A Partial Identification Subnetwork Approach to Discrete Games in Large Networks: An Application to Quantifying Peer Effects*

Tong Li† Li Zhao‡

June 10, 2016

Abstract

This paper studies identification and estimation of discrete games in large networks, with an application to peer effects on smoking in friend networks. Due to the presence of multiple equilibria, the model is not point identified. We adopt the partial identification approach by constructing moment inequalities on choice probabilities of subnetworks. Doing so not only significantly reduces the computational cost, but also enables us to find consistent estimator of the moment conditions even when the network is large and the friendship relationship structure varies significantly among networks. Monte Carlo

*This research uses data from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Information on how to obtain the Add Health data files is available on the Add Health website (http://www.cpc.unc.edu/addhealth). No direct support was received from grant P01-HD31921 for this analysis.

†Department of Economics, Vanderbilt University, tong.li@vanderbilt.edu.

‡Antai College of Economics and Management, Shanghai Jiao Tong University, li_zhao@sjtu.edu.cn.

studies are conducted to evaluate the performance of the subnetwork approach. In the application using the Add Health data, we find significant and positive peer effects on smoking.

1 Introduction

Peer effects play a central role in influencing individual behavior. In recent years, there is an exploding interest in studying interactions in social networks. For example, there is empirical evidence of peer effects on educational achievements (Zimmerman 2003, Calvó-Armengol, Patacchini and Zenou 2009), employment (Calvo-Armengol and Jackson 2004), health outcomes (Cohen-Cole and Fletcher 2008, Krauth 2006, Nakajima 2007, Badev 2013), risky behavior taking (Gaviria and Raphael 2001, Clark and Loheac 2007), adoption of new technology (Conley and Udry 2010), among others. Social interaction can be modeled as a system of equations where each equation is a regression of one person's action on the actions of his or her peers. This framework is widely used in studying peer effects on a continuous outcome. Whereas the identification of peer effects model with continuous outcomes has been studied by Manski (1993) and Bramoullé, Djebbari and Fortin (2009), identification and estimation issues of the model with discrete outcomes are not well addressed.

This paper develops an empirical method to study peer effects on discrete choices in large social networks. Our framework extends the linear network model in Bramoullé, Djebbari and Fortin (2009) to the case of binary outcomes. Our model belongs to a large and growing literature on discrete games of complete information, which includes entry game as a special case. It has been well known in the entry literature that due to the presence of multiple equilibria, estimating strategic interaction of discrete outcomes requires either strong assumptions or special econometric tools (Bjorn and Vuong 1984, Bresnahan and Reiss 1991a, Berry 1992, Tamer 2003, Ciliberto and Tamer 2009, Andrews, Berry and Jia 2004).[1] While both peer effects model and entry model study strategic interaction of discrete choices, existing methods for entry games is not suitable to estimate games in networks because the peer effects model is different from the entry model in a number of ways.

One empirical challenge that is new to games in networks is due to the large number of agents in a network. To the best of our knowledge, all applications in discrete games of complete information study strategic interaction among a handful of agents (Bjorn and Vuong

[1] For a survey on econometric approaches to games with multiple equilibria, see de Paula (2013).

1984, Bajari, Hong and Ryan 2010, Jia 2008, Krauth 2006, Soetevent and Kooreman 2007). There are various reasons why identification and inference of large games are difficult. First, point identification relies on the knowledge of all the equilibria of a game. In practice the set of equilibria is usually calculated by enumerating all outcomes of the game and checking whether each of them is an equilibrium. The number of outcomes grows at an exponential rate of the number of agents. Therefore, obtaining the set of equilibrium for games in networks is computationally demanding. Second, when a game is played by many agents, the sets of equilibria vary significantly across games, making it harder to find a reasonable assumption of equilibrium selection mechanism that is needed for point identification. Also with a large number of agents, if the individual shocks are dependent, the computation of the joint likelihood becomes intractable if not impossible. Third, existing partial identification approaches could not handle large games as they check moment conditions of all outcomes of the game. Each moment condition needs to be consistently estimated by a large number of networks of the same outcome. Since the number of outcomes is large, the number of networks needed for constructing moment conditions is enormous. In practice we may not have enough number of networks of the same outcome to construct moment conditions for all outcomes of the game.

The variation in the network structure adds further challenges in estimating peer effects. An individual's action depends on whom he or she connects with. This is in contrast to entry games, in which the "network" is fixed in the sense that every firm interacts with all others firms in a market. The peer effects model has an additional variation in friend relationship. The moment conditions, in the partial identification approach, need to be constructed for each network structure. For the same reason as above, there may not be enough observations of the same friend relation to calculate empirical probabilities, therefore partial identification approaches based on moment conditions of full networks are not feasible in practice.

The novelty of this paper is to address computational and consistency issues by partially identifying peer effects via subnetworks. Though the number of outcomes of a full network is enormous, the number of outcomes in subnetworks can be tractable. There is one additional

issue we need to address. Because people in a subnetwork interact with people outside the subnetwork, moment conditions of subnetworks need to consider these potential interactions. We seek conditions that will hold regardless of what actions people outside networks choose. Since moment conditions do not rely on the information outside the subnetwork, the number of moments needed to check depends on the features of the subnetwork only.

In the Monte Carlo study, we demonstrate that moment conditions of subnetworks are not only tractable, but also informative. The subnetwork approach successfully excludes parameter values that are far from the true parameters of the data generating process. By using the Monte Carlo examples, we also illustrate the factors that influence the performance of our subnetwork approach. Generally, the subnetwork approach performs better if the number of links connected to individuals inside and outside networks is small. This is because our approach cuts the dependence between agents inside and outside the subnetwork in exchange for computation tractability. Since many of the real world applications have sparse networks, our approach will be well suited for these applications.

The final part of this paper studies peer effects on smoking using data from the National Longitudinal Study of Adolescent to Adult Health (Add Health). Add Health is a nationwide survey of health related questions. This data set also contains information on friend nomination, from which we could form friend networks. The friend network revealed in Add Health data is very sparse. Using our econometric method, we find significant and positive peer effects of smoking.

This paper contributes to the peer effects literature by proposing a computationally feasible way of estimating peer effects on discrete outcomes. Our framework is closely related to Manski (1993) and Bramoullé, Djebbari and Fortin (2009) except that we consider discrete actions, and is also related to Krauth (2006) and Soetevent and Kooreman (2007), which both consider discrete choice models with social interactions assuming that the observed choices represent an equilibrium of the game played by all the interacting agents to get around the multiple equilibria problem.[2] Multiple equilibria is not an issue if individuals choose contin-

[2]de Paula (2009) considers inference in a synchronization game with social interactions, where the model

uous actions. As described in Bramoullé, Djebbari and Fortin (2009), the action of friends' friend could be served as an instrumental variable for the peer effect in the linear model. However, the instrumental variable approach could not be extended to the case of discrete choices. In the peer effects literature, most of the empirical studies on discrete choices follow Brock and Durlauf (2001) and Brock and Durlauf (2007), who model individual behavior as a best response to the expected behavior of peer group and study aggregate behavioral outcomes with social interactions imbedded in individual decisions.

This paper also adds to the growing literature of identification and inference of discrete games of complete information. For the best of our knowledge, all discussions of discrete games focus on small number of agents. For tractable number of players, point identification could be achieved in symmetric entry games (Berry 1992), by assuming or estimating equilibrium selection mechanism (Bjorn and Vuong 1984, Bajari, Hong and Ryan 2010), or by a large support condition (Tamer 2003). Partial identification approaches are discussed in Andrews, Berry and Jia (2004), Ciliberto and Tamer (2009), Galichon and Henry (2011), Beresteanu, Molchanov and Molinari (2011) and Henry, Meango and Queyranne (2015). We contribute to this literature by considering games in large but sparse networks. Though sharp identification is achievable in Galichon and Henry (2011), Beresteanu, Molchanov and Molinari (2011) and Henry, Meango and Queyranne (2015), because of the variation in network structure and the large size of the network, sharp identification is extremely difficult in the case we consider. Therefore, this paper exploits necessary but not sufficient conditions of the model as in Ciliberto and Tamer (2009).

