

Identification of Local Interaction Models with Imperfect Location Data

Timothy G. Conley, University of Chicago
Giorgio Topa, New York University*

February 20, 2002

Abstract

This paper considers the problem of estimating a local interaction model defined at the level of individual agents, in the absence of perfect information about agent locations in the relevant socio-economic space. We consider two types of data limitations: one in which individual locations are measured with error, the other in which location information is correct but only up to some spatial region (e.g., Census tracts or zip codes).

Our results indicate that the model parameters are not identified almost surely in the former situation, whereas identification is preserved in the latter. We also suggest a potential estimation strategy that can yield useful information about the likely range of parameters even when point identification fails.

JEL: J64, R12, C21.

Keywords: Local interactions, economic distance, spatial aggregation, unemployment.

*We would like to thank the Editor, an anonymous referee, Alberto Bisin, Adeline Delavande, Donghoon Lee, Efe Ok, Vincenzo Quadrini, and seminar participants at Minnesota and Wisconsin for very useful comments and suggestions. Topa gratefully acknowledges financial support from the NSF and the C.V. Starr Center for Applied Economics at NYU. Conley gratefully acknowledges support from the NSF. The authors are of course responsible for all errors. Corresponding author: Giorgio Topa, Department of Economics, New York University, 269 Mercer Street, NY, NY 10003. *Giorgio.Topa@nyu.edu*.

1 Introduction

This paper is concerned with estimation of local interaction models without access to perfect information about individuals' locations. In estimating local interaction models, one typically needs to have some information about agents' locations in a socio-economic space in order to determine who is likely to interact with whom.¹ However, it is often the case that such data are imperfect: either locations are measured with error or only imprecise information is available, for example locations are known only up to census tracts or zip codes. Therefore, it is important to determine under what conditions one can still identify the model parameters in the presence of these limitations in the available data.² We examine identification in the context of a particular local interaction model, with location information that is error-ridden and with imprecise location information. We then suggest a potential estimation strategy.

We consider two types of an interaction model called a contact process. First we examine a simple model where binary responses depend on neighbors' outcomes and then a straightforward generalization that allows for individuals' characteristics to impact responses as well. We are directly motivated by an interest in models of the labor market where agents may find jobs thanks to tips or referrals by their social contacts (see e.g. Montgomery (1991) or Topa (2001)). Therefore we present the model referring to the binary outcomes as reflecting employment status. However, the analysis is directly relevant for many other applications including models of social pathologies (such as crime, drug use, or teenage pregnancy); peer effects in education; cultural transmission and the evolution of preferences.³

We begin by defining a dynamic local interaction model with agents arranged on an integer lattice, where each agent's outcome is a binary variable, that generates a continuous time Markov process. We discuss existence and uniqueness of a stationary distribution as well as a few of its properties. We then show that the model parameters are identified from a set of simple cross-sectional moment conditions, if one has access to data on individuals' outcomes and their exact locations. These moment conditions are a set of linear 'balance equations' for the flows into and out of specific employment patterns, that must hold at the stationary distribution.

Two particular types of imperfect location information are then considered. The first type occurs when agents' locations are mismeasured. We model this location distortion as being a random map from the set of agents' integer locations to a (potentially) new set of integer locations. This type of distortion of locations will be observationally equivalent to a transformation of the stationary joint distribution of agents' states into a different distribution. We demonstrate that, under fairly general

¹Conley and Topa (forthcoming) discuss several socio-economic distance metrics that may be good proxies for the dimensions along which information networks among agents are constructed.

²Brock and Durlauf (forthcoming) present an excellent treatment of identification issues and estimation strategies for a general class of local interaction models.

³See, as a very incomplete list of examples, Bertrand et al. (2000), Glaeser et al. (1996), Goolsbee and Klenow (forthcoming).

conditions, the true parameters cannot be recovered from this resulting distribution using our particular set of moment conditions, almost surely with respect to the random map that describes the distortion process. We conjecture that this result holds more generally for any set of moment conditions derived using a balance equation approach.

The second type of imperfect location information occurs when agents' locations are known only up to a region. With binary outcomes for the agents and no covariates, this is equivalent to having only aggregate-level information on the number of people with each outcome within a region. We investigate whether the true parameters can be recovered from moments of these region aggregates. We provide a numerical characterization of the map between the interaction model parameters and the first two moments of aggregates for several region structures for an interesting subset of the parameter space. These results strongly suggest that the interaction parameters can be recovered from these moments, i.e. that the interaction process is locally identified for these types of regions. Of course this is not a proof that there are no non-trivial sets of parameters consistent with these moments. Therefore we briefly present an estimation strategy for the model parameters that would be useful even if they are not identified – there is a nonsingleton set of parameters consistent with the moments we specify. Test statistics from a standard GMM estimator can be used to estimate confidence sets for the structural parameter(s) that are consistent with the data.

The combination of our results suggests that researchers may be better off using more plausibly correct but imprecise measures of agents' locations if they are not confident about their definitions of agents' exact locations. For example, suppose data were available on workers' social contacts from a survey that detailed the network structure of many randomly selected individuals in an urban area. These data might provide a description of agents' links to each other, though one likely to contain errors or omissions. Rather than using the link information directly as a measure of relative locations in socio-economic space to estimate a model, it may be preferable to use say moments of Census tract aggregate outcomes in estimation. The link information may be best used only to define the most relevant geometry for the network.

The rest of the paper is organized as follows. Section 2 sets up the dynamic local interaction model and describes the properties of the invariant distribution. Section 3 demonstrates a method to recover model parameters when individual locations are known. Section 4 discusses identification with distorted location information. Section 5 reports our numerical exercises characterizing the map between model parameters and moments of region aggregates, and presents potential estimation strategies. Finally, conclusions and potential directions for future work are briefly discussed in Section 6.

