Federal Deposit Insurance Corporation • Center for Financial Research

Working Paper Series

Estimation of Discrete Choice Network Models with Missing Outcome Data

Previously circulated as "Estimation of Spillover Effects in Home Mortgage Delinquencies with Sampled Loan Performance Data"

Denghui Chen
Office of the Comptroller of the Currency

Hua Kiefer
Federal Deposit Insurance Corporation

Xiaodong Liu
University of Colorado Boulder

This Version: April 2022

Published as: "Estimation of Discrete Choice Network Models with Missing Outcome Data" *Regional Science and Urban Economics,* November 2022, Volume 97, 1038355, 1-14. Available online.

FDIC CFR WP 2021-03

fdic.gov/cfr

Estimation of Discrete Choice Network Models with Missing Outcome Data*

Denghui Chen[†] Hua Kiefer[‡] Xiaodong Liu[§]

April 25, 2022

*A previous version of this paper was circulated under the title "Estimation of Spillover Effects in Home Mortgage Delinquencies with Sampled Loan Performance Data". We thank the Co-Editor Stephen Ross for valuable suggestions. We also thank Xudong An, Ryan Goodstein, Amanda Heitz, Mark Kutzbach, Ajay A. Palvia, Jon Pogach, Alexander Ufier, Ioan Voicu, Chiwon Yom, Calvin Zhang, and the seminar participants of the Federal Deposit Insurance Corporation and the Federal Reserve Bank of Philadelphia for helpful comments. The views and opinions expressed in this paper do not necessarily reflect the views of the Federal Deposit Insurance Corporation, the Office of the Comptroller of the Currency, any federal agency, or the United States, and do not establish supervisory policy, requirements, or expectations. This research uses data from Add Health, a program project designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris, and funded by a grant P01-HD31921 from the National Institute of Child Health and Human Development, with cooperative funding from 17 other agencies. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Persons interested in obtaining data files from Add Health should contact Add Health, Carolina Population Center, 123 W. Franklin Street, Chapel Hill, NC 27516-2524 (addhealth@unc.edu). No direct support was received from grant P01-HD31921 for this analysis.

[†] Office of the Comptroller of the Currency, denghui.chen@occ.treas.gov.

[‡] Federal Deposit Insurance Corporation, hkiefer@fdic.gov.

[§] Department of Economics, University of Colorado Boulder, xiaodong.liu@colorado.edu.

Abstract

This paper studies the problem of missing observations on the outcome variable in a discrete choice network model. The research question is motivated by an empirical study of the spillover effect of home mortgage delinquencies, where mortgage repayment decisions can only be observed for a sample of all the borrowers in the study region. We show that the nested pseudo-likelihood (NPL) algorithm can be readily modified to address this missing data problem. Monte Carlo simulations indicate that the proposed estimator works well in finite samples and ignoring this issue leads to a severe downward bias in the estimated spillover effect. We apply the proposed estimation procedure using data on single-family residential mortgage delinquencies in Clark County of Nevada in 2010, and find strong evidence of the spillover effect. We also conduct some counterfactual experiments to illustrate the importance of consistently estimating the spillover effect in policy evaluation.

Keywords: missing data, mortgage defaults, networks, NPL, rational expectation.

JEL: C21, R31

1 Introduction

The past decades have seen a fast progress in the theoretical development of network models. Yet, the empirical applications of network models are still limited due to the high cost of collecting network data. Moreover, most existing estimation methods for network models require that the whole network, instead of a random sample of the network nodes or links, to be observed by the researcher, which escalates the difficulty of data collection. Hence, it is of great practical importance to develop econometric methods to analyze network models with partially observed or sampled network data. The current literature on this topic can be divided into two research strands. The first strand focuses on partially observed or completely unobserved network links (see, e.g., Liu 2013, Chandrasekhar & Lewis 2016, de Paula et al. 2019, Hardy et al. 2019, Lewbel et al. 2019, Breza et al. 2020, Boucher & Houndetoungan 2020, Griffith 2020), while the second strand focuses on the missing data problem in the outcome or covariates of network nodes (see, e.g., Sojourner 2013, Wang & Lee 2013*a*,*b*, Boucher et al. 2014, Liu et al. 2017). Our paper contributes to the second research strand by studying the problem of missing observations on the outcome variable in a discrete choice network model. It complements the studies by Boucher et al. (2014), Wang & Lee (2013*a*,*b*) and Liu et al. (2017) for the same missing data problem in linear network models.

The research question in this paper is motivated by an empirical study of the spillover effect of mortgage delinquencies. To establish a direct link connecting mortgage repayment decisions of neighboring homeowners, we build an empirical model based on the discrete choice network model in Lee et al. (2014), where a mortgage borrower's repayment decision depends not only on neighboring foreclosures in the previous time period (the contagion effect in Towe & Lawley 2013) but also on the rational expectation of neighbors' repayment decisions in the current time period (the spillover effect in Chomsisengphet et al. 2018).[1] An

[1] In the literature, "contagion effects" and "spillover effects" are often used interchangeably. To distinguish between the time-lagged effect of past foreclosures from the contemporaneous effect of current default decisions, we refer to the former as the contagion effect (as in Towe & Lawley 2013), and the latter as the spillover effect.

underlying assumption in Lee et al. (2014) is that the researcher can observe the outcomes and covariates of all individuals in the network. Although this assumption is quite common for network models, it is not realistic for a mortgage repayment behavior study. More specifically, the outcome variable in this empirical model is defined as being 90 days past due or worse (90+ DPD). This information is only available in loan performance data, which is usually collected by the mortgage servicer serving the loans and only covers a portion of all the active mortgage borrowers in the study region depending on the mortgage servicer's market share. Ignoring this missing data issue, by treating the sampled borrowers in the loan performance data as the full population of all the active mortgage borrowers in the study region, may lead to a biased estimate of the spillover effect.