Last but not the least, our work contributes to the new literature on the econometrics of networks. One topic in this area focuses on network formation; some studies model network formation as a complete information game (Sheng 2012, Uetake 2012) and others model network as an incomplete information game (Leung 2015).[3] Badev (2013) studies both network formation and interactions in networks. This paper contributes to the literature by studying

can be viewed as a simultaneous duration model with multiple decision makers and interdependent durations.

[3]Graham (2014) formalizes an empirical model of network formation that allows for detecting homophily when agents are heterogeneous.

interactions in networks. Our work is related to Sheng (2012) in the sense that both papers explore information from subnetworks to conduct inference. The objectives of our paper and Sheng (2012) are different because her work considers network formation while we study interactions in networks. While our approach does not attempt to model explicitly network formulation as in Sheng (2012), since our approach allows the individual shocks to be dependent, it provides a way to take into account network formulation as the dependence among individual shocks could capture the possibility that individuals with the same attributes are more likely to be friends.

The rest of paper is organized as follows. Section 2 describes our econometric framework and discusses the empirical challenges caused by the network features. Section 3 starts with an example of constructing moment conditions for a 2-person subnetwork game when the full network is of size 4. Then we discuss identification and inference in general cases. Section 4 is devoted to studying the performance of the subnetwork approach. Section 5 conducts an empirical exercise of peer effects on smoking. Section 6 concludes.

2 Model

2.1 Model Setup

Network A network can be described as a graph of nodes and edges. Each node represents an agent, which can be a person or a firm. Each edge connects one pairs of nodes and represents a relationship, such as friendship.

Let $V = \{1, 2, ..N\}$ be the set of agents in the network. Links are non-directional.[4] Let $g_{ij} = g_{ji} = 1$ if i and j are connected, $g_{ij} = g_{ji} = 0$ otherwise. The collection of links forms an $n \times n$ matrix called G, filled with zeros and ones. We study interaction between agents, taking the formation of network as given.[5]

[4]Our model can easily extend to a directional network. Because our focus is to study peer effects in a network, the assumptions of non-directional friend relationship is more appropriate.

[5]Estimating network formation is computationally intensive and often relies on partial identification approach. See Sheng (2012) and Uetake (2012). In this paper, we assume network is exogeneouly determined.

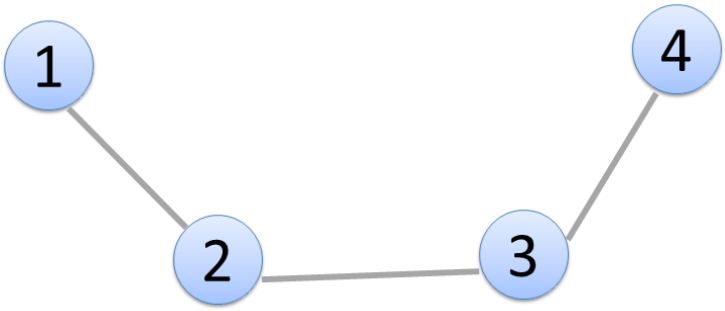


Figure 1: A Graph of 4-Person Network

Figure 1 illustrates a friend network of 4 individuals. Each link denotes a friend relationship. In this example $g_{12} = g_{21} = 1$ because person 1 and 2 are connected. $g_{13} = g_{31} = 0$ because person 1 and 3 are not connected.

In the discussion later, we will develop our identification strategy using the information about subnetworks. A subnetwork consists of a subset of agents and the links associated with these agents. The subnetwork A contains three types of information: 1) a set of players A; 2) the links between agents in A: $G_A = \{g_{ij}\}$, $(i,j) \in A$; and 3) the number of links connecting to agents outside the subnetwork $n_A = \{n_{A,i}\}$ for $\forall i \in A$, where $n_{A,i} = \sum g_{ij} \cdot 1(j \notin A)$.

For example, we may be interested in the subnetwork that consists of agents 2 and 3. Let $A = \{2,3\}$ denote the set of agents inside the subnetwork. $G_A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ reveals that agents 2 and 3 are connected. $n_A = [1,\ 1]$ because both agents 2 and 3 have one link connecting to agents outside the subnetwork. In inference, we will use (A, G_A, n_A) to construct moment inequalities.

Utility Function Agents play a simultaneous game of complete information. Each individual i chooses a binary action $y_i \in \{0,1\}$. Normalize the utility of action 0 to 0. The utility of the alternative action is affected by person i's characteristics $\mathbf{x}_i$, individual shock ϵ_i, and

the average actions taken by individuals that are connected with i:

$$u(y_i, \mathbf{x}_i, y_{-i}; \boldsymbol{\beta}, \gamma) = \boldsymbol{\beta}\mathbf{x}_i + \gamma \frac{\sum_{j \in V} g_{ij} y_j}{\sum_{j \in V} g_{ij}} + \epsilon_i, \tag{1}$$

where γ represents the endogenous (or peer) effect (Manski 1993, Soetvent and Kooreman 2007). Assume $\gamma > 0$, in which case γ is also called a social multiplier. Under the framework of games of complete information, $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, ...\epsilon_N)$ is observed by all game players, and is assumed to follow a joint normal distribution with zero mean and variance-covariance matrix V that has unit diagonal terms and possibly nonzero off-diagonal terms allowing for correlation among the individual shocks/types.[6]

Agent i chooses the action that has a higher utility, hence

$$y_i = 1(u(y_i, \mathbf{x}_i, y_{-i}; \boldsymbol{\beta}, \gamma) > 0).$$

In this paper, we adopt the standard form of utility function that assumes a linear functional form and homogeneous effects. This is the model used in early studies, e.g. Bresnahan and Reiss (1991a) and Berry (1992). Our identification result can extend to non-linear utility functions and cases where the interaction effects are heterogeneous. For example, we can let γ vary across i or j as in Ciliberto and Tamer (2009). In addition, our framework could be extended to the case of negative strategic interactions, only by a minor change in moment conditions. It is also feasible to allow for a correlation between shocks, if we consider the correlation as one additional parameter to be estimated. For the illustrative purpose, in what follows, we keep the simple form of utility function with a positive and homogeneous interaction effect γ.

Throughout the paper, we consider the benchmark model that has the utility function given in (1). Our approach can be readily extended to the following general form of the

[6]Such a correlation, if positive, represents the homophily principle, as discussed in Easley and Kleinberg (2010), McPherson, Smith-Lovin and Cook (2001) and Liu and Xu (2015). Assuming that agents know which equilibrium solution is selected to play and that the equilibrium selection mechanism is a deterministic function of $\mathbf{x}$, Liu and Xu (2015) consider the Brock and Durlauf (2001) model with homophily and proposes semiparametric estimation. Our approach can be extended to the model in Liu and Xu (2015) when multiple equilibria are allowed.

utility function

$$u(y_i, \mathbf{x}_i, y_{-i}; \boldsymbol{\beta}, \gamma) = \boldsymbol{\beta}\mathbf{x}_i + \gamma\psi\left(\frac{\sum_{j\in V} g_{ij}y_j}{\sum_{j\in V} g_{ij}}\right) + \delta\frac{\sum_{j\in V} g_{ij}\mathbf{x}_j}{\sum_{j\in V} g_{ij}} + \epsilon_i,$$

where $\psi(\cdot)$ is an increasing function, and δ represents the contextual or exogenous effect (Manski 1993).

It should also be noted that we focus on games of complete information. Actions are made in response to actual actions of others rather than to the belief of actions derived from the distribution of other players' types. Complete information is a reasonable assumption in applications such as the peer effects model, because an individual plays best response to the realized action of his or her peers. The second reason why we focus on complete information is because the equilibrium solutions to complete and incomplete games differ from each other significantly. To the best of our knowledge, the bound estimation approach has not yet be fully developed to estimate games of incomplete information.