2 Data Generation Model

We start by examining a model in which agents are homogeneous: in particular, the rate at which agents enter and exit unemployment is the same across agents. The model is a version of the contact process, which was introduced by Harris (1974) and has been extensively studied in the literature on interacting particle systems.⁴ There is a set $M \subseteq Z$ of agents, arranged at discrete locations on the line.⁵ Time flows continuously from 0 to ∞ in the model. The state of each agent i at time t , y_{it} , is her employment status: $y_{it} \in E \equiv \{1, 0\}$, where 1 denotes the employed state and 0 the unemployed state. Therefore, the state of the system at each point in time is a configuration of employment states $y \in \mathcal{Y} \equiv E^M$.

The dynamics of the model are specified by a collection of transition rates into and out of unemployment, defined as follows: any agent i switches from employed to unemployed at a normalized exponential rate of one, and makes the reverse transition from unemployment to employment at exponential rate $\gamma + \lambda \cdot \sum_{j:|i-j|=1} y_{jt}$. The scalars γ and λ are positive parameters. In other words, an agent is more likely to exit unemployment when more of her neighbors are currently employed. As in Topa (2001), the interpretation is that agents may find jobs thanks to tips or referrals by those among their social contacts who are presently employed (this is essentially an informal mutual insurance arrangement that can be sustained even under limited commitment). However, informal exchanges are not the only channel through which agents become employed: even if all her neighbors are unemployed, agent i can still become employed at rate γ .

The transition rates described above uniquely define a continuous time Markov process y_t with state space \mathcal{Y} over the set of locations. Liggett (1985) provides the details of the construction. Essentially, the key feature of the model that allows construction of the Markov process over the lattice is that the rate at which each site changes state only depends on the configuration of a *finite* number of other sites. It is possible to show that a stationary distribution exists and is unique, for any dimension $d \geq 1$.⁶

The stationary distribution has several useful properties (See Liggett (1999) for a thorough discussion). The data in the cross section are stationary and mixing and a central limit theorem (with \sqrt{n} convergence) is applicable due to Schonmann (1986). The covariance between the state of any pair of agents on the lattice $Cov(y_i, y_j)$ is non-negative, and is bounded above by a quantity that decays exponentially in the

⁴For an excellent treatment of this and other stochastic interacting systems, see Liggett (1985), (1999).

⁵In general, one can define a contact process on an integer lattice Z^d for any dimension $d \geq 1$. Later we will also consider two-dimensional examples.

⁶See Theorem 2.28, Ch. 6, in Liggett (1985) for the case in which agents are arranged on a line ($d = 1$); Bezuidenhout and Grimmett (1990) present a proof for all $d \geq 1$. When M is finite and time flows discretely, this result is straightforward – in particular, uniqueness derives from the fact that the transition matrix Q of the Markov process y_t is irreducible and aperiodic.

distance between agents.⁷

3 Identification with Perfect Location Information

We first consider identification of the structural parameters $\theta \equiv (\gamma, \lambda)$ when we perfectly observe individuals' outcomes and locations for a single cross section of agents. This is slightly complicated by the fact that we do not have a closed form solution for the stationary distribution. Therefore we do not frame our identification discussion in the context of a likelihood function, but rather we focus on identification given a set of moment conditions. Following Clifford and Sudbury (1979), we present a system of 'balance equations' that relate the probability of specific patterns of ones and zeros under the stationary distribution to the structural parameters.

Our definition of identification is that θ is identified if it is the unique solution over the entire parameter space of a set of moment equations. We will use the phrase local identification to refer to the case in which θ is the unique solution to a set of moment equations in a subset of the parameter space which includes θ .

We proceed by assuming we know the cross-sectional distribution of the process once it has reached its invariant distribution. Our demonstration of identification uses probabilities $\pi(\cdot)$ under the invariant distribution of sets of individual outcomes, such as the probability of a single individual being employed, the probability that two adjacent agents are both employed, and so on. At the invariant distribution, balance equations must hold that equate the flow into a specific configuration of ones and zeros to the flow out of that specific pattern. These equations define a system of linear equations in the unknown parameters θ and in the observed $\pi(\cdot)$.

Let $\pi(1)$ be the employment rate over the line, i.e. the fraction of ones (this is equivalent to the probability of being employed). Let $\pi(11)$ be the proportion of pairs of adjacent agents that are both employed. Let $\pi(10)$ be the proportion of pairs of adjacent agents where one is employed and the other is unemployed. Likewise, $\pi(100)$ is the proportion of strings of length three where only the first agent is employed and the other two are unemployed. Notice that because of the symmetry of the contact process, $\pi(10) = \pi(01)$, $\pi(100) = \pi(001)$, and so on.

By adapting an argument developed by Clifford and Sudbury (1979), we can show that at the stationary distribution the following balance equations must be satisfied:

$$[0] : \gamma\pi(0) + 2\lambda\pi(10) = \pi(1) \tag{1}$$

$$[00] : 2\gamma\pi(00) + 2\lambda\pi(100) = 2\pi(10) \tag{2}$$

The interpretation of these moment equations is very simple. For example, the left hand side of (2), also designated $[00]$, represents the rate at which strings of length two where both agents are unemployed disappear; the right hand side represents the

⁷See Liggett (1999), p. 7 and Theorem 2.30.

rate at which a [00] pattern appears. At the invariant distribution, the flow into [00] and the flow out of [00] must balance.

By using the identities $\pi(0) = 1 - \pi(1)$ and $\pi(00) = 1 - \pi(1) - \pi(10)$, equations (1) and (2) can be rewritten as

$$\begin{bmatrix} 1 - \pi(1) & 2\pi(10) \\ 1 - \pi(1) - \pi(10) & \pi(100) \end{bmatrix} \begin{bmatrix} \gamma \\ \lambda \end{bmatrix} = \begin{bmatrix} \pi(1) \\ \pi(10) \end{bmatrix}. \quad (3)$$

It is easy to show that the matrix in (3) is non-singular if $\pi(00) > 0$, $\pi(10) > 0$, and either $\pi(101) > 0$ or $\pi(000) > 0$. The probabilities $\pi(1)$, $\pi(10)$ and $\pi(100)$ are recoverable from an entire cross section of individual level data. Therefore, the system of linear equations (3) has a unique solution for θ , so the two structural parameters are identified in this case.