In this paper, we show that, by supplementing the loan performance data with public records on covariates of all the borrowers in the study region, the nested pseudo-likelihood (NPL) algorithm (Aguirregabiria & Mira 2007) can be readily modified to address this missing data issue. Our Monte Carlo simulations indicate that the proposed estimator works well in finite samples and ignoring this missing data issue leads to a severe downward bias in the estimated spillover effect. Using empirical data on single-family residential mortgage delinquencies in Clark County of Nevada in 2010, we find evidence for both a time-lagged contagion effect (Towe & Lawley 2013) and a contemporaneous spillover effect (Chomsisengphet et al. 2018). Consistent with the Monte Carlo simulations, we find that the spillover effect is underestimated when the missing data problem is left unaddressed. We complement our estimation effort with two counterfactual studies to illustrate the importance of consistently estimating the spillover effect in policy evaluation. In the first study, we hypothetically remove properties in foreclosure, one at a time, from the data, and calculate the corresponding reduction in the aggregate delinquency level. In the second study, we introduce a positive utility shock, which can be interpreted as a mortgage payment reduction, to all residents in the study region, and plot the percentage reduction in delinquency rates as the shock increases. In both counterfactual studies, we find that the overall reduction in mortgage delinquencies tends to be understated when the contemporaneous spillover effect is ignored or underestimated due to the missing data problem.

Our empirical findings contribute to a large literature on mortgage defaults and corresponding neighborhood effects. In this literature, some work has shown that mortgage defaults have a significant and highly localized impact on house prices in the neighborhood (Immergluck & Smith 2006, Schuetz et al. 2008, Harding et al. 2009, Campbell et al. 2011, Hartley 2014, Gerardi et al. 2015); while other work has been focusing on the impact of negative equity on default likelihood (Deng et al. 2000, Foote et al. 2008, Bhutta et al. 2010, Elul et al. 2010, Calomiris et al. 2013, Gerardi et al. 2018). Nevertheless, with a few exceptions, little work has been done to study the interaction of neighboring mortgage borrowers' default decisions. Towe & Lawley (2013) relate a homeowner's default decision to the *observed* default decisions of the neighbors in the previous time period (i.e., neighboring foreclosures). Munroe & Wilse-Samson (2013) investigate the impact of a completed foreclosure on future neighboring foreclosure filings. Gupta (2019) studies the contagion effect of foreclosures triggered by an interest rate increase. Huang et al. (2021) develop an exogenous proxy for the fraction of mortgages in negative equity based on the timing of foreclosures in a neighborhood, and use it to estimate the spillover effect of foreclosures. The most close work to ours is Chomsisengphet et al. (2018), which establishes a direct connection between neighboring mortgage borrowers' contemporaneous default decisions. However, in Chomsisengphet et al. (2018), the aforementioned missing data problem is left unaddressed.

The rest of the paper proceeds as follows. Section 2 describes the model, NPL estimation strategy and Monte Carlo simulation experiments. Section 3 presents the data, empirical results and counterfactual studies. Section 4 concludes.

2 Model and NPL Estimation

2.1 Model

Consider a network with a set of n individuals $\mathcal{N} \equiv \{1, \cdots, n\}$. The topology of the network is represented by an $n \times n$ adjacency matrix $W = [w_{ij}]$. Let $y_i \in \{0, 1\}$ denote the dichotomous choices of individual $i \in \mathcal{N}$, X_i denote a row vector of exogenous covariates, and $F(\cdot)$ denote a distribution function. Lee et al. (2014) propose a binary choice network model, where, in the rational expectation equilibrium, the probability of $y_i = 1$ is given by

$$p_i \equiv \Pr(y_i = 1) = F(X_i\beta + \lambda \sum\nolimits_{j \in \mathcal{N} \setminus \{i\}} w_{ij} p_j). \tag{1}$$

In Equation (1), the spatial lag term $\sum_{j \in \mathcal{N} \setminus \{i\}} w_{ij} p_j$ is the weighted sum of the expected outcomes of individual i's connections, with the coefficient λ capturing the network spillover effect.

To motivate the general econometric model defined in Equation (1), we consider a random utility model for home mortgage delinquencies. As in a standard random utility model, the utility of delinquency ($y_i = 1$) is normalized to zero, and the utility of making loan payments ($y_i = 0$) is given by

$$\epsilon_i - X_i\beta - \lambda \sum\nolimits_{j \in \mathcal{N} \setminus \{i\}} w_{ij} y_j, \tag{2}$$

where ϵ_i is an i.i.d. idiosyncratic shock with the distribution function $F(\cdot)$. In the empirical study, $F(\cdot)$ is the standard logistic function, and $w_{ij} = w_{ij}^* / \sum_{j \in \mathcal{N} \setminus \{i\}} w_{ij}^*$, where w_{ij}^* is a known constant capturing the geographical proximity between i and j. More specifically, as the literature suggests the spillover effect of distressed properties is very local and decays rapidly with distance (e.g., Campbell et al. 2011, Gerardi et al. 2015, Cohen et al. 2016), we assume $w_{ij}^* = 1/d_{ij}$ if i and j are within a cutoff distance (say, 0.5 mile), where d_{ij} denotes the geographical distance between i and j, and $w_{ij}^* = 0$ otherwise. Thus, $\sum_{j \in \mathcal{N} \setminus \{i\}} w_{ij} y_j$ is the distance-weighted delinquency rate in mortgage borrower i's neighborhood, with its coefficient λ representing the spillover effect of mortgage delinquencies.