Equilibrium We focus on pure strategy Nash equilibrium. An outcome is a Nash equilibrium if all players play best response to each other. Let $\mathbf{x}$ be the matrix of observed characteristics of all agents and let $\boldsymbol{\epsilon}$ be the vector of unobserved characteristics. The possible Nash equilibrium of a game is determined by the utility function, which is a function of individual characteristics $(\mathbf{x}, \boldsymbol{\epsilon})$ and the set of parameters $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \gamma\}$. Let $\mathcal{E}(\boldsymbol{\epsilon}, \mathbf{x}; \boldsymbol{\theta})$ denote the set of Nash equilibria, $\mathcal{E}(\boldsymbol{\epsilon}, \mathbf{x}; \boldsymbol{\theta})$ is defined as

$$\mathcal{E}(\boldsymbol{\epsilon}, \mathbf{x}; \boldsymbol{\theta}) = \{y \in \{0,1\}^N : \ y_i = 1(\boldsymbol{\beta}\mathbf{x}_i + \gamma\frac{\sum_{j\in V} g_{ij}y_j}{\sum_{j\in V} g_{ij}} + \epsilon_i > 0), \ \forall i \in V\},$$

$\mathcal{E}(\boldsymbol{\epsilon}, \mathbf{x}; \boldsymbol{\theta})$ may contain one or more outcomes, depending on the realization of $\boldsymbol{\epsilon}$ and $\mathbf{x}$. In the example above, both $y_1 = (0,0,0,0)$ and $y_2 = (1,1,1,1)$ are equilibria of games such that $-(\boldsymbol{\beta}\mathbf{x}_i + \gamma) < \epsilon_i < -\boldsymbol{\beta}x_i$, $\forall i \in V$. There are many other combinations of multiple equilibria for games with different utility profiles.[7]

[7]For example, both $y_1 = (0,0,0,0)$ and $y_3 = (0,1,1,0)$ are equilibria of games such that $\epsilon_i < -(\boldsymbol{\beta}\mathbf{x}_i + \frac{1}{2})$, $\forall i \in \{1,4\}$ and $-(\boldsymbol{\beta}\mathbf{x}_i + \frac{1}{2}) < \epsilon_i < -\boldsymbol{\beta}\mathbf{x}_i$, $\forall i \in \{2,3\}$.

The presence of multiple equilibria is not a special feature of games in network, but rather a common feature in discrete games. Various approaches have been proposed to estimate discrete games (Bjorn and Vuong 1984, Bresnahan and Reiss 1991a, Berry 1992, Tamer 2003, Andrews, Berry and Jia 2004, Ciliberto and Tamer 2009, Galichon and Henry 2011, and Beresteanu, Molchanov and Molinari 2011). An entry game can be thought of as a special case of games in networks, where all agents connect to each other. When agents are not necessarily linked to all other agents, the property of multiple equilibria is much harder to describe, because the set of equilibria depends on the network structure as well. More importantly, in applications, we usually deal with networks with a large number of nodes and varying sizes. In the next subsection we elaborate on the reasons that make the estimation of games in network more challenging.

2.2 Empirical Challenges of Estimating Discrete Games in Large Networks

As discussed in Bresnahan and Reiss (1991a), Tamer (2003) and Ciliberto and Tamer (2009), when a game has multiple equilibria, the probability of outcomes are not well-defined without information on equilibrium selection. Traditional methods such as the likelihood approach and the method of moments cannot be used. Games in networks have other features that complicate identification and estimation. The network we consider contains more than a few agents. Data may contain networks of different sizes and structures. In this subsection we will explain why existing approaches for estimating discrete games are not easily extendable for games in networks.

In principle, point identification could be achieved if equilibrium selection mechanism is given (Bjorn and Vuong 1984, Jia 2008), or estimated (Bajari, Hong and Ryan 2010, Narayanan 2013). Point identification requires the calculation of all the equilibria of a game. A standard way of obtaining the set of equilibria is to enumerate all the possible outcomes of the game and check if each of them is an equilibrium. This method is not feasible in practice if the number of agents is large, because the number of possible outcomes that need

to be checked grows at an exponential rate. For example, in a binary game with 20 agents, the number of possible outcomes is $2^{20} \approx 10^6$. In the data we consider, the majority of friend networks have sizes range from 70 to 90. It will be computationally costly to check such large number of outcomes for a network. Moreover, even if we are able to find ways to calculate all the possible equilibria of all games, point identification is still questionable because the composition of multiple equilibria varies significantly across games. It is hard to justify that the assumed selection mechanism is the true selection mechanism of the data generating process.

The partial identification approach does not rely on assumptions on equilibrium selection or calculation of all equilibria of the model. Thus it is more suitable for the model that is considered in this paper. Following the idea of Ciliberto and Tamer (2009), moment conditions could be constructed by imposing bounds on the probability of the game outcomes. In the 4 person example, $y = \{0, 0, 0, 0\}$ will be observed only if every players play best response. The upper bound of observing y is the probability that $u_i < 0$, $\forall i$, therefore $\Pr(y) \leq \Pr(u_i < 0, \forall i)$. In a game with N players, there are a total of 2^N upper bounds that could be used to construct moment conditions.

There are a number of concerns that can arise from constructing bounds like this. First, because the number of possible outcomes increases exponentially, the number of moment inequalities that we need to check grows at the exponential rate as well. In a network with 70 agents, the total number of outcomes of full network is more than 10^{21}. It is computationally infeasible to check so many number of moment conditions. More seriously, each individual moment condition cannot be consistently estimated if the number of outcomes is enormous. We consider the case where the number of agents is large in data, but we do not require the number of networks to be large. There will not be enough networks of the same outcome available in data to construct the empirical probability of each outcome, therefore moment conditions cannot be verified.

Another problem arising from games of network is the variation of structure among networks. In an entry game, each market has the same number of potential entrants, whose

decision is affected by all the rest of players in the market. In the peer effects model, each disconnected network is an analogue of a market in an entry game. However, the number of people and the friendship relationship among them are generally different across disconnected networks. The outcome of a network depends on how players are linked. If a moment inequality is placed on the outcome of the full network, there may not be enough number of networks of the same size and structure to construct moment conditions of the game.

In this paper, instead of constructing moment conditions of outcomes of full networks, we address the computation and consistency issues by exploring properties of subnetworks. Because there are links that connect agents inside and outside the subnetwork, we have to consider the interaction between agents inside and outside the subnetwork as well. The novelty of our method is to find conditions of subnetwork that are satisfied regardless of which actions people outside the network take. By this additional relaxation in constructing upper bounds, the moment conditions can be easily verified. We detail our identification strategy in the next section.

3 Identification and Inference via Subnetworks

3.1 An Illustration

In the 4-person peer effects model in Figure 1, assuming no control variable $\mathbf{x}$, agents' actions are characterized by the following set of equations:

$$\begin{aligned} y_1 &= 1(\gamma \cdot y_2 + \epsilon_1 > 0), \\ y_2 &= 1(\gamma \cdot \frac{y_1 + y_3}{2} + \epsilon_2 > 0), \\ y_3 &= 1(\gamma \cdot \frac{y_2 + y_4}{2} + \epsilon_3 > 0), \\ y_4 &= 1(\gamma \cdot y_3 + \epsilon_4 > 0). \end{aligned}$$

As before, assume $\epsilon \overset{i.i.d.}{\sim} N(0,1)$. We temporarily assume $\gamma > 0$.

