in bias and variance for the individual-based sample. This is because $\rho_x$ (and thus $\gamma$) is identified in the individual-based sample from the effect of $\mathbf{x}_{gi}$ on $\bar{y}_{gi}$ relative to its effect on $y_{gi}$, as discussed in Section 3.3. When $\mathbf{x}_{gi}$ has little effect on $y_{gi}$, the effect is similar to that of a weak instrument in IV estimation. Row 5 shows results from a variation on the baseline experiment in which there are four explanatory variables (independent from one another with a $N(0,1)$ distribution) rather than just one. Because the previous experiment revealed that explanatory power matters, the coefficients on the $\mathbf{x}$ vector are set to $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0.5$ so that $var(\beta\mathbf{x})$ of the model is unchanged from the baseline experiment. As the table shows, this has little effect on the estimates of $\rho$ and $\gamma$. Row 6 shows the results froom a variation in which there are four binary explanatory variables. Because the structural model assumes normally distributed explanatory variables, it is slightly misspecified here in a way that is likely in applied work. As the results show, this form of misspecification has little effect on the estimators.

**Correlation between x and $\epsilon$:** Rows 7-10 report the results from simulations in which there is correlation between observables and unobservables. As the results here show, such a correlation will introduce a bias in the SML estimates. Whether the estimated peer effect is biased up or down appears to depend on two factors: whether the sample is individual-based or group-based, and how large the within-group correlation is relative to the individual-level correlation. At least in the cases presented here, the SML estimator still does well relative to the naive estimator, though it obviously would be inferior to some other estimator which is consistent under a correlation between observables and unobservables.

**Wrong equilibrium selection rule:** Rows 11-14 report the results from simulations in which the equilibrium selection rule is misspecified by the researcher. In all four cases, the estimates are calculated assuming that the low-activity equilibrium is already selected. This type of misspecification has minimal effect on the resulting parameter estimates, at least when the true selection rule falls within these three categories.

**Selection rule free estimation:** Rows 15-17 show results from selection rule free estimation of the model as described in Section 4.2. In each of these three experiments, the actual selection rule is the low-activity rule. The quantities reported are the median (across 100 simulations) of the lower bound and upper bound estimates of $\gamma$. The median is reported here because the distribution of estimated upper bounds has a strong positive skew. As these results show, selection rule free estimation is able to provide informative bounds on the endogenous effect for the group-based sample (especially when the true endogenous effect is zero), but is less informative for the individual-based sample.

**Contextual effects:** Rows 18-19 show the results when there is a contextual effect rather than an endogenous effect. While both estimators mistakenly interpret at least some of the contextual effect as an endogenous effect, the estimated peer effect from an individual-based sample has much lower bias. This appears to be in part because the correlation in observables is being identified from the correlation between a respondent's characteristics and the average choice of his or her peers, so the contextual effect produces an upwardly biased estimate of $\rho$ which then reduces the effect on the estimated $\gamma$. In estimating the model from a group-based sample, the contextual effect has little impact on the estimated $\rho$ (note that the estimates are approximately correct), and so will not reduce the bias in $\gamma$.

**Inconsistent reporting:** Rows 20-21 show the results for a case in which there is inconsistent reporting of one's own behavior and that of one's peers, as modeled in Section 4.3. In row 20, this inconsistent reporting is incorporated into the estimator, and $p_r$ is estimated. In row 21, the inconsistent reporting is not incorporated into the estimator (i.e., it is implicitly and incorrectly assumed that $p_r = 1$). As the table shows, the parameter estimates are quite inaccurate when inconsistent reporting is not accounted for, but accurate when it is.

# 6 Application: Youth smoking

This section provides a brief illustrative application to the estimation of close friend influence on youth smoking. The decision of young people to smoke cigarettes is a natural place to look for social interaction effects. Smoking is a social activity, and one in which many believe that peer influence is very strong, particularly among young people. In addition, the decision to start smoking as a teenager has profound and life-long consequences, and is thus a major concern for public health policy.

## 6.1 Background

The current consensus in the public health literature is that peers are critical influences in the decision to start smoking. For example, Wang et al. (2000, p. 1241) state that the "[s]moking literature has indicated that the influence of peers has been the single most important factor related to smoking acquisition," while Tyas and Pederson (1998, p. 416) report that "one of the most consistent findings in the literature is that of the social influence of peers and others on adolescent smoking." Unfortunately, this consensus is built on exactly the type of methodologically flawed research which has been criticized by Manski (1993) and others.

Several recent papers (Norton, Lindrooth and Ennett 1998, Gaviria and Raphael 2001, Powell, Tauras and Ross 2003) have used a simple IV method to estimate endogenous social effects in youth smoking at the classroom and/or neighborhood level. Although their estimates of the endogenous effect vary substantially, all three papers find that the bias in the naive estimator is actually negligible, i.e., there is no correlated effect. This finding is somewhat disconcerting, given that longitudinal data (Engels, Knibbe, Drop and de Haan 1997, Wang, Eddy and Fitzhugh 2000) indicates that young smokers in low-smoking peer groups are likely to switch into higher-smoking peer groups over time. Such behavior will tend to produce substantial correlated effects, but the IV estimates find no evidence of them.