Note that these equations do not exhaust all the information contained in the cross section. One can write down similar balance equations for any string of adjacent sites and for any pattern of ones and zeros. Therefore the model parameters are clearly over-identified.

3.1 Two-dimensional Case

The identification strategy discussed above can be easily extended to two-dimensional lattices, i.e. $M \subseteq Z^2$. Each agent location is now simply a vector of two integer coordinates. As before, if we have access to individual level data, we can use a pair of balance equations to identify the parameters θ . In particular, we can rewrite (1) and (2) as follows:

$$[0] : \gamma\pi(0) + 4\lambda\pi(10) = \pi(1) \quad (4)$$

$$[00] : 2\gamma\pi(00) + \lambda \left[2\pi(100) + 4\pi \begin{pmatrix} 1 & \\ 0 & 0 \end{pmatrix} \right] = 2\pi(10) \quad (5)$$

where $\pi \begin{pmatrix} 1 & \\ 0 & 0 \end{pmatrix}$ is the fraction of ‘‘L’’ shaped sets of three agents with one corner employed and the other sites unemployed. These balance equations give rise to the following system of linear equations in the unknown parameters θ and in the proportions $\pi(\cdot)$:

$$\begin{bmatrix} 1 - \pi(1) & 4\pi(10) \\ 1 - \pi(1) - \pi(10) & \pi(100) + 2\pi \begin{pmatrix} 1 & \\ 0 & 0 \end{pmatrix} \end{bmatrix} \begin{bmatrix} \gamma \\ \lambda \end{bmatrix} = \begin{bmatrix} \pi(1) \\ \pi(10) \end{bmatrix}, \quad (6)$$

which again has a unique solution for θ provided that the matrix in (6) is non-singular.

3.2 Addition of Covariates

In this subsection we wish to extend the previous analysis to the case in which agents differ in terms of several characteristics that may affect their transition rates into and out of employment (for simplicity, we return to the case of one-dimensional lattices). Formally, each agent i is indexed by a vector of characteristics X_i that are constant over time and have a discrete support.⁸ Such covariates may include (among others) education levels, other measures of human capital, occupation, race, ethnicity.

Agent characteristics are assumed to affect their entry rate into employment (again, we normalize the exit rate from employment to unity). In particular, an unemployed agent enters employment at rate $\gamma(X_i) + \lambda(X_i) \cdot \sum_{j:|i-j|=1} y_j$. Conditional on a given realization of X over the lattice, a unique stationary distribution of the Markov process y_t exists, as in the homogeneous case.

Assumption 1. *The distribution of X is isotropic, and the probability that pairs of adjacent agents have the same realization of X is strictly positive.*

The first part of this assumption ensures that $\pi(10) = \pi(01)$ or $\pi(100) = \pi(001)$, so that we can use balance equations similar to (3) in our identification argument; the second part ensures that we can condition on specific patterns of covariates to retrieve our model parameters.

A key feature of our model is that the transition rates of each agent i are only affected by her own covariates, and not by those of her neighbors. This simplifies the analysis a great deal. We assume that the $\gamma(\cdot)$, $\lambda(\cdot)$ functions are linear in X :

$$\begin{aligned}\gamma(X) &= \gamma_0 + \gamma_1 X_1 + \dots + \gamma_K X_K \\ \lambda(X) &= \lambda_0 + \lambda_1 X_1 + \dots + \lambda_K X_K\end{aligned}$$

In order to show that the structural parameters of the model are identified, the strategy is to use the same balance equations as in the case without covariates, but conditioning on specific patterns of characteristics. By appropriately choosing these conditioning sets, we can state the following Proposition.

Proposition 1. *Under Assumption 1, all the parameters in $\gamma(\cdot)$, $\lambda(\cdot)$ are identified.*

Proof: see the Appendix.

⁸These can include both individual characteristics and characteristics of the area in which i resides (e.g., the presence of parks and other amenities, the quality of the school district, or the crime rate in the area).

4 Locations Distorted by Measurement Error

We now consider the case in which individual agents' locations are not fully observed by the econometrician, but rather they are measured with error. We model this location distortion as a random map from the set of agents' integer locations to a (potentially) new set of integer locations. In other words, agents' locations may be mismeasured by any number of positions on the line (we focus here on one-dimensional lattices, although the same argument can be easily extended to sets of agents in Z^d , $d > 1$).

Let π denote a vector of true $\pi(\cdot)$ probabilities from the stationary distribution of the model. Let $\tilde{\pi}$ denote the corresponding vector of probabilities for the transformed locations. Then the relationship between true and mismeasured probabilities can be described as a linear mapping represented by a matrix A . First, consider probabilities involving pairs of agents. Then the typical element a_{kl} of A is defined as $a_{kl} = P(k|l)$: this is the proportion of agents measured to be at distance k , when they in fact reside at distance l from each other. For pairs of agents with outcomes of zero, one can write

$$\begin{bmatrix} \tilde{\pi}(00) \\ \tilde{\pi}(0, \cdot, 0) \\ \tilde{\pi}(0, \cdot, \cdot, 0) \\ \dots \end{bmatrix} = \begin{bmatrix} P(1|1) & P(1|2) & P(1|3) & \dots \\ P(2|1) & P(2|2) & P(2|3) & \dots \\ P(3|1) & P(3|2) & P(3|3) & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix} \begin{bmatrix} \pi(00) \\ \pi(0, \cdot, 0) \\ \pi(0, \cdot, \cdot, 0) \\ \dots \end{bmatrix}, \quad (7)$$

where $\pi(0, \cdot, 0) \equiv \pi(000) + \pi(010)$, $\pi(0, \cdot, \cdot, 0) \equiv \sum_x \sum_y \pi(0, x, y, 0)$, and so on.