Consider a subnetwork that consists of agents 2 and 3. Our goal is to find moment inequalities for outcomes of the subnetwork, for example, the probability of observing the

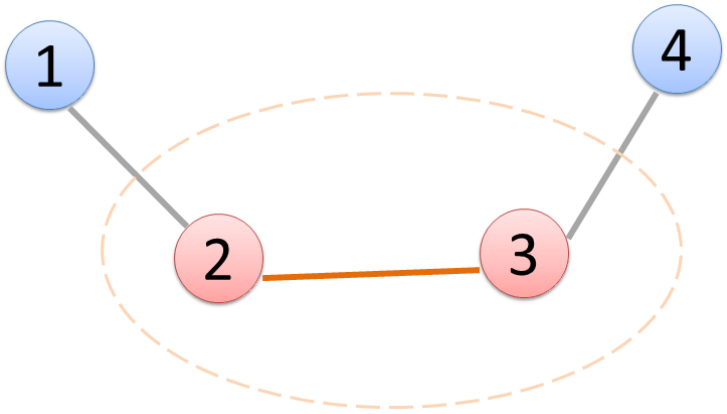


Figure 2: An Illustration of Subnetwork

event

$$A := (y_2 = 0, y_3 = 1).$$

$(y_2 = 0, y_3 = 1)$ is observed if and only if one of the following outcomes of the full network is observed.

$$\begin{aligned}
B_1 &:= (y_1 = 0, y_2 = 0, y_3 = 1, y_4 = 0), \\
B_2 &:= (y_1 = 0, y_2 = 0, y_3 = 1, y_4 = 1), \\
B_3 &:= (y_1 = 1, y_2 = 0, y_3 = 1, y_4 = 0), \\
B_4 &:= (y_1 = 1, y_2 = 0, y_3 = 1, y_4 = 1).
\end{aligned}$$

By checking the best response functions for each agents, it is verifiable that B_1 is an Nash equilibrium of a game if and only if $\epsilon \in R_1$, where

$$R_1 := \{\epsilon \in \mathbb{R}^4 : \gamma \cdot 0 + \epsilon_1 < 0; \gamma \cdot \frac{1}{2} + \epsilon_2 < 0; \gamma \cdot 0 + \epsilon_3 > 0; \gamma \cdot 1 + \epsilon_4 < 0\}.$$

If $\epsilon \notin R_1$, B_1 cannot be observed because it is not a Nash equilibrium of the game. If $\epsilon \in R_1$, the game may have other equilibria in addition to B_1. Whether or not observing B_1 depends

on how multiple equilibria are selected. Therefore $\epsilon \in R_1$ is a necessary condition of observing B_1.

Similarly, $\epsilon \in R_2$ is a necessary condition of observing B_2, where

$$R_2 := \{\epsilon \in \mathbb{R}^4 : \gamma \cdot 0 + \epsilon_1 < 0; \gamma \cdot \frac{1}{2} + \epsilon_2 < 0; \gamma \cdot \frac{1}{2} + \epsilon_3 > 0; \gamma \cdot 1 + \epsilon_4 > 0\};$$

$\epsilon \in R_3$ is a necessary condition of observing B_3, where

$$R_3 := \{\epsilon \in \mathbb{R}^4 : \gamma \cdot 0 + \epsilon_1 > 0; \gamma \cdot 1 + \epsilon_2 < 0; \gamma \cdot 0 + \epsilon_3 > 0; \gamma \cdot 1 + \epsilon_4 < 0\};$$

And $\epsilon \in R_4$ is a necessary condition of observing B_4, where

$$R_4 := \{\epsilon \in \mathbb{R}^4 : \gamma \cdot 0 + \epsilon_1 > 0; \gamma \cdot 1 + \epsilon_2 < 0; \gamma \cdot \frac{1}{2} + \epsilon_3 > 0; \gamma \cdot 1 + \epsilon_4 > 0\}.$$

Define H as

$$H := \{\epsilon \in \mathbb{R}^4 : \gamma \cdot \frac{1}{2} + \epsilon_2 < 0; \gamma \cdot \frac{1}{2} + \epsilon_3 > 0\}.$$

It is easy to check that $R_i \subset H, \forall i = 1, 2, 3, 4$.

Putting these together, we get

$$\begin{aligned} \Pr(A) &= \Pr(B_1 \textit{ or } B_2 \textit{ or } B_3 \textit{ or } B_4) \\ &\leq \Pr(\epsilon \in (R_1 \cup R_2 \cup R_3 \cup R_4)) \\ &\leq \Pr(\epsilon \in H). \end{aligned} \tag{2}$$

By replacing H and A with their expressions, the following inequality for the probability of the outcome in subnetwork $A = \{2, 3\}$ is satisfied :

$$\Pr(y_2 = 0, y_3 = 1) \leq \Pr(\gamma \cdot \frac{1}{2} + \epsilon_2 < 0; \gamma \cdot \frac{1}{2} + \epsilon_3 > 0). \tag{3}$$

Note that because $y_2 = 1(\gamma \cdot \frac{y_1+y_3}{2} + \epsilon_2 > 0)$ and $y_3 = 1$, $\gamma \cdot \frac{1}{2} + \epsilon_2 < 0$ is a necessary condition for $y_2 = 0$ when $y_1 = 0$. Similarly, because $y_3 = 1(\gamma \cdot \frac{y_2+y_4}{2} + \epsilon_3 > 0)$ and $y_2 = 0$, $\gamma \cdot \frac{1}{2} + \epsilon_3 > 0$ is a necessary condition for $y_3 = 1$ when $y_4 = 1$. The upper bound is coincident

with the conditions that requires each player i inside the subnetwork to play best response to the hypothetical scenario such that player i's all friends outside subnetwork A take the same action as player i does.

3.2 Moment Conditions in General Cases

For the full network consists of notes $V = \{1, 2, ...N\}$, let y_V denote the outcome of all players in V. y_V is a Nash equilibrium of a game if each individual in V plays best response given the actions of all other players in the network. Player i chooses action 1 if the utility of choosing action 1 is positive, because the utility of the alternative is normalized to 0. Player i chooses the alternative if the utility of action 1 is negative.

Let

$$\begin{aligned} R(y_V; x; \theta) \quad := \quad & \{\epsilon \in \mathbb{R}^N : u(y_i, x_i, y_{-i}; \theta) \geq 0, \ \forall y_i = 1, i \in V; \\ & u(y_i, x_i, y_{-i}; \theta) \leq 0, \ \forall y_i = 0, i \in V\}, \end{aligned} \tag{4}$$

denote the set of games of which y_V is a Nash equilibrium. If the utility function takes the form as in Eq. (1) , $R(y_V; x; \theta)$ is equivalent to

$$R(y_V; x; \theta) = \{\epsilon \in \mathbb{R}^N : (2 \cdot y_i - 1) \cdot (\beta x_i + \gamma \frac{\sum_{j \in V} g_{ij} y_j}{\sum_{j \in V} g_{ij}} + \epsilon_i) \geq 0, \ \forall i \in V\}. \tag{5}$$

y_V will only be observed if it is an equilibrium of the game. Hence

$$\Pr(y_V | \mathbf{x}) \leq \Pr(\boldsymbol{\epsilon} \in R(y_V; \mathbf{x}, \boldsymbol{\theta})). \tag{6}$$

This is an example of moment inequalities of full network. As discussed in the previous section, moment conditions of full network cannot be consistently estimated if the size of network is large, or if the interaction matrix varies across networks. Therefore we need to seek alternative moment conditions.

Recall that a subnetwork contains three types of information: the list of agents A, the connects among agents in the subnetwork G_A, and the number of connections to agents

outside the subnetwork n_A. Theorem 1 shows how moments of subnetwork are bounded above by moments predicted by the model. Moment inequalities like this could be used to make partial identification of the original model.