As mortgage delinquencies (90+ DPD) cannot be directly observed by other borrowers, we assume borrowers make delinquency decisions y_i simultaneously. We further assume that $X = (X_1', \cdots, X_n')'$ and distribution of ϵ_i are common knowledge among all borrowers in the area, but the realization of ϵ_i is privately observed by borrower i. In the random utility model, borrower i goes delinquent on loan payments if the *expected* utility of $y_i = 0$, given the information set $\mathcal{I}_i = \{W, X, \epsilon_i\}$, is less than zero, i.e.

$$\mathrm{E}(\epsilon_i - X_i\beta - \lambda \sum\nolimits_{j \in \mathcal{N} \backslash \{i\}} w_{ij} y_j | \mathcal{I}_i) < 0$$

or, equivalently,

$$\epsilon_i < X_i\beta + \lambda \sum\nolimits_{j \in \mathcal{N} \backslash \{i\}} w_{ij} \mathrm{E}(y_j | \mathcal{I}_i).$$

As the distribution function of ϵ_i is $F(\cdot)$, borrower i's probability of delinquency is

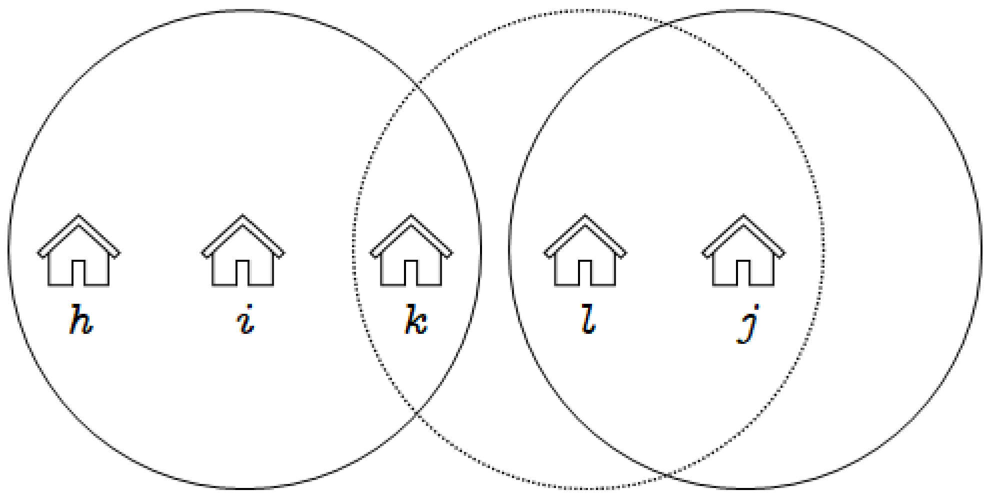
$$p_i \equiv \Pr(y_i = 1) = F(X_i\beta + \lambda \sum\nolimits_{j \in \mathcal{N} \backslash \{i\}} w_{ij} \mathrm{E}(y_j | \mathcal{I}_i)).$$

In the rational expectation equilibrium (Brock & Durlauf 2001*a*,*b*), borrower i's expectation on borrower j's delinquency decision, i.e. $\mathrm{E}(y_j | \mathcal{I}_i)$, should be equal to the mathematical probability for borrower j to be delinquent, i.e. p_j. Therefore, the equilibrium of the random utility model is given by Equation (1). Lee et al. (2014) provide a sufficient condition for the existence of a unique solution to the fixed point problem defined in Equation (1). In the case where $F(\cdot)$ is the standard logistic function and $\sum_{j \in \mathcal{N} \backslash \{i\}} w_{ij} = 1$ for all $i \in \mathcal{N}$, Equation (1) has a unique solution if $|\lambda| < 4$.

To highlight the importance of the spillover effect, we consider the following example. Suppose X_i is the number of foreclosures initiated in the previous period in borrower i's neighborhood.[2] In the absence of the spillover effect, i.e., $\lambda = 0$, the *direct* marginal effect

[2] It is worth pointing out that, in the empirical study, the initiation of foreclosure is indicated by the notice of default (NOD) or the notice of trustee sale (NOTS) filed in the county office. Thus, different from the

Figure 1: The Spillover Effect of Home Mortgage Delinquencies



of X_i on p_i is

$$\frac{\partial p_i}{\partial X_i} = f(X_i\beta)\beta,$$

where $f(x) = \partial F(x)/\partial x$. For a borrower j that is far from i, the *indirect* marginal effect of X_i on p_j is

$$\frac{\partial p_j}{\partial X_i} = 0.$$

That is, when the neighborhoods of borrowers i and j (represented by the solid circles in Figure 1) do not overlap, foreclosures in borrower i's neighborhood have no impact on borrower j's delinquency decision. On the other hand, when $\lambda \neq 0$, the *direct* marginal effect of X_i on p_i is

$$\frac{\partial p_i}{\partial X_i} = (1 + \lambda\psi_{ii}) f_i \beta, \tag{3}$$

and the *indirect* marginal effect of X_i on p_j is

$$\frac{\partial p_j}{\partial X_i} = \lambda\psi_{ji} f_j \beta, \tag{4}$$

delinquency decision in current period that is unobservable to the neighbors, the foreclosure in the previous period is publicly observable.

with $f_i = f(X_i\beta + \lambda \sum_{j \in \mathcal{N} \setminus \{i\}} w_{ij} p_j)$ and ψ_{ij} denoting the (i, j)th element of the matrix

$$\Psi = W(I - \lambda \mathrm{diag}\{f_i\}W)^{-1}\mathrm{diag}\{f_i\},$$

where $\mathrm{diag}\{f_i\}$ is an $n \times n$ diagonal matrix with the ith diagonal element being f_i. Equation (4) implies that, when $\lambda \neq 0$, foreclosures in borrower i's neighborhood may affect borrower j's delinquency decision even if they are far from each other. This can be seen in Figure 1. Suppose house h goes into foreclosure. Knowing that a foreclosure in borrower i's neighborhood has an impact on borrower i's delinquency risk, borrower k will adjust the delinquency decision accordingly. As borrower k is in borrower l's neighborhood (represented by the dotted circle in Figure 1), borrower k's delinquency risk will affect borrower l, which will in turn affect borrower j. Thus, as a result of the chain reaction, a borrower's delinquency decision can be influenced by a foreclosure in a far away neighborhood.