The same matrix A also describes the mapping between π and $\tilde{\pi}$ for the other possible outcomes involving pairs of agents: $\pi(10)$, $\pi(01)$, and $\pi(11)$. Analogs of equation (7) hold for π 's corresponding to sets of more than two agents. We use the notation \mathcal{A} to denote the entire set of A and its higher-order analogs. We characterize the mapping between π and $\tilde{\pi}$ only up to this transformation, treating as equivalent any maps that result in the same \mathcal{A} .

As an illustration, suppose that agents' locations may be mismeasured by at most one position on the integer line. So agents that are in reality at distance three from each other may appear to the researcher to be as close as one unit apart (i.e., they are adjacent), or as far as five units apart. Then, for π probabilities involving pairs of agents, the matrix A becomes band-diagonal:

$$A = \begin{bmatrix} P(1|1) & P(1|2) & P(1|3) & 0 & \dots & \dots \\ P(2|1) & P(2|2) & P(2|3) & P(2|4) & 0 & \dots \\ P(3|1) & P(3|2) & P(3|3) & P(3|4) & P(3|5) & 0 \\ 0 & P(4|2) & P(4|3) & P(4|4) & P(4|5) & \dots \\ \dots & 0 & P(5|3) & P(5|4) & P(5|5) & \dots \\ \dots & \dots & 0 & P(6|4) & P(6|5) & \dots \\ \dots & \dots & \dots & 0 & P(7|5) & \dots \\ \dots & \dots & \dots & \dots & 0 & \dots \end{bmatrix}$$

We assume that the mapping described by \mathcal{A} is a random draw from a distribution over potential maps from π into $\tilde{\pi}$. Define as $f_{kl}^s(\cdot)$ the marginal pdf for each nonzero element a_{kl}^s of every $A_s \in \mathcal{A}$ (obviously the support of a_{kl}^s is $[0, 1]$). We make the following assumption.

Assumption 2. *Each marginal density $f_{kl}^s, \forall k, l, s$ does not exhibit any point masses on the interval $[0, 1]$. In other words, we require f_{kl}^s to be non-atomic.*

Consider the identification argument in Section 3. The true parameter vector $\theta_0 \equiv (\gamma_0, \lambda_0)$ is identified from the system of equations (3), exploiting balance equations (1) - (2). Let us now replace the correct $\pi(\cdot)$ probabilities with the mismeasured $\tilde{\pi}(\cdot)$:

$$\begin{bmatrix} 1 - \tilde{\pi}(1) & 2\tilde{\pi}(10) \\ 1 - \tilde{\pi}(1) - \tilde{\pi}(10) & \tilde{\pi}(100) \end{bmatrix} \begin{bmatrix} \gamma \\ \lambda \end{bmatrix} = \begin{bmatrix} \tilde{\pi}(1) \\ \tilde{\pi}(10) \end{bmatrix}. \quad (8)$$

We immediately obtain the following Proposition.

Proposition 2. *Under Assumption 2, the solution of system (8) is $\tilde{\theta} \neq \theta_0$ almost surely w.r.t. the probability measure over the random maps \mathcal{A} .*

Proof: Geometrically, (8) defines the intersection of two lines in \mathfrak{R}^2 . As we change the $\pi(\cdot)$ probabilities from π to $\tilde{\pi}$, the new intersection $\tilde{\theta}$ will be different than θ_0 almost surely, as long as the densities f_{kl}^s do not have any point masses, which is guaranteed by Assumption 2. ■

Therefore, under mild conditions, the true parameter values θ_0 cannot be recovered from the moment conditions in (8), using the mismeasured $\tilde{\pi}$. We conjecture that this result holds more generally for any set of moment conditions derived using a balance equation approach.

5 Locations Known only up to Region

In this Section we wish to address the issue of whether or not we can identify the model parameters if we only have access to imprecise but correct locational information. In other words, we only know the location of agents up to some geographic region like zip codes or Census tracts. We assume we can correctly assign agents to the region in which they reside, although we do not observe their specific location within the region.

For simplicity, we consider the model without covariates and for three region structures, two in which agents are on the line and one in which agents live on the plane. On the line, we investigate a regular structure where regions are segments with the the same number of agents each (e.g., [1 2], [3 4], [5 6] ...). For brevity we will only

discuss an illustration corresponding to two-agent regions. Qualitatively identical results obtain for size four, five, ten and twenty regions, leading us to conjecture that they will hold for arbitrarily sized regular regions. Also on the line, we study an irregular structure where agents are grouped as follows: [1], [2], [3 4], [5], [6], [7 8], [9], [10], [11 12] ... On the plane, we study square regions of four agents each.

We have not been able to develop an analytic demonstration of identification in these cases. Therefore, we resort to numerical methods to illustrate the map between model parameters and first and second moments of the number of employed agents in a region. In a region m , let $Y_{m,t}$ denote the total number of employed agents, i.e. $Y_{m,t} = \sum_{i \in m} y_{it}$. The strategy is to numerically approximate the first two moments of $Y_{m,t}$ for a grid of values of θ in a reasonable range, plot these surfaces, and study the sets of values θ that give rise to a given pair of values for the specified moments. Local identification is attained if these sets are singletons for any chosen value of the moments.