In addition to the methodological differences, the application presented here differs from the previous IV studies by investigating the influence of close friends rather than larger and more formal groups. Estimates of close friend effects appear frequently in the public health literature on youth smoking (1995, 2002, 2003), but have been absent from the economics literature.

## 6.2 Data

The data source is the 1993 Teenage Attitudes and Practices Survey (TAPS), an individual-based survey conducted by the Centers for Disease Control (CDC) to learn more about the determinants of smoking and other risky behavior among U.S. teens. The outcome variable is an indicator of whether the respondent is what the CDC defines as a "current smoker" – someone who has smoked a cigarette in the past 30 days. The social group is the respondent's four best same-sex friends, and the measure of peer behavior is the number of those friends that smoke, as reported by the respondent. In addition, the TAPS provides standard background variables including race, age, parental smoking, exposure to information about the risks of smoking, and participation in sports and religious services. Wang et al., (1995) also use this data to estimate (by the naive method) close friend effects on smoking. After dropping observations without information on the endogenous variables, the sample has 8,192 respondents, all high-school age. The model is estimated with the correction for inconsistent reporting described in Section 4.3. The fraction of teens that admit to smoking is about 20%, while the average fraction of their friends who smoke is about 27%. The estimated rate of truthful reporting by smokers is thus $\hat{p}_r = \frac{0.20}{0.27} = 0.74$.

## 6.3 Results

Table 2 reports basic results. The first column in the table displays coefficient estimates from a naive probit, while the second column shows SML parameter estimates for the structural model. The naive probit results suggest a large endogenous social effect ($\hat{\gamma} = 1.891$). The structural model estimates imply a large correlated effect ($\hat{\rho} = 0.621$) and thus a much smaller endogenous effect ($\hat{\gamma} = 0.225$).

Because the structural model includes the underreporting correction, the magnitudes of the coefficient estimates are not directly comparable in the two columns, nor are they directly comparable with coefficient estimates from other studies which use different functional forms. To characterize the results in a way that is comparable, consider a representative individual, i.e one with observed characteristics such that his or her probability of being a self-reported smoker is equal to the average (0.20), and who has no close friends who smoke. According to each model, by how many percentage points will this representative individual's probability of being a self-reported smoker increase if one close friend becomes a smoker?

I calculate this quantity for both the naive model and the structural model, as well as for other models estimated in the literature. In this paper, the naive model predicts an increase in smoking probability by 16 percentage points (from 20% to 36%) as a result of one friend becoming a smoker. The structural model structural model predicts an increase in both actual and self-reported smoking of 1 to 2 percentage points. Previous articles in the literature on close friend peer effects, all of which use naive estimators, find even larger responses than the naive estimator here. Wang et al. (1995), also using the TAPS data, find that the probability of being a smoker increases by 37-53 percentage points, depending on age. Lloyd-Richardson et al. (2002) find that the probability of being a smoker increases by 34 percentage points, and Norton et al. (2003) find that it increases by 18 percentage points. Clearly the basic estimates from the structural model imply that close friends are less influential than previously thought.

The starkness of this result is reduced but not eliminated when the equal correlation assumption is relaxed. Figure 3 shows the results of relaxing this assumption, using the methodology described in Section 4.1. The figure shows a number of results. First, even if one imposes only the very weak restriction that peer groups exhibit at least as much similarity as would be the case if they were constructed by random assignment ($\rho_\epsilon \geq 0$), the data indicate the peer effect is lower than implied by the naive model. If one is willing to impose the slightly stronger restriction that $\rho_\epsilon \in [0.311, 0.621]$ (i.e., the correlation in unobservables is no more than the estimated correlation in observables, and no less than half the estimated correlation in observables), then this restriction produces bounds of approximately $[0.000, 0.909]$.

These results should be viewed as suggesting that close friend effects on youth smoking are weaker than implied by previous studies, rather than proving that close friend effects are near zero. There are some limitations in the TAPS data, particularly a lack of state identifiers so that cigarette prices can be incorporated into the model. In addition, space and scope limitations preclude a detailed analysis of the data here. Ongoing research applies the methodology developed here to richer data sets on youth smoking. and with more detailed analysis. The results here clearly merit further investigation.

# 7 Conclusion

Many economists are skeptical of the empirical importance of social interaction effects and suspect that their magnitude has been overstated by researchers who fail to account for correlated effects. Researchers have made some progress on this issue recently, but each of the main approaches in the literature has significant limitations. The approach developed here does not require centralized manipulation of the economic environment to generate experimental data, nor does it require the existence of a suitable instrumental variable. Instead, it merely requires standard survey data with some information on peer behavior. Although the estimation method itself is somewhat complex, a complex research design is not required. The available computer code can simply be applied to the data as a direct substitute for the naive estimator. This feature is important in a world where policy-oriented researchers, needing a point estimate for the peer effect in a given situation, will use the naive estimator if all of the alternatives require complex research designs.