2.2 NPL estimation with missing outcome data

The main difficulty in estimating Equation (1) is that $p = (p_1, \cdots, p_n)'$ is not observable. Lee et al. (2014) suggest to use the nested fixed point (NFXP) algorithm (Rust 1987), with an internal subroutine that solves the fixed point problem given by Equation (1) for p, to implement the maximum likelihood (ML) estimation. To bypass the computational burden of the NFXP algorithm, which repeatedly solves the fixed point problem at each candidate parameter value in the search for the maximum of the log-likelihood function, Lin & Xu (2017) adopt the computationally more efficient NPL algorithm (Aguirregabiria & Mira 2007).

Both Lee et al. (2014) and Lin & Xu (2017) assume that the researcher can observe the outcomes and exogenous characteristics of all individuals in the network. In our empirical study of home mortgage delinquencies, the exogenous variables X_i include the home ownership, house square footage, number of bedrooms, property value, loan-to-value (LTV) ratio, and number of foreclosures initiated in the previous period in borrower i's neighborhood.[3] All this information is public and available for all borrowers in a disclosure state.[4] On the other hand, the outcome variable y_i is defined as being 90+ DPD, and this information is only available in the loan performance data. In the empirical study, we use the loan performance data assembled by a government agency that regulates several national mortgage servicers. Similarly to other popular residential mortgage databases that are commercially available (e.g., CoreLogic or BlackKnight), the coverage of this data depends on the mortgage servicer's market share. For the specific study region of our empirical analysis, Clark County of Nevada, this data covers about 26% of the single-family residential mortgages. In other words, among all the mortgage repayment decisions in the population, about 26% of them are observed and recorded in our data.

More generally, suppose we can observe the exogenous variables X_i for all $i \in \mathcal{N}$, and the outcome variable y_i for $i \in \mathcal{N}^*$, where $\mathcal{N}^*$ is a random sample of $\mathcal{N}$ with the sample size given by $n^* = |\mathcal{N}^*|$. If one drops individuals with missing outcome data and only uses the sample $\mathcal{N}^*$ in the estimation, then the estimated model becomes

$$\Pr(y_i = 1) = F(X_i\beta + \lambda \sum\nolimits_{j \in \mathcal{N}^* \setminus \{i\}} w_{ij} p_j^*), \tag{5}$$

where p_j^* is the solution of the fixed point problem

$$p_i^* = F(X_i\beta + \lambda \sum\nolimits_{j \in \mathcal{N}^* \setminus \{i\}} w_{ij} p_j^*), \tag{6}$$

for $i \in \mathcal{N}^*$. Comparing Equation (5) with Equation (1), we can see that the exclusion of individuals with missing outcome data introduces a measurement error to the spatial lag term $\sum_{j \in \mathcal{N} \setminus \{i\}} w_{ij} p_j$. The measurement error comes from two sources. First, some neighbors of individual i are omitted from the spatial lag term. Second, the equilibrium delinquency

[3] Both property value and LTV ratio are recorded on the loan origination date in the publicly available transaction data. The initiation of foreclosure indicated by the NOD or NOTS is also publicly available.

[4] The study region of our empirical analysis is in Nevada, which is a disclosure state.

probability obtained from Equation (6) is miscalculated. Take the network in Figure 1 as an example. Suppose we do not observe the delinquency decision of borrower k, i.e., $k \notin \mathcal{N}^*$. If we exclude borrower k from the network in the estimation, then the connection between borrowers i and j is cut off. As a result, the interdependence of i and j's delinquency decisions will be attributed to some other confounding factors, leading to an underestimated spillover effect. Hence, simply excluding individuals with missing outcome data may lead to biased estimation results. In the following, we propose a modified NPL algorithm to address this missing data problem.

Let $\theta = (\lambda, \beta')'$. The modified NPL algorithm starts from an initial value $p^{(0)} = (p_1^{(0)}, \cdots, p_n^{(0)})'$ and takes the following iterative steps:

Step 1 Given $p^{(t-1)} = (p_1^{(t-1)}, \cdots, p_n^{(t-1)})'$, obtain $\widehat{\theta}^{(t)} = \arg\max \ln L(\theta; p^{(t-1)})$, where

$$\begin{aligned}\ln L(\theta; p^{(t-1)}) \quad = \quad & \sum_{i \in \mathcal{N}^*} \{y_i \ln F(X_i\beta + \lambda \textstyle\sum_{j \in \mathcal{N}\setminus\{i\}} w_{ij} p_j^{(t-1)}) \\ & +(1 - y_i) \ln[1 - F(X_i\beta + \lambda \textstyle\sum_{j \in \mathcal{N}\setminus\{i\}} w_{ij} p_j^{(t-1)})]\}.\end{aligned}$$

Step 2 Given $\widehat{\theta}^{(t)}$, update $p^{(t)} = (p_1^{(t)}, \cdots, p_n^{(t)})'$ according to

$$p_i^{(t)} = F(X_i\widehat{\beta}^{(t)} + \widehat{\lambda}^{(t)} \textstyle\sum_{j \in \mathcal{N}\setminus\{i\}} w_{ij} p_j^{(t-1)}). \tag{7}$$

Repeat Steps 1 and 2 until the process converges.