Theorem 1 *Consider a simultaneous game of complete information in network* $V = \{1, 2, ...N\}$ *with the utility function*

$$y_i = 1(\boldsymbol{\beta}\mathbf{x}_i + \gamma \frac{\sum_{j \in V} g_{ij} y_j}{\sum_{j \in V} g_{ij}} + \epsilon_i > 0),$$

where $\gamma > 0$ *and* $\epsilon_i \overset{i.i.d.}{\sim} N(0,1)$, $\forall i \in V$. *Let* A *be a subset of* V. *Define*

$$\begin{aligned} & H(y_A, \mathbf{x}, G_A, n_A; \boldsymbol{\theta}) \\ := \quad \{\epsilon \in \ & \mathbb{R}^N : (2 \cdot y_i - 1) \cdot (\boldsymbol{\beta}\mathbf{x}_i + \gamma \frac{\sum_{j \in V} g_{ij}[1(j \in A) \cdot y_j + 1(j \notin A) \cdot y_i]}{\sum_{j \in V} g_{ij}} + \epsilon_i) \geq 0, \ \forall i \in A\} \\ = \quad & \{\epsilon \in \mathbb{R}^N : (2 \cdot y_i - 1) \cdot (\boldsymbol{\beta}\mathbf{x}_i + \gamma \frac{\sum_{j \in A} g_{ij} \cdot y_j + n_{A,i} \cdot y_i}{\sum_{j \in A} g_{ij} + n_{A,i}} + \epsilon_i) \geq 0, \ \forall i \in A\} \end{aligned}$$

The following inequality holds for any $A \subset V$ *:*

$$\Pr(y_A | \mathbf{x}, G_A, n_A; \boldsymbol{\theta}) \leq \Pr(\epsilon \in H(y_A, \mathbf{x}, G_A, n_A; \boldsymbol{\theta})). \tag{7}$$

$H(y_A, \mathbf{x}, G_A, n_A; \boldsymbol{\theta})$ is the key innovation of our subnetwork approach. Besides that the conditions of $\boldsymbol{\epsilon}$ are placed on the agents in the subnetwork rather than all agents, what is special about $H(y_A, \mathbf{x}, G_A, n_A; \boldsymbol{\theta})$ is that y_j in $R(y_V, \mathbf{x}; \boldsymbol{\theta})$ is replaced by $1(j \in A) \cdot y_j + (j \notin A) \cdot y_i$. Mathematically, this new term takes value y_j for agent j inside the subnetwork and takes value y_i for j outside the subnetwork. In other words, for individual i, we assume all of his or her friends outside the subnetwork takes the same value as i takes, regardless of what their true actions are. This is because when we focus on actions in the subnetwork, we look for conditions that will be satisfied regardless of what action agents outside the subnetwork take. When we construct the upper bound for $\Pr(y_A)$, we seek actions that will make y_A most likely to happen. When $\gamma > 0$, individual i will gain extra utility of taking an action if a larger percent of his or her friends take the same action. Following this intuition, for $i \in A$

who has friends outside the subnetwork, $y_i = 0$ is more likely to occur if i's "outside friends" all take action 0; alternatively, $y_i = 1$ is more likely to occur if i's "outside friends" all take action 1, that is why we replace the actions of agents outside the network by the action of player i in the definition of $H(y_A, \mathbf{x}, G_A, n_A; \boldsymbol{\theta})$.

As an extension of Theorem 1, we can also consider the case where the interaction effect is negative. In this case, we replace the actions of agents outside the network by the opposite of player i'th action. This is summarized in the next corollary:

Corollary 2 *Consider a simultaneous game of complete information in network* $V = \{1, 2, ...N\}$ *with the utility function*

$$y_i = 1(\boldsymbol{\beta}\mathbf{x}_i - \gamma \frac{\sum_{j \in V} g_{ij} y_j}{\sum_{j \in V} g_{ij}} + \epsilon_i > 0), \tag{8}$$

where $\gamma > 0$ *and* $\epsilon_i \overset{i.i.d.}{\sim} N(0,1)$, $\forall i \in V$. *Let* A *be a subset of* V. *Define*

$$\begin{aligned} & \tilde{H}(y_A, \mathbf{x}, G_A, n_A; \boldsymbol{\theta}) \\ := \quad & \{\epsilon \in \mathbb{R}^N : (2 \cdot y_i - 1) \cdot (\boldsymbol{\beta}\mathbf{x}_i - \gamma \frac{\sum_{j \in V} g_{ij}[1(j \in A) \cdot y_j + 1(j \notin A) \cdot (1 - y_i)]}{\sum_{j \in V} g_{ij}} + \epsilon_i) \geq 0, \ \forall i \in A\} \\ = \quad & \{\epsilon \in \mathbb{R}^N : (2 \cdot y_i - 1) \cdot (\boldsymbol{\beta}\mathbf{x}_i - \gamma \frac{\sum_{j \in A} g_{ij} \cdot y_j + n_{A,i} \cdot (1 - y_i)}{\sum_{j \in A} g_{ij} + n_{A,i}} + \epsilon_i) \geq 0, \ \forall i \in A\} \end{aligned}$$

The following inequality holds for any $A \subset V$ *:*

$$\Pr(y_A | \mathbf{x}, G_A, n_A; \boldsymbol{\theta}) \leq \Pr(\boldsymbol{\epsilon} \in \tilde{H}(y_A, \mathbf{x}, G_A, n_A; \boldsymbol{\theta})). \tag{9}$$

3.3 Inference and Estimation

Our inference procedure is based on subnetworks. For each subnetwork A, the model predicts moment inequality (7) if the interaction effect is positive.

Suppose we are interested in subnetwork A, let y_A denote all the possible outcomes of A. The identified set is

$$\Theta_I = \{\theta : \Pr(y_A | \mathbf{x}, G_A, n_A) \leq \Pr(\boldsymbol{\epsilon} \in H(y_A; \mathbf{x}, G_A, n_A, \boldsymbol{\theta})) \ \forall y_A \in Y_A\},$$

where $\Pr(y_A|\mathbf{x}, G_A, n_A)$ is the choice probability of the data.

Let $m(y_A, \mathbf{x}, G_A, n_A; \boldsymbol{\theta}) = P(y_A|\mathbf{x}, G_A, n_A) - \Pr(\boldsymbol{\epsilon} \in H(y_A, \mathbf{x}, G_A, n_A; \boldsymbol{\theta}))$, the model predicts $m(y_A, \mathbf{x}, G_A, n_A; \boldsymbol{\theta}) \leq 0$. Let $\mathbf{m}(\mathbf{x}, G_A, n_A; \boldsymbol{\theta})$ be the vector of of these moment conditions for all outcomes of subnetwork A. Our inference procedure uses the objective function

$$Q(\boldsymbol{\theta}) = E_{(\mathbf{x}, G_A, n_A)}||(\mathbf{m}(\mathbf{x}, G_A, n_A; \boldsymbol{\theta}))_+||$$

where $(x)_+ = \max(x, 0)$, and $||\cdot||$ is the Euclidean norm.[8]

To make inference, we use the sample analogue of $Q(\theta)$. We randomly select S ($S \to \infty$) subnetworks of the same size as A. The sample analogue of criterion function is

$$Q_S(\boldsymbol{\theta}) = \frac{1}{S}\sum_{s=1}^{S} ||(\mathbf{m}(\mathbf{x}_s, G_{A_t}, n_{A_t}; \boldsymbol{\theta}))_+||.$$

Inference could be made by subsampling as discussed in Chernozhukov, Hong and Tamer (2007) and Ciliberto and Tamer (2009). The confidence set is

$$\hat{\Theta}_I = \{\boldsymbol{\theta} : \ S \cdot (Q_S(\boldsymbol{\theta}) - \min_k Q_S(k)) \leq c_\tau(\boldsymbol{\theta})\}.$$

4 Monte Carlo Study

In this section, we conduct a sequence of Monte Carlo experiments to study finite sample properties of our subnetwork approach. For simplicity, we make inference on one parameter. The utility function contains the interaction effect only,

$$u(y_i, \mathbf{x}_i, y_{-i}; \boldsymbol{\beta}, \gamma) = \gamma \frac{\sum_{j \in V} g_{ij} y_j}{\sum_{j \in V} g_{ij}} + \epsilon_i.$$

[8] Clearly our approach offers considerable advantage when allowing the individual shocks to be dependent, as what is required here to compute is the multiple integral in $\Pr(\boldsymbol{\epsilon} \in H(y_A; \mathbf{x}, G_A, n_A, \boldsymbol{\theta}))$ with the dimension equal to the number of agents in the subnetwork. If the number of agents is less than or equal to 3, the integral is straighforward to calculate. Otherwise, one can use the GHK simulator to approximate the integral (see, e.g. Geweke (1991), Bőrsch-Supan and Hajivassiliou (1993), and Keane (1994)).