In practice, we simulate the model in discrete time, on both one- and two-dimensional lattices. On the line, we take a string of $N = 400$ agents, with the neighbor structure specified in Section 2: each agent at position i has two neighbors at positions $i - 1$ and $i + 1$. The lattice is a torus, i.e. agent 1 has neighbors at positions N and 2, whereas agent N has neighbors $N - 1$ and 1. In two dimensions, we take a square finite integer lattice of size 20×20 (again, $N = 400$), with the same neighbor structure: each agent (i_1, i_2) now has four neighbors at locations $(i_1, i_2 \pm 1)$, $(i_1 \pm 1, i_2)$.⁹

Since the model is simulated in discrete time, the process becomes a first order Markov chain with the following transition probabilities for each agent:

$$\Pr(y_{i,t+1} = 0 | y_{it} = 1) = \delta \tag{9}$$

$$\Pr(y_{i,t+1} = 1 | y_{it} = 0; y_{-i,t}) = \delta \left(\gamma + \lambda \sum_{j \in N_i} y_{jt} \right) \tag{10}$$

where N_i denotes the set of neighbors for agent i . For simplicity, we normalize δ to be $\delta = 0.1$. We then simulate the model for the following grid of values for θ : $\gamma \in [0.1, 1.3]$ with step size 0.03, $\lambda \in [0.5, 3.5]$ with step size 0.05.¹⁰

For each value of the parameters we simulate the process for 60 iterations, to attempt to reach the invariant distribution. We then record the configuration $\{y_{it}\}$ over the lattice at the level of individual agents, as well as the configuration of employment

⁹Again, the lattice is a torus, so for example agent $(1, 1)$ has neighbors $(1, 2)$, $(1, 20)$, $(2, 1)$, $(20, 1)$.

¹⁰This grid covers a reasonable range of probabilities of staying employed (or unemployed) for a certain number of periods. Given δ , the probability of staying employed for four periods is 0.66. On the other hand, the probability of remaining unemployed for four periods ranges from 0.96 (with both γ and λ at their lower bounds and no employed neighbors) to about zero (when both γ and λ are at their upper bounds, and all neighbors are employed). These probabilities seem sensible if a period represents say three to six months.

totals $\{Y_{m,t}\}$ for the regions. We can then compute the mean and variance of $Y_{m,t}$ across regions. The population moments for a given choice of θ are approximated by taking the average over 10,000 repetitions of the above procedure.

A representative selection of the simulation results is displayed in Figures 1 and 2. Figure 1 reports surface plots for the mean and variance of $Y_{m,t}$ for the three region structures described above: the first row is for two agent regions on the line, the second row is for the irregular regional structure on the line; the third row is for size four square regions in two dimensions. Figure 2 reports the corresponding contour plots for the same moments.

Although the hook-shaped contour lines for the variance suggest that there may be scope for multiple intersections with the contour lines for the mean, in fact the slopes are such that the intersections between the two sets of contour lines are singletons over our grid. Therefore, the parameters are locally identified using the first and second moments of $Y_{m,t}$. The same result holds true for any of the regional structures listed above, and in both one and two dimensions. Thus, numerical simulations are strongly suggestive that local identification is attained even when information about agents' locations is available only at the level of regions.¹¹

5.1 Potential Estimation Strategy

In view of the fact that this investigation falls short of a proof that the model parameters are identified using the first two moments of $Y_{m,t}$, we are motivated to briefly discuss a strategy for obtaining confidence intervals for the model parameters that remains valid even if identification fails. Let R_N be a vector containing the sample mean and variance of $Y_{m,t}$ for a sample of size N and $m(\theta_0)$ be the vector of corresponding population moments for the contact process with parameters θ_0 (in practice $m(\cdot)$ would have to be numerically approximated as above). Defining $h_N(\theta) = m(\theta) - R_N$, the value of θ that minimizes $\|h_N(\theta)\|$ is an obvious candidate estimator of θ_0 . When θ_0 is identified by these moments, this estimator will be consistent for θ_0 .

A method of moments approach can still be useful even if there is not point identification, i.e. there are other values of θ in addition to θ_0 that generate the same values for the expectation and variance of $Y_{m,t}$. The strategy of 'inverting' a quadratic form test statistic will still yield valid confidence intervals even if point identification fails. Schonmann (1986) provides a central limit result for cross sectional data from this process, so we know that $h_N(\theta_0)$ will have a limiting distribution when properly normalized:

$$\sqrt{N}h_N(\theta_0) \xrightarrow{d} N(0, V).$$

¹¹Of course, when using individual data, mean and variance alone are not sufficient to identify (γ, λ) , since the outcome variable y_t is a binary variable. In this case, an additional moment is needed, such as the correlation between agents at distance one.

The matrix V can be consistently estimated with its sample analog \widehat{V}_N , and we can construct the quadratic form:

$$J_N(\theta) = \sqrt{N}h_N(\theta)^\top \widehat{V}_N^{-1} \sqrt{N}h_N(\theta).$$

$J_N(\theta)$ will have a limiting χ^2 distribution when $\theta = \theta_0$. So consider forming the set of all values θ for which $J(\theta)$ is less than an appropriate critical value J_c from this distribution. Let $\Gamma_N = \{\theta : J_N(\theta) < J_c\}$. For large N the set Γ_N will have the appropriate coverage probability – it is a confidence interval. For example, if J_c is the 95th percentile of the limit distribution then the probability that $\theta_0 \in \Gamma_N$ will be approximately 95% for large N . This remains true even if there are multiple values of θ that generate the same moments for $Y_{m,t}$ so that θ_0 is not identified. The consequence of other solutions to the moment equations is that the set Γ_N will not shrink to the point θ_0 as sample size grows. Nevertheless, Γ_N is a very useful characterization of the set of models that are consistent with the data.¹²

6 Conclusion

This paper examines the question of identification and estimation of local interaction models in the absence of perfect information about individual agents' locations. This is an important issue given the variety of applications in which socio-economic distances between agents are potentially mismeasured, or the only data on locations that are available to the researcher are defined at some level of spatial aggregation.

We first show that the parameters of a particular interaction model are identified from a single cross-section if one observes individual agents' outcomes and their exact locations. Our demonstration of identification works even if agents are assumed to be heterogeneous, i.e. are characterized by a vector of covariates that determine their transitions between states, as long as the distribution of such covariates is isotropic over the lattice.