Several avenues for further research remain, in addition to empirical applications. An ongoing research project investigates social interaction effects in youth smoking in more detail. While much remains to be done, the results here indicate that simulation-based structural estimation can be a valuable tool in the empirical analysis of social interactions.

# References


**Akerlof, George and Rachel E. Kranton**, "Economics and Identity," *Quarterly Journal of Economics*, 2000, *115* (3), 715–753.

**Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber**, "Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools," Working Paper 7831, NBER 2000.

**Becker, Gary S. and Kevin M. Murphy**, *Social Economics: Market Behavior in a Social Environment*, Harvard University Press, 2001.

**Bertrand, Marianne, Erzo F.P. Luttmer, and Sendhil Mullainathan**, "Network effects and welfare cultures," *Quarterly Journal of Economics*, 2000, *115*, 1019–1056.

**Bhat, Chandra R.**, "Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences," *Transportation Research*, 2003, *37* (9), 837–855.

**Brock, William A. and Steven N. Durlauf**, "Discrete choice with social interactions," *Review of Economic Studies*, 2001, *68* (2), 235–260.

__ **and** __ , "Interactions-based models," in James J. Heckman and Edward Leamer, eds., *Handbook of Econometrics, Volume 5*, North-Holland, 2001.

**Engels, Rutger C.M.E., Ronald A. Knibbe, Maria J. Drop, and Ypie T. de Haan**, "Homogeneity of cigarette smoking within peer groups: Influence or selection?," *Health Education and Behavior*, 1997, *24*, 810–811.

**Epple, Dennis and Holger Sieg**, "Estimating equilibrium models of local jurisdictions," *Journal of Political Economy*, 1999, *107* (4), 645–680.

**Evans, William N., Wallace E. Oates, and Robert M. Schwab**, "Measuring peer group effects: A study of teenage behavior," *Journal of Political Economy*, 1992, *100* (5), 966–991.

**Gaviria, Alejandro and Steven Raphael**, "School-based peer effects and juvenile behavior," *Review of Economics and Statistics*, 2001, *83* (2), 257–268.

**Glaeser, Edward L., Bruce I. Sacerdote, and José A. Scheinkman**, "Crime and Social Interactions," *Quarterly Journal of Economics*, 1996, *111* (2), 507–548.

__ , __ , **and** __ , "The Social Multiplier," Discussion Paper 1968, Harvard Institute of Economic Research 2002.

**Gouriéroux, Christian and Alain Monfort**, *Simulation-Based Econometric Methods*, Oxford University Press, 1996.

**Hajivassiliou, Vassilis, Daniel McFadden, and Paul Ruud**, "Simulation of multivariate normal rectangle probabilities and their derivatives: Theoretical and computational results," *Journal of Econometrics*, 1996, *72* (1), 85–134.

**Heckman, James J. and Guilherme Sedlacek**, "Heterogeneity, aggregation, and market wage functions: An empirical model of self-selection in the labor marker," *Journal of Political Economy*, 1985, *93* (6), 1077–1125.

**Hoxby, Caroline M.**, "Peer effects in the classroom: Learning from gender and race variation," Working Paper 7867, NBER 2000.

**Ioannides, Yannis M. and Jeffrey E. Zabel**, "Interactions, neighborhood selection, and housing demand," Working Paper, Tufts University 2002.

**Jovanovic, Boyan**, "Observable implications of models with multiple equilibria," *Econometrica*, 1989, *57* (6), 1431–1437.

**Katz, Lawrence F., Jeffrey R. Kling, and Jeffrey B. Liebman**, "Moving to Opportunity in Boston: Early results of a randomized mobility experiment," *Quarterly Journal of Economics*, 2001, *116* (2), 607–654.

**Krauth, Brian V.**, "Social interactions in small groups," Working Paper, Simon Fraser University 2001. Available at `http://www.sfu.ca/~bkrauth`.

__ , "Maximum likelihood estimation of social interaction effects with nonrandom group selection," Working Paper, Simon Fraser University 2004. Available at `http://www.sfu.ca/~bkrauth`.

**Kremer, Michael and Dan M. Levy**, "Peer effects from alcohol use among college students," Working Paper 9876, NBER 2001.

**Lloyd-Richardson, Elizabeth E., George Papandonatos, Alessandra Kazura, Cassandra Stanton, and Raymond Niaura**, "Differentiating stages of smoking intensity among adolescents: Stage-specific psychological and social influences," *Journal of Consulting and Clinical Psychology*, 2002, *70* (4), 998–1009.