It is worth noting that the updating rule for delinquency probabilities given by Equation (7) depends only on X_i, but not on y_i. As we observe X_i for all $i \in \mathcal{N}$, Equation (7) calculates the updated delinquency probabilities for all $i \in \mathcal{N}$, which allows us to obtain the spatial lag term in the log-likelihood function free of measurement error. Kasahara & Shimotsu (2012) show that a key determinant of the convergence of the NPL algorithm is the contraction property of Equation (1), which is ensured by the condition $|\lambda| < 4$ in the case with $F(\cdot)$ being the standard logistic function and $\sum_{j \in \mathcal{N}\setminus\{i\}} w_{ij} = 1$ for all $i \in \mathcal{N}$. When the NPL algorithm converges, the NPL estimator $\widehat{\theta}$ satisfies $\widehat{\theta} = \arg\max \ln L(\theta; \widehat{p})$, where

$$\begin{aligned}
\ln L(\theta; \widehat{p}) &= \sum_{i \in \mathcal{N}^*} \{y_i \ln F(X_i \beta + \lambda \sum\nolimits_{j \in \mathcal{N} \setminus \{i\}} w_{ij} \widehat{p}_j) \\
&\quad + (1 - y_i) \ln[1 - F(X_i \beta + \lambda \sum\nolimits_{j \in \mathcal{N} \setminus \{i\}} w_{ij} \widehat{p}_j)]\},
\end{aligned}$$

and $\widehat{p} = (\widehat{p}_1, \cdots, \widehat{p}_n)'$ is the solution of the system of equations

$$\widehat{p}_i = F(X_i \widehat{\beta} + \widehat{\lambda} \sum\nolimits_{j \in \mathcal{N} \setminus \{i\}} w_{ij} \widehat{p}_j),$$

for $i \in \mathcal{N}$. Under some standard regularity conditions, it follows by a similar argument as in Aguirregabiria & Mira (2007) that the proposed NPL estimator is consistent and asymptotically normal.

The estimation of the asymptotic variance of the NPL estimator $\widehat{\theta}$ also needs to take this missing data issue into consideration. Let $\widehat{\Omega}$ be an $n \times n$ diagonal matrix with the ith diagonal element being $\widehat{f}_i^2 / [\widehat{F}_i (1 - \widehat{F}_i)]$, where $\widehat{F}_i = F(X_i \widehat{\beta} + \widehat{\lambda} \sum_{j \in \mathcal{N} \setminus \{i\}} w_{ij} \widehat{p}_j)$ and $\widehat{f}_i = f(X_i \widehat{\beta} + \widehat{\lambda} \sum_{j \in \mathcal{N} \setminus \{i\}} w_{ij} \widehat{p}_j)$. Let J be a $n^* \times n$ selector matrix such that JX collects elements in $X = (X_1', \cdots, X_n')'$ corresponding to $i \in \mathcal{N}^*$. The asymptotic variance of $\widehat{\theta}$ can be estimated by

$$(\widehat{\Sigma}_1 + \widehat{\lambda} \widehat{\Sigma}_2')^{-1} \widehat{\Sigma}_1 (\widehat{\Sigma}_1 + \widehat{\lambda} \widehat{\Sigma}_2)^{-1}$$

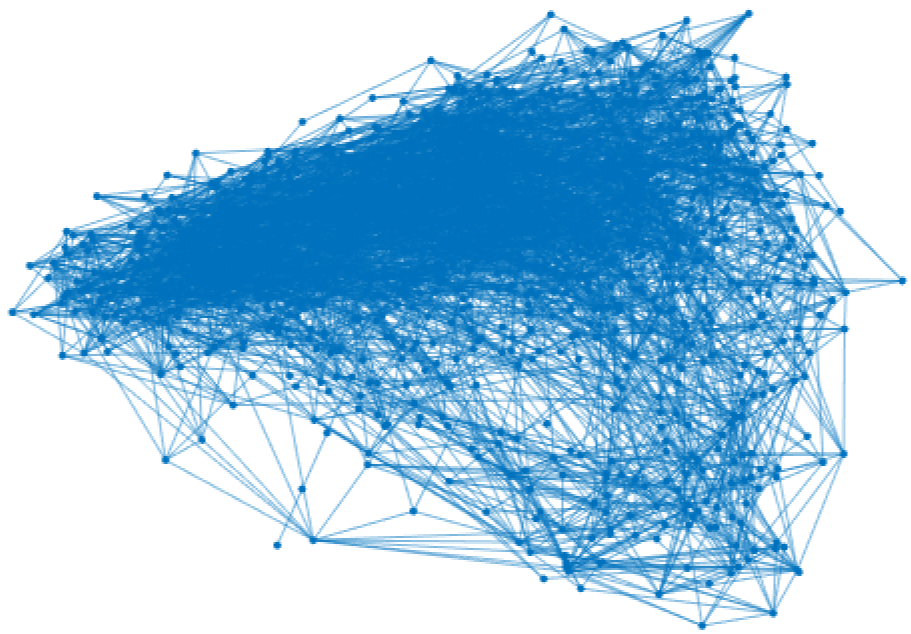
where

$$\begin{aligned}
\widehat{\Sigma}_1 &= [X, W\widehat{p}]' J' J \widehat{\Omega} J' J [X, W\widehat{p}], \\
\widehat{\Sigma}_2 &= [X, W\widehat{p}]' J' J \widehat{\Omega} J' J W (I_n - \widehat{\lambda} \text{diag}\{\widehat{f}_i\} W)^{-1} \text{diag}\{\widehat{f}_i\} [X, W\widehat{p}].
\end{aligned}$$

2.3 Monte Carlo simulations

To investigate the finite sample performance of the proposed NPL algorithm, we conduct a simulation study. In the data generating process, we consider both generated and empirical

Figure 2: Friendship Network



networks. The generated networks provide stylized facts on how the missing data bias is affected by the network configuration (e.g., a sparse network v.s. a dense network). The empirical network shows how the proposed estimator performs in a more realistic setting.