Table 1: Upper Bounds and Parameter Values

Outcome	Emp. Prob.	Upper Bound						
(y_1, y_2)	$P(y_A)$	$\gamma = 0.1$	$\gamma = 0.5$	$\gamma = 1$	$\gamma = 1.5$	$\gamma = 2$	$\gamma = 3$	$\gamma = 5$
$(1,1)$	0.642	0.291	0.478	0.708	0.870	0.955	0.997	1
$(0,1)$	0.102	0.240	0.201	0.154	0.113	0.080	0.033	0.003
$(1,0)$	0.100	0.239	0.185	0.110	0.052	0.020	0.001	0
$(0,0)$	0.155	0.250	0.250	0.250	0.250	0.250	0.250	0.250

The true parameter is $\gamma_0 = 1$. We set the number of networks to be 1000. For each network, we generate individual shocks ϵ and calculate its set of equilibria. If the game has multiple equilibria, each one is selected with equal probability. Inference is based on a 80% confidence interval of 200 re-samples. Upper bounds are approximated by 1000 simulations.

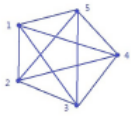
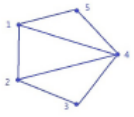
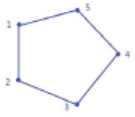
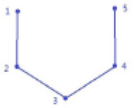
First, we find that upper bounds on outcomes of subnetwork are informative. Table 1 collects the empirical probabilities of outcomes of subnetwork $\{1,2\}$, and upper bounds calculated by different hypothetical values of parameter γ in a 4-person network described in Figure 1. The numbers in blue show cases where moment inequalities are violated. When γ is too low, the upper bound predicts too small share of outcome $(y_1, y_2) = (1,1)$; when γ is too large, it predicts too small share of outcomes $(y_1, y_2) = (0,1)$ and $(y_1, y_2) = (1,0)$. Only when the value is near the true value $\gamma_0 = 1$ all the moment inequalities are satisfied.

The second sets of example show the impact of the structure of full network on the performance of our approach. Table 2 collects the identified set of four networks using four choices of subnetworks. From the left to the right, we generate full networks with decreasing number of edges. The performance of our approach improves when there are fewer links connecting individuals inside and outside subnetworks. For example, in the first network, agent 1 is connected with all others in the network, but in the last network, agent 1 is connected to agent 2 only. When we construct upper bounds for subnetwork $\{1,2\}$, we relax our upper bounds more in the first network because we ignore many connections. Our upper

bounds are tighter if the network is sparse.

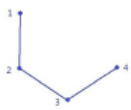
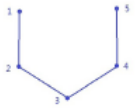
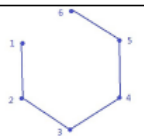
Table 2 also illustrates the relationship between the choice of subnetwork and the size of the confidence interval. From the top to the bottom, we increase the size of subnetworks. When the size of subnetwork increases, the confidence interval shrinks. This is because more information is used to construct moment conditions. However, it is not the case that the performance is the best if we make inference based on the full-subnetwork, which has the largest number of agents. This is because with a large number of outcomes, the probability of each individual outcome may not be precisely estimated for a fixed sample size.

Table 2: Identified Set of Varying Full and Sub Networks

Subnetwork	Identification Set			
				
{1,2}	(0.834, 3.001)	(0.716, 3.001)	(0.680, 2.669)	(0.771, 1.601)
{1,2,3}	(0.823, 2.987)	(0.737, 3.001)	(0.758, 1.659)	(0.780, 1.337)
{1,2,3,4}	(0.830, 1.861)	(0.832, 1.578)	(0.812, 1.612)	(0.877, 1.240)
{1,2,3,4,5}	(1.027, 1.037)	(0.900, 0.903)	(0.827, 1.210)	(0.911, 1.001)

The last sets of examples show the performance of our approach when the size of full network increases. As shown in Table 3, for a given choice of subnetwork, confidence interval increases slightly with the size of full network, but the magnitude is small. This is because for a given choice of subnetwork, the links whose end points are outside the subnetwork do not affect moment conditions. The moment conditions are still informative even if the network is of large size.

Table 3: Full Network Size and Confidence Interval

Subnetwork	Identification Set			
				
{1,2}	(0.743, 1.276)	(0.712, 1.605)	(0.771, 1.601)	(0.734, 1.678)
{1,2,3}	(0.899, 1.081)	(0.834, 1.410)	(0.780, 1.337)	(0.846, 1.391)
{1,2,3,4}		(0.865, 1.136)	(0.877, 1.240)	(0.904, 1.350)

5 Application

In this application, we study peer effects on smoking using the data from the National Longitudinal Study of Adolescent to Adult Health (Add Health). Add Health is a nationwide longitudinal survey of adolescent health. We use Wave I in Home Survey, which collects social, economic and physical information of teenagers from grades 7 to 12 in year 1993 and 1994. Add Health is one of the most commonly used data set in studying peer effects (e.g. Gaviria and Raphael 2001 on juvenile behavior, Calvó-Armengol, Patacchini and Zenou 2009 on education, Trogdon, Nonnemaker and Pais 2008 on overweight, Cohen-Cole and Fletcher 2008 on obesity). What is special about Add Health is that respondents are asked to nominate their friends. The information on family, social background, activities together with friend nomination provides a unique opportunity to study interactions among friends controlling for other social and economic influences.

Table 4 reports summary statistics of our key variables. There are a total of 20745 observations in the survey. The key dependent variable, smoke, is an indicator of whether a correspondent smokes on a regularly basis. On average 26% of the respondents are smokers. The control variables include age, gender, race, parents' smoking behavior and household income. We further include a dummy variable indicating whether a person's school has prevention programs for smoking.

Table 4: Summary Statistics

	Full Sample		Regression Sample	
	Mean	Stdev	Mean	Stdev
Obs	20,745		8,865	
Smoke	0.26	0.44	0.26	0.44
Grade	9.67	1.64	9.73	1.63
log(Income)	3.51	0.82	3.57	0.82
Gender	0.49	0.50	0.49	0.50
Race	0.72	0.45	0.62	0.49
Parent_smoke	0.26	0.44	0.24	0.43
School_program	0.92	0.27	0.93	0.26
NumMF	0.80	1.12		
NumFF	0.85	1.13		

The goal of our study is to identify factors that influence smoking decision, especially the peer effects. Our model is as follows

$$y_i = 1(\beta \mathbf{x}_i + \gamma \frac{\sum_{j \in V} g_{ij} y_j}{\sum_{j \in V} g_{ij}} + \epsilon_i),$$

where V is the set of all individuals in the data set, $\epsilon_i \sim i.i.d.\ N(0,1)$. $y_i = 1$ if person i smokes, $y_i = 0$ otherwise. $\mathbf{x}_i$ are control variables. We set $g_{ij} = g_{ji} = 1$ for $\{i, j\} \in V$ as long as one of persons i and j nominates the other as a friend.

Our estimation starts with a naive Probit estimation of smoking. The sample contains individuals that have at least one friend. Coefficients, standard errors and marginal effects at mean are reported in Table 5. People are more likely to smoke if their peers or parents smoke. The rate of smoking also differs by gender and race. Because in the Probit regression, the coefficient for grade, parents' smoking, and race are significant, we include these three control variables in our partial identification approach.