**Manski, Charles F.**, "Identification of endogenous social effects: The reflection problem," *Review of Economic Studies*, 1993, *60* (3), 531–542.

**Milgrom, Paul and John Roberts**, "Rationalizability, learning, and equilibrium in games with strategic complementarities," *Econometrica*, 1990, *58*, 1255–1277.

**Moffitt, Robert A.**, "Policy interventions, low-level equilibria, and social interactions," in Steven N. Durlauf and H. Peyton Young, eds., *Social Dynamics*, MIT Press, 2001, pp. 45–82.

**Norton, Edward C., Richard C. Lindrooth, and Susan T. Ennett**, "Controlling for the endogeneity of peer substance use on adolescent alcohol and tobacco use," *Health Economics*, 1998, *7*, 439–453.

__ , __ , **and** __ , "How measures of perception from survey data lead to inconsistent regression results: Evidence form adolescent and peer substance use," *Health Economics*, 2003, *12*, 139–148.

**Powell, Lisa M., John A. Tauras, and Hana Ross**, "Peer Effects, Tobacco Control

Policies, and Youth Smoking Behavior," Working Paper, University of Illinois - Chicago 2003.
**Sacerdote, Bruce**, "Peer effects with random assignment: Results for Dartmouth roommates," *Quarterly Journal of Economics*, May 2001, *116* (2), 681–704.
**Tamer, Elie T.**, "Empirical strategies for estimating discrete games with multiple equilibria," Working Paper, Princeton University 2002.
__ , "Incomplete simultaneous discrete response model with multiple equilibria," *Review of Economic Studies*, 2002. Forthcoming.
**Topa, Giorgio**, "Social Interactions, Local Spillovers, and Unemployment," *Review of Economic Studies*, 2001, *68*, 261–295.
**Train, Kenneth**, *Discrete Choice Methods with Simulation*, Cambridge University Press, 2002.
**Tyas, Suzanne L. and Linda L. Pederson**, "Psychosocial factors related to adolescent smoking: A critical review of the literature," *Tobacco Control*, 1998, *7*, 409–420.
**U.S. Department of Health and Human Services, National Center for Health Statistics**, *Teenage Attitudes and Practices Survey, 1993: United States*, Ann Arbor MI: Inter-university Consortium for Political and Social Research, 1995. Computer file.
**Wang, Min Qi, Eugene C. Fitzhugh, R. Carl Westerfield, and James M. Eddy**, "Family and peer influences on smoking behavior among American adolescents: An age trend," *Journal of Adolescent Health*, March 1995, *16*, 200–203.
**__ , James M. Eddy, and Eugene C. Fitzhugh**, "Smoking acquisition: Peer influence and self-selection," *Psychological Reports*, 2000, *86*, 1241–1246.