Next, we discuss the question of identification in the presence of two particular types of imperfect information on locations. When individual locations are mismeasured, we show that using our set of moment conditions one cannot recover the structural parameters of the model almost surely, w.r.t. the probability distribution of mappings that describe the distortion process. On the other hand, when locations are measured correctly but only up to some spatial region, numerical simulations strongly suggest that local identification is attained, using moment conditions involving region level aggregates. Taken together, our results indicate that it may be better to use correct but imprecise data on locations than more detailed but potentially incorrect information about individual locations.

One weakness of this paper is that we consider only one very restrictive class of parametric interaction models. It is unsurprising that we can obtain identification

¹²See Hansen, Heaton and Yaron (1996) and Hu (2000) for examples of very similar methods.

with such a tight parameterization when locations are correct if imprecise. An important question for future research is what aspects of the joint distribution of agents' outcomes are nonparametrically identified when location data are only available up to region. Extensions of the methods in Cross and Manski (forthcoming) to the kind of dependent data generated by local interaction models may allow us to nonparametrically characterize the restrictions placed upon the conditional distribution of y given agents' own characteristics and their neighbors' outcomes by the distributions that are recoverable from region-level data.

7 Appendix

Proof of Proposition 1 Suppose we fix a configuration of covariates \overline{X} . Under Assumption 1, the balance equations (1) and (2) can be modified as follows:

$$[0] : \gamma(\overline{X})\pi(0|\overline{X}) + 2\lambda(\overline{X})\pi(10|\cdot\overline{X}) = \pi(1|\overline{X}) \quad (11)$$

$$[00] : 2\gamma(\overline{X})\pi(00|\overline{X}\overline{X}) + 2\lambda(\overline{X})\pi(100|\cdot\overline{X}\overline{X}) = 2\pi(10|\overline{X}\overline{X}) \quad (12)$$

The notation $\pi(\cdot|\cdot)$ denotes the proportion of specific configurations of ones and zeros, given a specific value \overline{X} for the covariates at those sites. For example, $\pi(0|\overline{X})$ indicates the proportion of agents in state [0] among agents that are characterized by \overline{X} . Likewise, $\pi(10|\overline{X}\overline{X})$ indicates the proportion of pairs of neighboring agents in state [10] among pairs of adjacent agents that have the same value $\overline{X}\overline{X}$ of characteristics. On the other hand, $\pi(10|\cdot\overline{X})$ refers to the fraction of pairs of neighboring agents in state [10], among pairs of neighboring agents where the second one is characterized by \overline{X} .

The idea is simply to use replicas of equations (11) - (12), conditioning on different values of the covariates, to recover the single parameters (γ_k, λ_k) , $k = 0, 1, \dots, K$. As an illustration, suppose $K = 1$. As long as X_1 can take two distinct values, we can consider sets of agents in the following distinct configurations of covariates: $\overline{X}^a \equiv (X_1 = a)$; $\overline{X}^b \equiv (X_1 = b)$. We then have two distinct sets of equations (11) - (12) for \overline{X}^a and \overline{X}^b , respectively. We use the following notation: $\pi_0^a \equiv \pi(0|\overline{X}^a)$; $\pi_{00}^a \equiv \pi(00|\overline{X}^a\overline{X}^a)$; and so on. Then, using the identity $\pi_0^z \equiv 1 - \pi_1^z$ for every \overline{X}^z , we obtain the following system of equations:

$$\begin{bmatrix} 1 - \pi_1^a & 2\pi_{10}^a & a(1 - \pi_1^a) & a2\pi_{10}^a \\ \pi_{00}^a & \pi_{100}^a & a\pi_{00}^a & a\pi_{100}^a \\ 1 - \pi_1^b & 2\pi_{10}^b & b(1 - \pi_1^b) & b2\pi_{10}^b \\ \pi_{00}^b & \pi_{100}^b & b\pi_{00}^b & b\pi_{100}^b \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \lambda_0 \\ \gamma_1 \\ \lambda_1 \end{bmatrix} = \begin{bmatrix} \pi_1^a \\ \pi_{10}^a \\ \pi_1^b \\ \pi_{10}^b \end{bmatrix} \quad (13)$$

This is a linear system in $(\gamma_0, \lambda_0, \gamma_1, \lambda_1)$ that has a unique solution for the unknown parameters, provided that the matrix in (13) is non-singular. One can proceed in a similar fashion to recover all pairs (γ_k, λ_k) , for any K . ■

References

- [1] Bertrand, Marianne, Erzo F.P. Luttmer and Sedhil Mullainathan (2000), “Network Effects and Welfare Cultures”, *Quarterly Journal of Economics*, Vol. 115, 1019-1055.
- [2] Bezuidenhout, Carol and Geoffrey Grimmett (1990), “The Critical Contact Process Dies Out”, *Annals of Probability*, Vol. 18, No. 4, 1462-1482.
- [3] Brock, William and Steven Durlauf (forthcoming), “Interactions-Based Models”, in *The Handbook of Econometrics*, Vol. 5, J.J. Heckman and E. Leamer, Eds.
- [4] Clifford, Peter and Aidan Sudbury (1979), “On the Use of Bounds in the Statistical Analysis of Spatial Processes”, *Biometrika*, Vol. 66, No. 3, 495-503.
- [5] Conley, Timothy G. and G. Topa (forthcoming), “Socio-Economic Distance and Spatial Patterns in Unemployment”, *Journal of Applied Econometrics*.
- [6] Cross, Philip and C. Manski (forthcoming), “Regressions, Short and Long”, *Econometrica*.
- [7] Glaeser, Edward L., Bruce Sacerdote and José A. Scheinkman (1996), “Crime and Social Interactions”, *Quarterly Journal of Economics*, Vol. 111, 507-548.
- [8] Goolsbee, Austan D. and Peter J. Klenow (forthcoming), “Evidence on Learning and Network Externalities in the Diffusion of Home Computers”, *Journal of Law and Economics*.
- [9] Granovetter, Mark S. (1995), *Getting a Job: A Study of Contacts and Careers*, Cambridge, MA: Harvard University Press.
- [10] Hansen, Lars P., John Heaton and Amir Yaron (1996) “Finite Sample Properties of Alternative GMM Estimators” *Journal of Business and Economic Statistics* Vol. 14 No. 3 pp. 262-280.
- [11] Harris, Theodore E. (1974), “Contact Interactions on a Lattice”, *Annals of Probability*, vol. 2, 969-988.
- [12] Hu, LuoJia (2000) “Estimating a Censored Dynamic Panel Data Model with an Application to Earnings Dynamics”, unpublished manuscript, Northwestern University.
- [13] Liggett, Thomas M. (1985), *Interacting Particle Systems*, New York: Springer Verlag.
- [14] Liggett, Thomas M. (1999), *Stochastic Interacting Systems: Contact, Voter and Exclusion Processes*, New York: Springer Verlag.