The generated network has two spatial layouts based on the rook contiguity and the queen contiguity. More specifically, we allocate $n = 2500$ spatial units into a lattice of 50×50 squares. Under the rook contiguity, $w^*_{ij} = 1$ if the squares containing i and j share a common side and $w^*_{ij} = 0$ otherwise. Under the queen contiguity, $w^*_{ij} = 1$ if the squares containing i and j share a common side or vertex and $w^*_{ij} = 0$ otherwise. For both spatial layouts, the adjacency matrix is given by $W = [w_{ij}]$ with $w_{ij} = w^*_{ij} / \sum_{j \in \mathcal{N} \setminus \{i\}} w^*_{ij}$.

The empirical network is pulled from the Add Health survey, which collected data on the social environment of students in grades 7-12 from roughly 130 private and public schools in the Unites States in the academic year 1994-95. In the Add Health survey, every student attending the sampled schools on the interview day was asked to identify their friends (up to five males and five females) from the school roster. We use School #56 in the Add Health survey for the simulation study. After removing isolated students with no friends, the remaining 1546 students in School #56 are directly or indirectly connected in the friendship network as shown in Figure 2. On average, the students nominated 4.71 friends, with a standard deviation of 2.85. Let $w_{ij}^* = 1$ if student i nominated student j as a friend and $w_{ij}^* = 0$ otherwise. The adjacency matrix of the friendship network is given by $W = [w_{ij}]$ with $w_{ij} = w_{ij}^*/\sum_{j\in\mathcal{N}\setminus\{i\}} w_{ij}^*$.

In the simulation study, $F(\cdot)$ is the standard logistic function, and $X_i = (1, x_{2i})$, where x_{2i} is a scalar that is generated from a uniform distribution on $[-1, 1]$. We set $\lambda = 1$ and $\beta = (\beta_1, \beta_2)' = (-1, 2)'$ in the data generating process. We experiment with different sampling rates $n^*/n \in \{0.75, 0.5, 0.25\}$, and assume that X_i is observable for all $i \in \mathcal{N}$ while y_i is observable only for $i \in \mathcal{N}^*$.

We consider two NPL estimators in the simulation study. The NPL-1 estimator excludes individuals with missing observations on the outcome variable and only uses the sample $\mathcal{N}^*$ for the estimation. Hence, the NPL-1 estimator is likely to be biased due to the measurement error in the spatial lag term as explained in Section 2.2. The NPL-2 estimator follows the modified NPL algorithm described in Section 2.2. We conduct 1000 simulation repetitions. The mean and standard deviation (SD) of the empirical distribution of the NPL estimates are reported in Table 1. With the NPL-1 estimator, the estimated spillover effect (λ) is downward biased, the estimated intercept (β_1) is upward biased, and the estimated slope (β_2) is essentially unbiased. The bias increases as the sampling rate decreases. Comparing the rook contiguity and the queen contiguity, we can see that the bias is larger when the underlying network is more sparse (i.e., under the rook contiguity). The direction and size of the NPL-1 estimator's bias are comparable between the generated rook/queen network and the empirical friendship network. On the other hand, the NPL-2 estimates are unbiased with both generated and empirical networks, even when the sampling rate is low ($n^*/n = 0.25$).

Table 1: Monte Carlo Simulation Results

	NPL-1			NPL-2		
	$\lambda = 2$	$\beta_1 = -1$	$\beta_2 = 2$	$\lambda = 2$	$\beta_1 = -1$	$\beta_2 = 2$
Rook contiguity						
$n^*/n = 0.75$	1.314(0.327)	-0.655(0.168)	2.026(0.112)	1.991(0.356)	-0.996(0.180)	2.001(0.115)
$n^*/n = 0.50$	0.625(0.290)	-0.294(0.152)	2.054(0.131)	1.973(0.441)	-0.987(0.223)	2.006(0.138)
$n^*/n = 0.25$	0.232(0.311)	-0.077(0.145)	2.061(0.189)	1.958(0.615)	-0.976(0.314)	2.004(0.198)
Queen contiguity						
$n^*/n = 0.75$	1.511(0.405)	-0.757(0.206)	2.006(0.109)	1.985(0.413)	-0.994(0.209)	1.999(0.109)
$n^*/n = 0.50$	0.894(0.418)	-0.445(0.215)	2.020(0.132)	1.955(0.495)	-0.978(0.250)	2.003(0.134)
$n^*/n = 0.25$	0.289(0.394)	-0.127(0.198)	2.033(0.188)	1.929(0.715)	-0.962(0.362)	2.006(0.193)
Friendship network						
$n^*/n = 0.75$	1.560(0.308)	-0.795(0.142)	1.991(0.133)	1.983(0.327)	-0.993(0.150)	2.002(0.133)
$n^*/n = 0.50$	1.103(0.353)	-0.577(0.163)	1.987(0.161)	2.005(0.417)	-1.002(0.195)	2.006(0.162)
$n^*/n = 0.25$	0.572(0.414)	-0.320(0.173)	1.995(0.238)	1.974(0.587)	-0.989(0.271)	2.023(0.242)

Mean(SD)