| | | | | SML Estimator | | | |
|---|---|---|---|---|---|---|---|
| | Actual Values | | Naive Probit | Sample of Individuals | | Sample of Groups | |
| **Description** | $\gamma$ | $\rho$ | $\hat{\gamma}$ | $\hat{\gamma}$ | $\hat{\rho}$ | $\hat{\gamma}$ | $\hat{\rho}$ |
| Baseline | 1.0 | 0.00 | 1.525 | 0.996 | 0.001 | 1.032 | -0.003 |
| | | | (0.176) | (0.181) | (0.044) | (0.145) | (0.020) |
| | 0.5 | 0.25 | 1.489 | 0.514 | 0.243 | 0.495 | 0.248 |
| | | | (0.154) | (0.160) | (0.050) | (0.163) | (0.034) |
| | 0.0 | 0.25 | 0.834 | 0.067 | 0.233 | 0.065 | 0.238 |
| | | | (0.155) | (0.087) | (0.032) | (0.091) | (0.028) |
| Alternative assumptions on $X$: | | | | | | | |
| Low explanatory power | 0.5 | 0.25 | 1.818 | 0.795 | 0.139 | 0.497 | 0.247 |
| ($\beta_1 = 0.1$) | | | (0.134) | (0.573) | (0.230) | (0.157) | (0.032) |
| Multiple $X$'s ($k = 4$) | 0.5 | 0.25 | 1.489 | 0.485 | 0.255 | 0.467 | 0.254 |
| | | | (0.159) | (0.156) | (0.051) | (0.143) | (0.034) |
| Binary $X$'s. ($k = 4$) | 0.5 | 0.25 | 1.467 | 0.443 | 0.268 | 0.448 | 0.258 |
| | | | (0.152) | (0.172) | (0.053) | (0.153) | (0.030) |
| Correlation between $X$ and $\epsilon$: | | | | | | | |
| $\rho(\beta\mathbf{x}_{gi}, \epsilon_{gi}) = 0.2$, $\rho(\beta\mathbf{x}_{gi}, \epsilon_{gj}) = 0.0$ | 0.5 | 0.25 | 1.390 | 0.588 | 0.200 | 0.442 | 0.260 |
| $\rho(\beta\mathbf{x}_{gi}, \epsilon_{gi}) = 0.2$, $\rho(\beta\mathbf{x}_{gi}, \epsilon_{gj}) = 0.05$ | 0.5 | 0.25 | 1.572 | 0.686 | 0.239 | 0.633 | 0.253 |
| $\rho(\beta\mathbf{x}_{gi}, \epsilon_{gi}) = 0.2$, $\rho(\beta\mathbf{x}_{gi}, \epsilon_{gj}) = 0.1$ | 0.5 | 0.25 | 1.804 | 0.809 | 0.275 | 0.845 | 0.246 |
| Wrong equilibrium selection rule: | | | | | | | |
| Actual = High-activity | 1.0 | 0.00 | 1.453 | 0.891 | 0.011 | 0.955 | -0.006 |
| | 0.5 | 0.25 | 1.477 | 0.513 | 0.242 | 0.444 | 0.257 |
| Actual = Random | 1.0 | 0.00 | 1.517 | 0.974 | 0.001 | 0.976 | -0.001 |
| | 0.5 | 0.25 | 1.482 | 0.485 | 0.253 | 0.475 | 0.250 |
| Selection rule free estimation: | | | | | | | |
| Median[Lower bound,Upper bound] | 1.0 | 0.00 | | $[0.145, > 4.0]$ | | $[0.388, 2.706]$ | |
| | 0.5 | 0.25 | | $[0.082, > 4.0]$ | | $[0.177, 1.537]$ | |
| | 0.0 | 0.25 | | $[0.037, 3.566]$ | | $[0.001, 0.073]$ | |
| Contextual effects: | | | | | | | |
| $u(1) - u(0) = 1.0x_{gi} + 0.5\bar{x}_{gi} + \epsilon_{gi}$ | 0.0 | 0.25 | 1.222 | 0.108 | 0.367 | 0.379 | 0.221 |
| | 0.0 | 0.00 | 0.489 | 0.111 | 0.088 | 0.368 | -0.016 |
| Inconsistent reporting: | | | | | | | |
| $p_r = 0.5$, correction made | 0.5 | 0.25 | 0.678 | 0.499 | 0.252 | | |
| $p_r = 0.5$, no correction | 0.5 | 0.25 | 0.678 | -0.003 | 0.308 | | |

Table 1: Monte Carlo results. Calculated using 100 trials; see text for details.

| **Variable Description** | **Naive Probit** | **SML Estimator** |
|---|---|---|
| Peer correlation ($\rho$) | – | 0.621 |
| | – | (0.087) |
| Peer effect ($\gamma$) | 1.891 | 0.225 |
| | (0.050) | (0.235) |
| Intercept | -2.495 | -2.354 |
| | (0.290) | (0.563) |
| Black | -0.478 | -0.721 |
| | (0.064) | (0.106) |
| Age (years) | 0.073 | 0.134 |
| | (0.017) | (0.032) |
| Parental smoking | 0.102 | 0.273 |
| | (0.037) | (0.040) |
| Taught risks in class | -0.043 | -0.176 |
| | (0.044) | (0.040) |
| Plays sports | -0.152 | -0.301 |
| | (0.037) | (0.043) |
| Attends religious services | -0.203 | -0.458 |
| | (0.037) | (0.052) |

Table 2: Regression results for teen smoking data. Estimated standard errors in parentheses. Standard errors for SML estimates are estimated from 50 bootstrap replications.

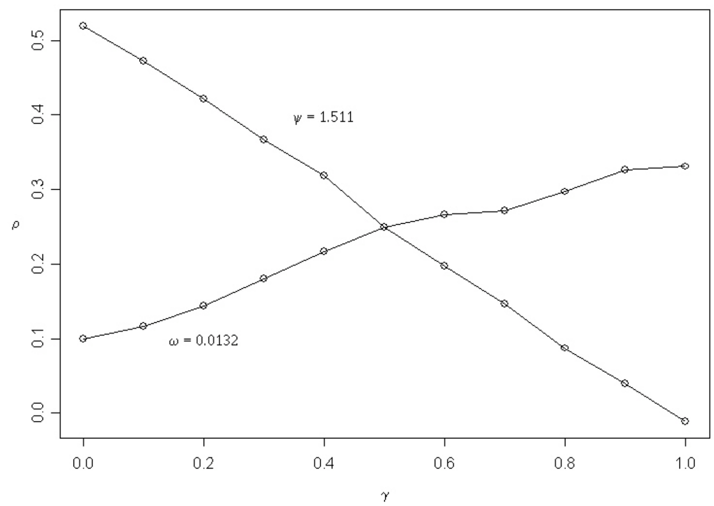


Figure 1: The upward-sloping curve is the set of $(\rho, \gamma)$ pairs such that $\hat{\omega} = 0.0132$. The downward-sloping curve is the set of $(\rho, \gamma)$ pairs such that $\hat{\psi} = 1.511$. Because these two curves intersect at only one point, these two relationships provide identification.