Table 5: Probit Estimation of Smoking

	(1)			(2)		
	Coef.	Std.Err.	Marg.Eff.	Coef.	Std.Err.	Marg.Eff.
Obs	6,421			8,865		
Peer_effect	1.259***	0.057	0.358	1.186***	0.048	0.338
Birth_year	−0.013	0.010	−0.004			
Grade	0.064***	0.015	0.020	0.076***	0.010	0.022
School_program	0.043	0.070	0.012			
Gender	0.022	0.036	0.006			
Log_income	−0.004	0.023	−0.001			
Race	0.420***	0.041	0.119	0.430***	0.033	0.122
Parent_smoke	0.248	0.042	0.070	0.240***	0.035	0.068
Constant	−0.930	0.937		−2.022***	0.101	

Results from the Probit estimation show that the marginal effect of peer effects is 30%, which means if half of a person's friends smoke, that person's probability of smoking would increase by 15%. However, Probit estimation could be misleading because it treats friend's smoking behavior exogenous. If there is indeed peer effects on smoking, an unobserved factor that influences person i's behavior can be correlated with the behavior of his/her friends', because his/her friends' actions are affected by his/her behavior via peer effects. To deal with the endogeneity problem of friends' actions, we next turn to our partial identification approach.

We make inference using subnetwork of size 2 for computational tractability. The sparsity of the data set also supports our choice of using small subnetworks. The last two rows of Table 4 show that the average number of male and female friends nominated by an individual is 0.80 and 0.85, respectively. Table 5 reports more information about friend nomination. About 2/3 of respondents nominate zero of one friend. 83% of people nominate two friends

Table 6: Number of Nominated Friends

	Obs	Frequency
NumF = 0	6045	29.13%
NumF = 1	6738	32.48%
NumF = 2	4453	21.47%
NumF $\leq$2	17206	83.09%

or fewer. This suggests that each individual only connects to a few people. Inference based on small subnetworks is therefore informative.

The moment conditions are tested by 11,492 pairs of agents that are friends. For subnetworks of 2 agents, there are two adjacency matrix G_A, either the two connects or disconnects. The upper bounds are higher if agents are not connected because the percentage of fiends outside subnetwork is larger. Moment inequalities are more likely to be violated for pairs of agents who are connected. Due to computational concerns, we only check inequalities for pairs of agents that are friends. This is equivalent to checking the conditional moment inequalities given a fixed G_A.

The final result is reported in Table 7. The confidence interval of peer effects estimated by the structural approach is (0.872, 1.493). This suggests positive and significant peer effects on smoking. Though the subnetwork approach concludes larger confidence intervals as compared to Probit estimation, it produces more convincing results because it takes into account the endogeneity of friends ' actions due to social interaction. It should also be noted that the criterion function evaluated at the Probit estimator is 0.016, while the minimum of criterion function is 0.003. Probit estimator lies outside the confidence set, and is therefore unlikely to be the true parameter of the data generating process.

Table 7: Partial Identification of Peer Effects in Friend Network

	Confidence Interval
Peer_effect	(0.872,1.493)
Grade	(-0.018,0.104)
Parent_smoke	(-0.150,0.290)
Race	(-0.001,0.391)
Constant	(-2.145,-1.027)

6 Conclusion

This paper studies identification and estimation of peer effects on binary choices in social networks. Our framework belongs to discrete games of complete information. As is well-known in the entry literature, identifying discrete games is difficult in general because the model often yields multiple equilibria. The inherited network feature of our model makes identification and estimation even more challenging.

The existing econometric methods that rely on the choice probabilities of the full game are not feasible in our case because we consider networks that are large and have varying friendship relationships. Not only does the number of outcomes grow exponentially, the number of observations of the same network structure is not sufficient to construct moment conditions. Therefore, we seek alternative moment conditions that can be consistently estimated.

The novelty of our identification strategy is the use of subnetworks. A subnetwork is a collection of agents, whose actions depend on actions of their friends, both inside and outside the subnetwork. Because we seek conditions that hold regardless the behavior of the agents outside the subnetwork, the bound of an outcome of a subnetwork is constructed by replacing the actions of the agents outside the subnetwork but connected to an individual inside by this individual's action. Because peer effects are positive, such bound gives the highest probability of observing an outcome. We therefore get a set of moment inequalities that could be used to partially identify the model.

Our estimation strategy is closely related to Ciliberto and Tamer (2009). The criterion function penalizes a parameter if upper bounds predicted by the parameters are less than the empirical choice probabilities of the subnetwork. The identified set is the set of parameters whose criterion function is less than a threshold, calculated by subsampling.

The Monte Carlo examples presented in this paper study the performance of our approach. The subnetwork approach is able to provide an informative inference on parameters of the model, especially when the network is sparse. We apply our identification and estimation strategy to study peer effects on smoking using the data from the National Longitudinal Study of Adolescent to Adult Health. The identified set suggests positive and significant peer effects on smoking.

The moment conditions used in this paper are only a small part of conditions implied by the model. The identification set could shrink further if more informative and easily verifiable conditions are considered. Sharp identification based on our approach is left for future research.

Appendix

Proof of Theorem 1: For a given subnetwork A, let y_A and y_{-A} denote the actions of all players inside and outside A. Define

$$R(y_A, y_{-A}|\mathbf{x}, G_A, n_A, \boldsymbol{\theta}) := \{\epsilon \in \mathbb{R}^N : \ (2 \cdot y_i - 1) \cdot (\boldsymbol{\beta}\mathbf{x}_i + \gamma \frac{\sum_{j \in V} g_{ij} y_j}{\sum_{j \in V} g_{ij}} + \epsilon_i) \geq 0 \ , \forall i \in A\}$$

A necessary condition for (y_A, y_{-A}) being an outcome of the game is that $\epsilon \in R(y_A, y_{-A}|\mathbf{x}, G_A, n_A, \boldsymbol{\theta})$. Otherwise players in A don't play best response to the actions of other players, therefore (y_A, y_{-A}) is not a Nash equilibrium of the game.

On the other hand, for $i \in A$ such that $y_i = 0$, we have $y_i \leq y_j$ for $\forall j \in V$, therefore

$$\boldsymbol{\beta}\mathbf{x}_i + \gamma \frac{\sum_{j \in V} g_{ij}[1(j \in A) \cdot y_j + 1(j \notin A) \cdot y_i]}{\sum_{j \in V} g_{ij}} + \epsilon_i \leq \boldsymbol{\beta}\mathbf{x}_i + \gamma \frac{\sum_{j \in V} g_{ij} \cdot y_j}{\sum_{j \in V} g_{ij}} + \epsilon_i.$$

When $y_i = 0$, $2 \cdot y_i - 1 = -1$. The above inequality is equivalent to

$$(2 \cdot y_i - 1) \cdot (\boldsymbol{\beta}\mathbf{x}_i + \gamma \frac{\sum_{j \in V} g_{ij}[1(j \in A) \cdot y_j + 1(j \notin A) \cdot y_i]}{\sum_{j \in V} g_{ij}} + \epsilon_i) \geq (2 \cdot y_i - 1)(\boldsymbol{\beta}\mathbf{x}_i + \gamma \frac{\sum_{j \in V} g_{ij} \cdot y_j}{\sum_{j \in V} g_{ij}} + \epsilon_i). \tag{10}$$

Similarly, for $i \in A$ such that $y_i = 1$, we have $y_i \geq y_j$ for $\forall j$, therefore

$$\boldsymbol{\beta}\mathbf{x}_i + \gamma \frac{\sum_{j \in V} g_{ij}[1(j \in A) \cdot y_j + 1(j \notin A) \cdot y_i]}{\sum_{j \in V} g_{ij}} + \epsilon_i \geq \boldsymbol{\beta}\mathbf{x}_i + \gamma \frac{\sum_{j \in V} g_{ij} \cdot y_j}{\sum_{j \in V \cdot} g_{ij}} + \epsilon_i,$$

which is equivalent to

$$(2 \cdot y_i - 1) \cdot (\boldsymbol{\beta}\mathbf{x}_i + \gamma \frac{\sum_{j \in V} g_{ij}[1(j \in A) \cdot y_j + 1(j \notin A) \cdot y_i]}{\sum_{j \in V} g_{ij}} + \epsilon_i) \geq (2 \cdot y_i - 1)(\boldsymbol{\beta}\mathbf{x}_i + \gamma \frac{\sum_{j \in V} g_{ij} y_j}{\sum_{j \in V} g_{ij}} + \epsilon_i). \tag{11}$$