Table 2: Variable Definitions and Summary Statistics

	Definition	RRP data		MM data	
		Mean	SD	Mean	SD
Dependent Variable					
delinquency	1 if 90+ DPD in 2010, and 0 otherwise.			0.19	0.39
Explanatory Variables					
neighbor foreclosures	# of foreclosures initiated in 2009 within the 0.1 mile neighborhood.	16.69	13.68	14.88	12.04
owner	1 if the property is occupied by the owner.	0.73	0.44	0.78	0.41
square footage	The property size in thousand square feet.	2.05	0.78	2.04	0.75
bedrooms	# of bedrooms of the property.	3.39	0.83	3.40	0.82
log property value	The logarithm of the property's value at the loan origination date.	12.35	0.52	12.27	0.51
LTV_60to80	1 if the LTV ratio at the loan origination date is between 60% and 80%.	0.28	0.45	0.31	0.46
LTV_80to100	1 if the LTV ratio at the loan origination date is between 80% and 100%.	0.56	0.50	0.52	0.50
LTV_gt100	1 if the LTV ratio at the loan origination date is greater than 100%.	0.07	0.25	0.07	0.26
# of observations		221,947		58,526	

3 Empirical Analysis

3.1 Data

Our main data sources are the Mortgage Metrics (MM) database and the Renwood Realty Property (RRP) database. The MM data, assembled by the Office of the Comptroller of the Currency (OCC) since January 2008, consists of loan-level origination and monthly performance information of residential first-lien mortgages serviced by seven national banks and a federal savings association regulated by the OCC. The RRP data covers over 151 million properties and 3,143 counties which translates into 99% of the U.S. population coverage.[5] We focus on the single-family residential mortgage repayment information in the MM data for Clark County of Nevada in 2010. The RRP data provide a wholistic coverage on the covariates of almost all single-family mortgage borrowers in that region, including those not in the MM data. Using the notation in Section 2.2, we consider the set of borrowers in the RRP data as $\mathcal{N}$ and that in the MM data as $\mathcal{N}^*$. In our study region, the RRP transaction data contains 221,947 loan records distributed across 155 census tracts,[6] whereas the MM sample only has 58,526 records.

The MM data is in a panel structure with monthly updated information for loan performance. The outcome variable of the empirical model – mortgage delinquency (90+ DPD) – is extracted from this data. It is worth pointing out that a mortgage delinquency is different from a foreclosure. The former is a decision made by a homeowner to not make a mortgage payment, while the latter is a legal process in which a lender attempts to recover the balance

[5]The RRP database consists of three types of data: (1) the transaction data, which provides a history of sales and financing activities on residential housing units, (2) the property tax assessment data collected from county (township) tax assessor's office, and (3) the pre-foreclosure data (e.g., public records of NOD and NOTS). We use mortgage transaction data (excluding cash transactions) in RRP to construct the pool of active mortgages in the study region. Although we do not know if a mortgage is paid off at the time of our analysis, we feel comfortable that our RRP mortgage data provides a reasonable proxy of the true "active" mortgage population given the fact that the average loan age of the mortgages in our study region is 5.2 years as of the end of 2009. We use the RRP tax assessment data for a complete set of housing characteristic measures. The pre-foreclosure data of RRP provides us information of the existing foreclosure filings, through which we can identify the contagion effect.

[6]We focus on census tracts where most single family homes are located by dropping census tracts with less than 1000 single-family loan records in the RRP data.

of a loan from a borrower who has stopped making payments to the lender by forcing the sale of the asset used as the collateral for the loan. Once a loan reaches a serious delinquency state, such as 90+ DPD, it is usually up to the state level laws and policies (e.g., the foreclosure law) as well as financial institute level programs (e.g., proprietary modification programs for loss mitigation) to determine how the foreclosure process proceeds. Because we are interested in a borrower's decision instead of the legal aspect of its consequence, we define the outcome variable as being 90+ DPD in 2010. On the other hand, it is well documented that mortgage delinquency decisions could be affected by neighboring foreclosures in the previous time period (Towe & Lawley 2013). Hence, we include the number of foreclosures initiated in 2009 in a borrower's neighborhood as a covariate in the empirical model. The initiation of foreclosure is indicated by the notice of default (NOD) or the notice of trustee sale (NOTS) filed in the county office. This information is publicly available in a disclosure state (e.g., Nevada) and contained in the RRP data. Other covariates in the empirical model includes the home ownership, house square footage, number of bedrooms, property value (on the loan origination date), and LTV ratio (on the loan origination date). All this information is also publicly available and contained in the RRP data. We match the loans in the MM data with those in the RRP data based on encrypted property IDs.[7]

Table 2 lists the definitions of the dependent variable and explanatory variables as well as their summary statistics for both the RRP and MM datasets. Overall, the summary statistics of the explanatory variables are comparable between the two datasets. In both datasets, the average number of neighbor foreclosures is 15~17. The majority of mortgage borrowers claimed to be the owners of their properties. The average size of the property is 2000 square feet, and the typical number of bedrooms is between 3 and 4. The average property value is about $220K. The number of borrowers with an initial LTV greater than 80% is slightly more than the number of borrowers with an initial LTV less than 80%.

[7] The encrypted property IDs were generated based on the actual address of each property. After the encrypted property IDs were generated, address information has been removed from both the RRP and MM data. Thus we, as the end data user, have no access to personally identifiable information.