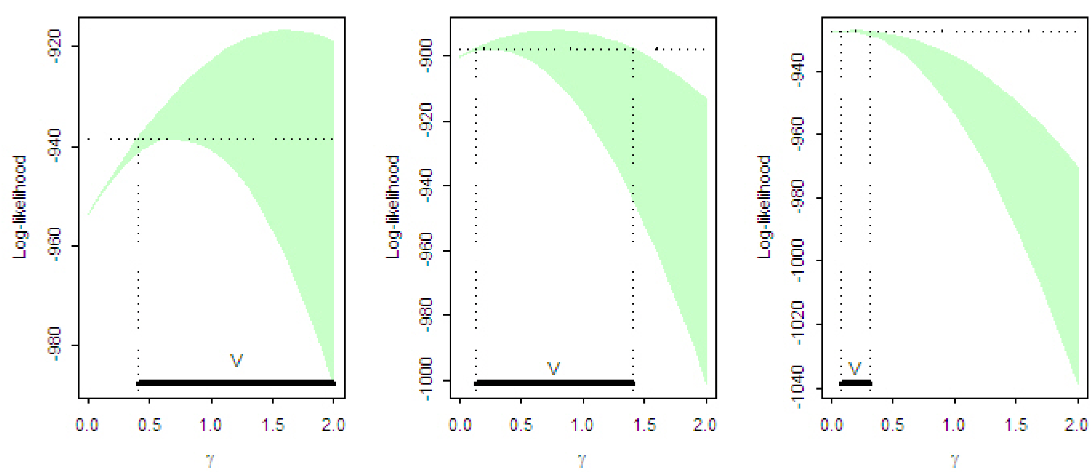


Figure 2: Mechanics of selection rule free estimation via likelihood bounds. All likelihood functions consistent with Nash equilibria lie within shaded area; dotted lines provide lower and upper bounds on peer effect $\gamma$. Graphs depict cases $(\rho, \gamma) = (0, 1)$, $(0.25, 0.5)$, and $(0.25, 0.0)$, respectively.

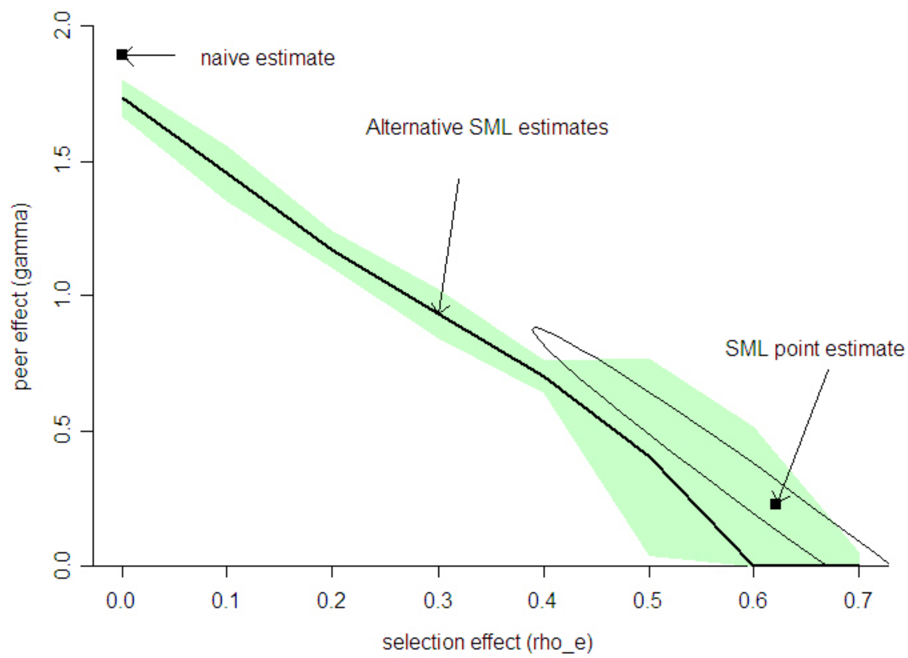


Figure 3: Results for teen smoking data which can be used to place bounds on social interaction effects. The individual points represent the point estimates from both the naive probit model and the structural SML model with the equal correlation restriction. The ellipse around the structural model estimate is an approximate 95% confidence region. The solid line represents the $\hat{\gamma}(\rho)$ function described in the text, and the shaded region is a pointwise asymptotic 95% confidence band for that function.