Note the left hand sides of inequalities (10) and (11) appear in the definition of $H(y_A; \mathbf{x}, G_A, n_A, \boldsymbol{\theta})$. The right hand sides appear in $R(y_A, y_{-A}|\mathbf{x}, G_A, n_A, \boldsymbol{\theta})$. If $\epsilon \in R(y_A, y_{-A}|\mathbf{x}, G_A, n_A, \boldsymbol{\theta})$, the left hands of inequalities (10) and (11) are greater than 0, therefore the left hands are greater than 0. Hence $\epsilon \in H(y_A; \mathbf{x}, G_A, n_A, \boldsymbol{\theta})$. In other words,

$$R(y_A, y_{-A}|\mathbf{x}, G_A, n_A, \boldsymbol{\theta}) \subset H(y_A; \mathbf{x}, G_A, n_A, \boldsymbol{\theta}).$$

If $\epsilon \notin H(y_A; \mathbf{x}, G_A, n_A, \boldsymbol{\theta})$, $\epsilon \notin R(y_A, y_{-A}|\mathbf{x}, G_A, n_A, , \boldsymbol{\theta})$ for $\forall y_{-A}$. As a consequence, there doesn't exist y_{-A} such that (y_A, y_{-A}) is an outcome of the game. Therefore, y_A cannot be observed. This suggests that

$$\Pr(\epsilon \notin H(y_A; \mathbf{x}, G_A, n_A, \boldsymbol{\theta})) \leq 1 - \Pr(y_A \mathbf{x}, G_A, n_A, \boldsymbol{\theta}),$$

or equivalently,

$$\Pr(y_A|\mathbf{x}, G_A, n_A, \theta) \leq \Pr(\epsilon \in H(y_A; \mathbf{x}, G_A, n_A, \boldsymbol{\theta})).$$

References

Andrews, D. W., S. Berry, and P. Jia (2004). Confidence regions for parameters in discrete games with multiple equilibria, with an application to discount chain store location. *Working Paper, Yale University*.

Badev, A. (2013). Discrete games in endogenous networks: Theory and policy. *Working Paper*.

Bajari, P., H. Hong, and S. P. Ryan (2010). Identification and estimation of discrete games of complete information. *Econometrica 78*(5), 1529–1568.

Beresteanu, A., I. Molchanov, and F. Molinari (2011). Sharp identification regions in models with convex moment predictions. *Econometrica 79*(6), 1785–1821.

Berry, S. T. (1992). Estimation of a model of entry in the airline industry. *Econometrica 60*(4), 889–917.

Bjorn, P. A. and Q. H. Vuong (1984). Simultaneous equations models for dummy endogenous variables: a game theoretic formulation with an application to labor force participation. *Working Paper, California Institute of Technology*.

Bramoullé, Y., H. Djebbari, and B. Fortin (2009). Identification of peer effects through social networks. *Journal of Econometrics 150*(1), 41–55.

Bresnahan, T. F. and P. C. Reiss (1990). Entry in monopoly market. *The Review of Economic Studies 57*(4), 531–553.

Bresnahan, T. F. and P. C. Reiss (1991a). Empirical models of discrete games. *Journal of Econometrics 48*(1), 57–81.

Bresnahan, T. F. and P. C. Reiss (1991b). Entry and competition in concentrated markets. *Journal of Political Economy 99*(5), 977–1009.

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

Brock, W. A. and S. N. Durlauf (2007). Identification of binary choice models with social interactions. *Journal of Econometrics 140*(1), 52–75.

Calvó-Armengol, A. and M. O. Jackson (2004). The effects of social networks on employment and inequality. *American Economic Review 94*(3), 426–454.

Calvó-Armengol, A., E. Patacchini, and Y. Zenou (2009). Peer effects and social networks in education. *Review of Economic Studies 76*(4), 1239–1267.

Chernozhukov, V., H. Hong, and E. Tamer (2007). Estimation and confidence regions for parameter sets in econometric models1. *Econometrica 75*(5), 1243–1284.

Ciliberto, F. and E. Tamer (2009). Market structure and multiple equilibria in airline markets. *Econometrica 77*(6), 1791–1828.

Clark, A. E. and Y. Loheac (2007). It wasn't me, it was them! social influence in risky behavior by adolescents. *Journal of Health Economics 26*(4), 763–784.

Cohen-Cole, E. and J. M. Fletcher (2008). Is obesity contagious? social networks vs. environmental factors in the obesity epidemic. *Journal of Health Economics 27*(5), 1382–1387.

Conley, T. G. and C. R. Udry (2010). Learning about a new technology: Pineapple in ghana. *The American Economic Review 100*(1), 35–69.

de Paula, A. (2009). Inference in a synchronization game with social interactions. *Journal of Econometrics 148*(1), 56–71.

de Paula, A. (2013). Econometric analysis of games with multiple equilibria. *Annual Review of Economics 5*(1), 107–131.

Easley, D. and J. Kleinberg (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press.

Galichon, A. and M. Henry (2011). Set identification in models with multiple equilibria. *Review of Economic Studies 78*(4), 1264–1298.

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

Geweke, J. (1991). Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints and the evaluation of constraint probabilities. In *Computing science and statistics: Proceedings of the 23rd symposium on the interface*, pp. 571–578. Citeseer.

Graham, B. S. (2014). An empirical model of network formation: detecting homophily when agents are heterogenous. *Working Paper, UC Berkeley*.

Henry, M., R. Méango, and M. Queyranne (2015). Combinatorial approach to inference in partially identified incomplete structural models. *Quantitative Economics 6*(2), 499–529.

Jia, P. (2008). What happens when wal-mart comes to town: An empirical analysis of the discount retailing industry. *Econometrica 76*(6), 1263–1316.

Keane, M. P. (1994). A computationally practical simulation estimator for panel data. *Econometrica 62*(1), 95–116.

Krauth, B. V. (2006). Simulation-based estimation of peer effects. *Journal of Econometrics 133*(1), 243–271.

Leung, M. (2015). Two-step estimation of network-formation models with incomplete information. *Journal of Econometrics 188*(1), 182–195.

Liu, N. and H. Xu (2015). Semiparametric inference on social interactions with homophily. *Working Paper, UT Austin*.

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

McPherson, M., L. Smith-Lovin, and J. M. Cook (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 415–444.

Nakajima, R. (2007). Measuring peer effects on youth smoking behaviour. *Review of Economic Studies 74*(3), 897–935.

Narayanan, S. (2013). Bayesian estimation of discrete games of complete information. *Quantitative Marketing and Economics 11*(1), 39–81.

Sheng, S. (2012). Identification and estimation of network formation games. *Working Paper*.

Soetevent, A. R. and P. Kooreman (2007). A discrete-choice model with social interactions: with an application to high school teen behavior. *Journal of Applied Econometrics 22*(3), 599–624.

Tamer, E. (2003). Incomplete simultaneous discrete response model with multiple equilibria. *Review of Economic Studies 70*(1), 147–165.

Trogdon, J. G., J. Nonnemaker, and J. Pais (2008). Peer effects in adolescent overweight. *Journal of Health Economics 27*(5), 1388–1399.

Uetake, K. (2012). Strategic network formation and performance in the venture capital industry. *Working Paper*.

Zimmerman, D. J. (2003). Peer effects in academic outcomes: Evidence from a natural experiment. *Review of Economics and Statistics 85*(1), 9–23.