- [15] Montgomery, James D. (1991), "Social Networks and Labor-Market Outcomes: Toward an Economic Analysis", *The American Economic Review*, vol. 81, no. 5, pp. 1408-1418.
- [16] Montgomery, James D. (1992), "Social Networks and Persistent Inequality in the Labor Market", unpublished manuscript, Northwestern University.
- [17] Schonmann, Roberto H. (1986), "Central Limit Theorem for the Contact Process", *Annals of Probability*, Vol. 14, No. 4, 1291-1295.
- [18] Topa, Giorgio (2001), "Social Interactions, Local Spillovers and Unemployment", *Review of Economic Studies*, Vol. 68, 261-295.

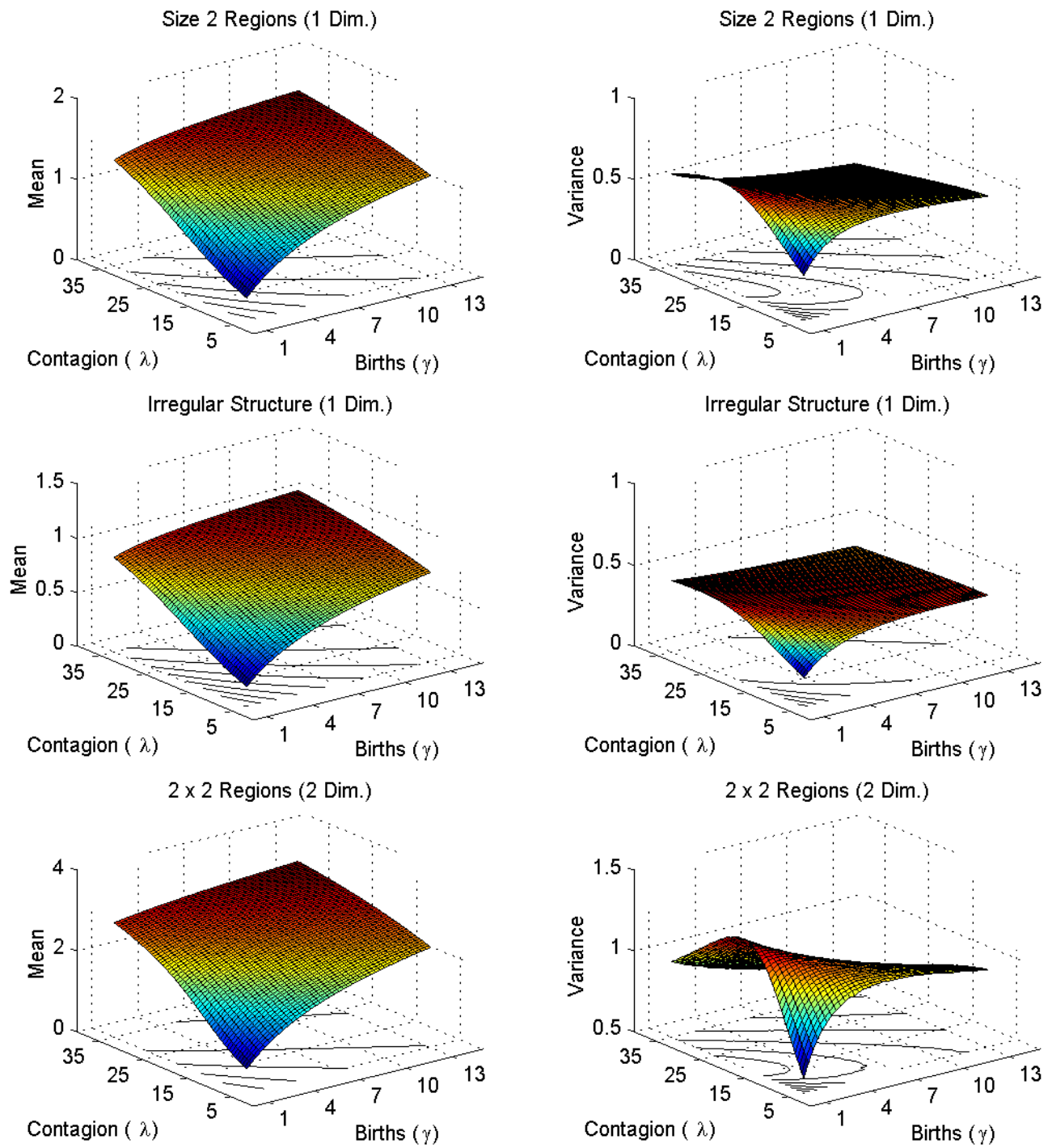


FIGURE 1 - Simulated Moments: Surface Plots

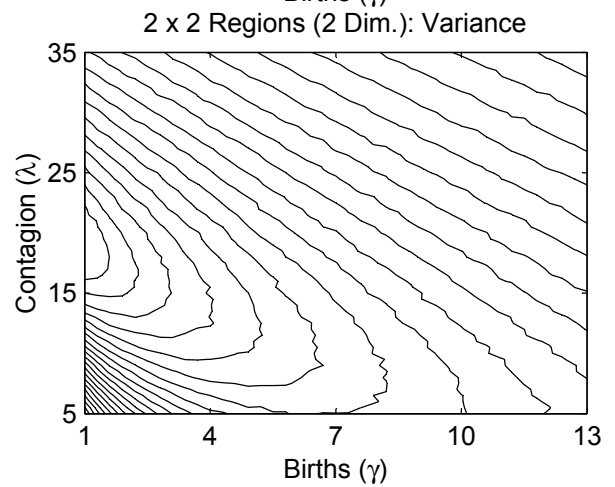
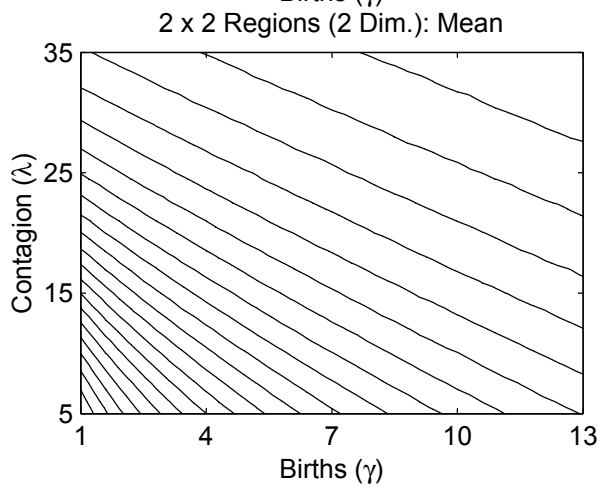
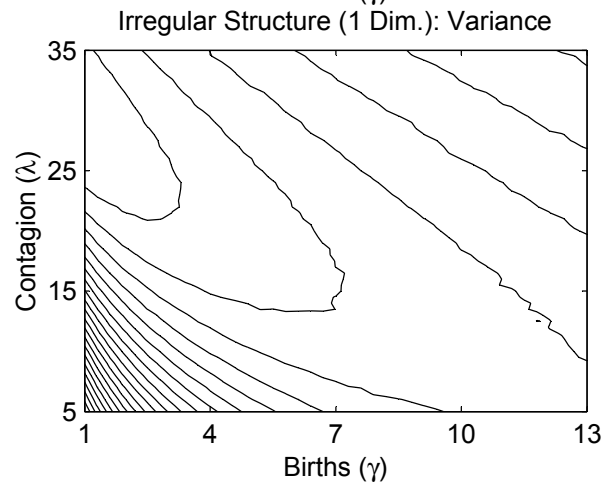
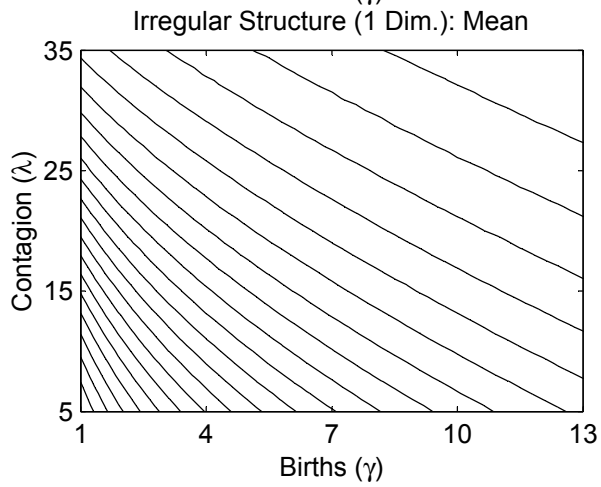
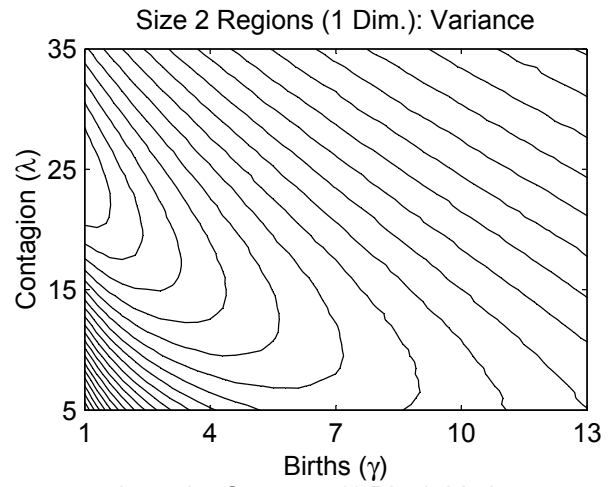
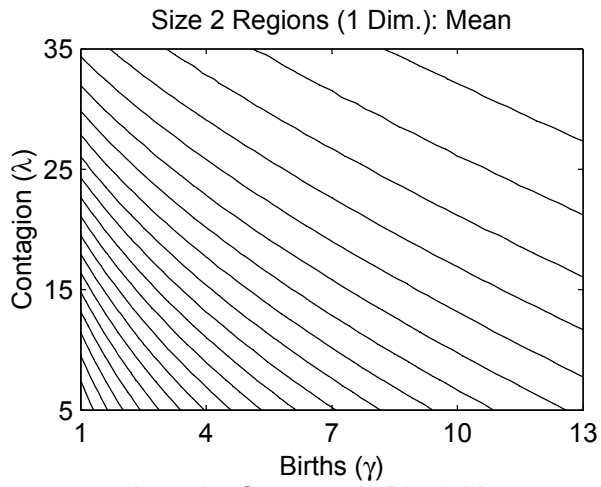


FIGURE 2 - Simulated Moments: Contour Plots