# A Additional not-for-publication notes

## A.1 Proof for Proposition 2

Proof: (1) We can simply construct the rectangle. Let:

$$\underline{c}_{gi} = \begin{cases} -\infty & \text{if } y_{gi} = 1 \\ \frac{1}{n_g - 1} \sum_{j \neq i} y_{gj} & \text{if } y_{gi} = 0 \end{cases} \qquad \bar{c}_{gi} = \begin{cases} \frac{1}{n_g - 1} \sum_{j \neq i} y_{gj} & \text{if } y_{gi} = 1 \\ \infty & \text{if } y_{gi} = 0 \end{cases}$$

and let $\underline{\mathbf{c}}_g = \left(\underline{c}_{g1}, \underline{c}_{g2}, \ldots, \underline{c}_{gn_g}\right)'$ and $\bar{\mathbf{c}}_g = (\bar{c}_{g1}, \bar{c}_{g2}, \ldots, \bar{c}_{gn_g})'$ Then $c(\mathbf{y}) = [\underline{\mathbf{c}}_g, \bar{\mathbf{c}}_g] \equiv \{\mathbf{c} \in R^{n_g} : \underline{\mathbf{c}}_g \leq \mathbf{c} \leq \bar{\mathbf{c}}_g\}$, which is a rectangle. (2) First, note that

$$C(Y) = \bigcap_{y \in \{0,1\}^n_g} \begin{pmatrix} [\underline{\mathbf{c}}(y), \bar{\mathbf{c}}(y)] & \text{if } y \in Y \\ [\underline{\mathbf{c}}(y), \bar{\mathbf{c}}(y)]^c & \text{if } y \notin Y \end{pmatrix}$$

Next, note that if $A = [a^L, a^H]$ and $B = [b^L, b^H]$ are both rectangles in $R^n$, then $A \cap B = [a^L \vee b^L, a^H \wedge b^H]$ is also a rectangle. In addition, if $A$ is a rectangle, then $A^c$ can be written as the union of a finite collection of (2*n) disjoint rectangles:

$$\begin{aligned} [a^L, a^H]^c &= [(-\infty, \ldots, -\infty), (a_1^L, \infty, \ldots, \infty)] \\ &\cup \; [(a_1^H, -\infty, \ldots, -\infty), (\infty, \ldots, \infty)] \\ &\cup \; [(a_1^L, -\infty, \ldots, -\infty), (a_1^H, a_2^L, \infty, \ldots, \infty)] \\ &\cup \; [(a_1^L, a_2^H, -\infty, \ldots, -\infty), (a_1^H, \infty, \ldots, \infty)] \\ &\cup \; \vdots \\ &\cup \; [(a_1^L, a_2^L, \ldots, a_{n-1}^L, a_n^H), (a_1^H, a_2^H, \ldots, a_{n-1}^H, \infty)] \end{aligned} \tag{27}$$

It follows that if $A$ and $B$ are rectangles then $A^c \cap B$ and $A^c \cap B^c$ can be expressed as the union of a finite collection of disjoint rectangles. Therefore, $C(Y)$ can be expressed as the union of a finite collection of disjoint rectangles. (3) follows from standard theorems on the conditional and marginal distribution of a linear function of a multivariate normal random vector. (4) We can write $\Pr(\mathbf{y}_g|\mathbf{X}_g;\theta)$ as (supressing the conditioning):

$$\begin{aligned} \Pr(\mathbf{y}_g|\mathbf{X}_g;\theta) &= \sum_Y \Pr(\mathbf{y}_g \text{ is selected}|Y_g = Y, \mathbf{X}_g;\theta) \Pr(Y_g = Y|\mathbf{X}_g;\theta) \\ &= \sum_Y \text{sel}(\mathbf{y}_g, Y) \Pr(\mathbf{c}_g \in C(Y)|\mathbf{X}_g;\theta) \end{aligned}$$

Since by the definition of the GHK simulator $\frac{1}{S}\sum_{s=1}^S GHK(M_g, \Sigma_g, C(Y)) \xrightarrow{p} \Pr(\mathbf{c}_g \in C(Y)|\mathbf{X}_g;\theta)$, the full GHK simulator (16) is consistent. The probability can also be written as:

$$\Pr(\mathbf{y}_g|\mathbf{X}_g;\theta) = \Pr(\mathbf{y}_g|\mathbf{y}_g \in Y_g, \mathbf{X}_g;\theta) \Pr(\mathbf{y}_g \in Y_g|\mathbf{X}_g;\theta)$$

Since by the definition of the GHK simulator, $\frac{1}{S}\sum_{i=1}^S \text{sel}(\mathbf{y}_g, Y(ghk(M_g, \Sigma_g, c(\mathbf{y}_g), \xi^s)) \xrightarrow{p} \Pr(\mathbf{y}_g|\mathbf{y}_g \in Y_g, \mathbf{X}_g;\theta)$ and $\frac{1}{S}\sum_{i=1}^S GHK(M_g, \Sigma_g, c(\mathbf{y}_g), \xi^s) \xrightarrow{p} \Pr(\mathbf{y}_g \in Y_g|\mathbf{X}_g;\theta)$, the GHK-frequency hybrid simulator (17) is also consistent. □

## A.2 Details on the "binary characteristics" Monte Carlo experiment

The "Binary Characteristics" experiment is complicated by there being many ways to generate a vector of binary variables with a particular mean vector and covariance matrix. The particular method used is as follows. For each explanatory variable $v$ and each group $g$, there is a group-specific probability $p_g^v$ drawn from a two-point distribution:

$$p_g^v = \begin{cases} p + \frac{\sqrt{\rho}}{2} & \text{with probability } 1/2 \\ p - \frac{\sqrt{\rho}}{2} & \text{with probability } 1/2 \end{cases} \tag{28}$$

Conditional on $p_g^v$, each group member's characteristic $x_{gi}^v$ is an independent draw from the Bernoulli($p_g^v$) distribution. Given these assumptions, $E(x_{gi}^v) = p$ and $corr(x_{gi}^v, x_{gj}^v) = \rho$. In order to keep the explanatory power of $x$ the same as the baseline experiment, the coefficients are set so that $E(\beta x_{g1}) = 0$ and $var(\beta x_{g1}) = 1$. Specifically, $p = 0.5$, $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 1$, and $\beta_0 = -2$. As usual, the vector of unobservables is multivariate normal with zero mean, unit variance, and correlation $\rho$ across group members.

## A.3 Details on comparison of previous studies

Because of differences in estimation methods across studies, and differences in standard reporting methods between researchers in health and in economics, some degree of subjective interpretation is involved. This section describes how each paper was interpreted to generate the estimates reported in Section 6.

**Wang et al. (1995)**

1. Data: Youth age 14-18 in the U.S. Teenage Attitudes and Practices Survey.
2. Peer group measure: Smoking rate of four same-sex best friends (0 smokers vs. 1-2 smokers vs. 3-4 smokers).
3. Method: Logistic regression, separately for each age, odds ratios reported.
4. Estimates: Odds ratio of 1-2 smokers vs 0 smokers among peers varies from 5.3 (Table 1, column 3) to 10.7 (Table 1, column 2). Therefore logit coefficient (on dummy variable "1-2 smokers among friends") varies from 1.67 to 2.37. Effect of an increase in best-friend smoking from zero to one is:

$$\begin{aligned} \Delta \Pr(y=1) &= \Lambda(\Lambda^{-1}(0.2) + 1.67) - 0.2 \\ &= 0.37 \\ \Delta \Pr(y=1) &= \Lambda(\Lambda^{-1}(0.2) + 2.37) - 0.2 \\ &= 0.53 \end{aligned}$$

**Lloyd-Richardson et al. (2002)**

1. Data: Grade 7-12 students in the U.S. National Longitudinal Survey of Adolescent Health.
2. Peer group measure: Smoking rate among 3 best friends.

3. Method: Logistic regression, odds ratios reported.
4. Estimates: Odds ratio for 1 peer who smokes vs. 0 peers who smoke was 4.68 (Table 2, column 2), implying logit coefficient of ln(4.68) = 1.543. Effect of a one person increase in best friend smoking is:

$$\begin{aligned} \Delta \Pr(y=1) &= \Lambda(\Lambda^{-1}(0.2) + 1.543) - 0.2 \\ &= 0.34 \end{aligned}$$

**Norton, et al. (2003)**

1. Data: Grade 9 students in a North Carolina study.
2. Peer group measure: Smoking rate of 3 best friends.
3. Method: Linear probability model. The researchers also included perceived smoking of peers (as distinct from their actual smoking) as an explanatory variable.
4. Estimates: LPM coefficient is 0.526 (Table 1, column 1). Effect of a one-person increase in peer smoking is :

$$\begin{aligned} \Delta \Pr(y=1) &= 0.526 * 0.33 \\ &= .18 \end{aligned}$$

**Krauth (current article)**

1. Data: Youth age 14-18 in the U.S. Teenage Attitudes and Practices Survey.
2. Peer group measure: Smoking rate of four same-sex best friends.
3. Method: Naive probit, structural estimator.
4. Estimates: Naive probit coefficient 1.891; structural coefficient estimates $\hat{\gamma} = 0.225$, $\hat{p}_r = 0.74$. Effect of a one-person increase in close-friend smoking predicted by naive model is:

$$\begin{aligned} \Delta \Pr(y=1) &= \Phi(\Phi^{-1}(0.2) + 1.891 * 0.25) - 0.2 \\ &= 0.16 \end{aligned} \tag{29}$$

The structural model predicts an increase in smoking probability of:

$$\begin{aligned} \Delta \Pr(y=1) &= \Phi(\Phi^{-1}(0.2) + 0.225 * 0.25) - 0.2 \\ &= 0.016 \end{aligned}$$

and an increase in self-reported smoking probability of:

$$\begin{aligned} \Delta \Pr(y=1) &= 0.74 * \Phi(\Phi^{-1}(0.27) + 0.225 * 0.25) - 0.2 \\ &= 0.014 \end{aligned}$$