In the empirical analysis, we treat a census tract as a spatial network. Thus, the scope of spatial interactions is restricted to the census tract level. It is natural to think that the interdependence of delinquency decisions is more likely to exist between nearby houses. We therefore adopt the conventional inverse-distance-based spatial weights and assign zero weights to houses located farther than a cutoff-distance apart. More specifically, we define the spatial weight as $w^*_{ij} = 1/d_{ij}$ if i and j are within a cutoff distance, where d_{ij} denotes the geographical distance between i and j, and $w^*_{ij} = 0$ otherwise. We normalize the spatial weight as $w_{ij} = w^*_{ij}/\sum_{j\in\mathcal{N}\setminus\{i\}} w^*_{ij}$, so that the spatial lag term $\sum_{j\in\mathcal{N}\setminus\{i\}} w_{ij}p_j$ can be interpreted as the distance-weighted average delinquency probability in a borrower's neighborhood. The radius of the neighborhood is given by the cutoff distance for $w^*_{ij} = 0$. In the empirical study, we experiment with different cutoff distances ranging from 0.5 miles to 0.1 miles and find the estimation results are robust. Figures 3-5 give a visualization of the average number of neighbors of each house with different cutoff distances for the census tracts used in the empirical analysis. Figure 6 plots the distribution of the delinquency rate in each house's neighborhood with different cutoff distances using the MM data. We can see that the distribution of the delinquency rate is quite stable with different cutoff distances.

3.2 Estimation results

The estimation results are reported in Table 3. The first column reports the standard logit estimates without accounting for the delinquency spillover effect. The second and third columns report the NPL estimates of Equation (1) with the delinquency spillover effect. The NPL-1 estimator falsely treat the borrowers in the MM data as the whole population and only uses the information on those borrowers and their properties to estimate the model. As explained in Section 2.2, the NPL-1 estimator is likely to be biased due to the measurement error in the spatial lag term. The NPL-2 estimator is the consistent estimator proposed in Section 2.2 for data with missing values on the dependent variable.

An implicit assumption of the general model given by Equation (1) is that the under-

Figure 3: Average Number of Neighbors in a Census Tract with Cutoff Distance of 0.5 miles

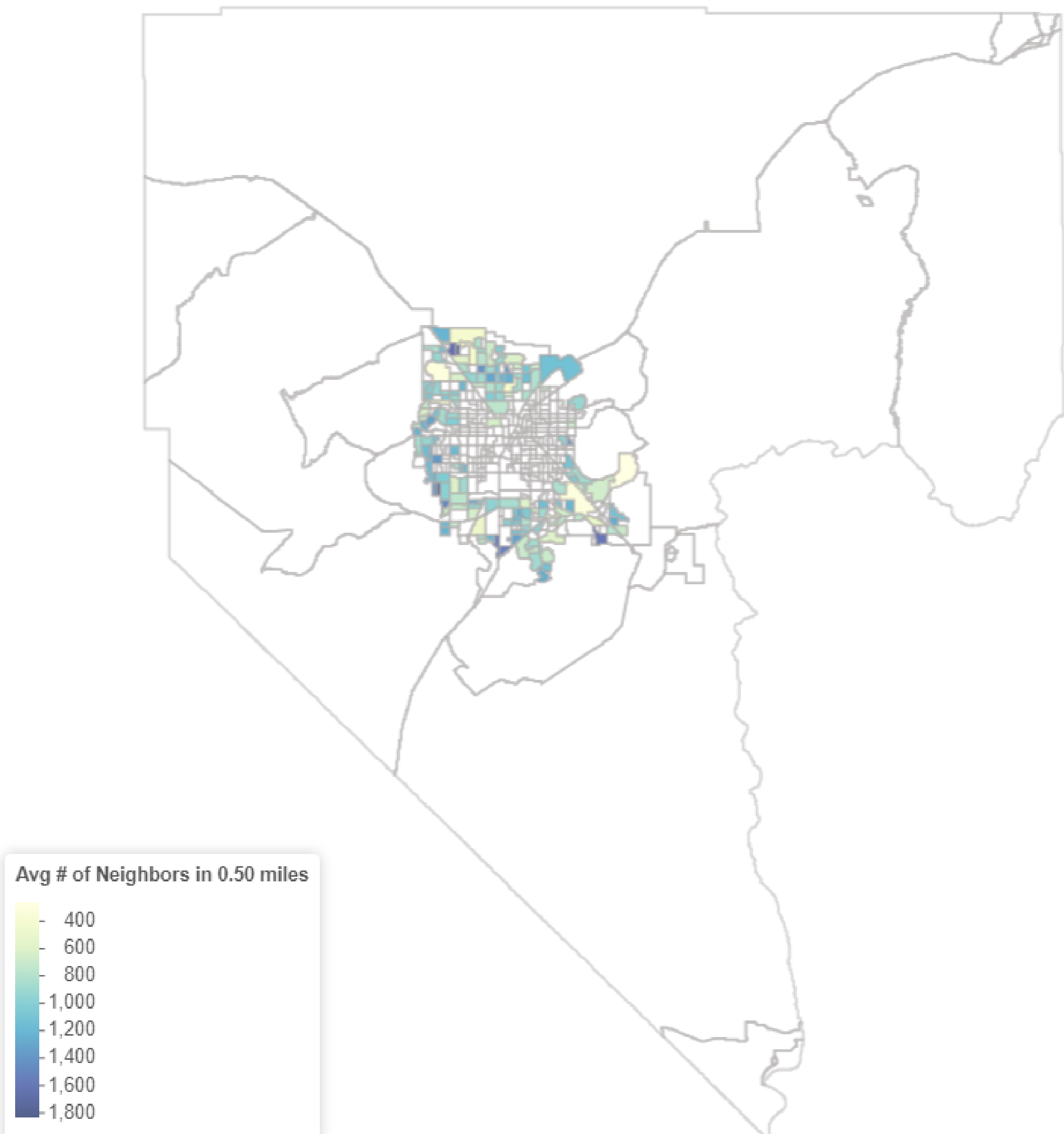


Figure 4: Average Number of Neighbors in a Census Tract with Cutoff Distance of 0.25 miles

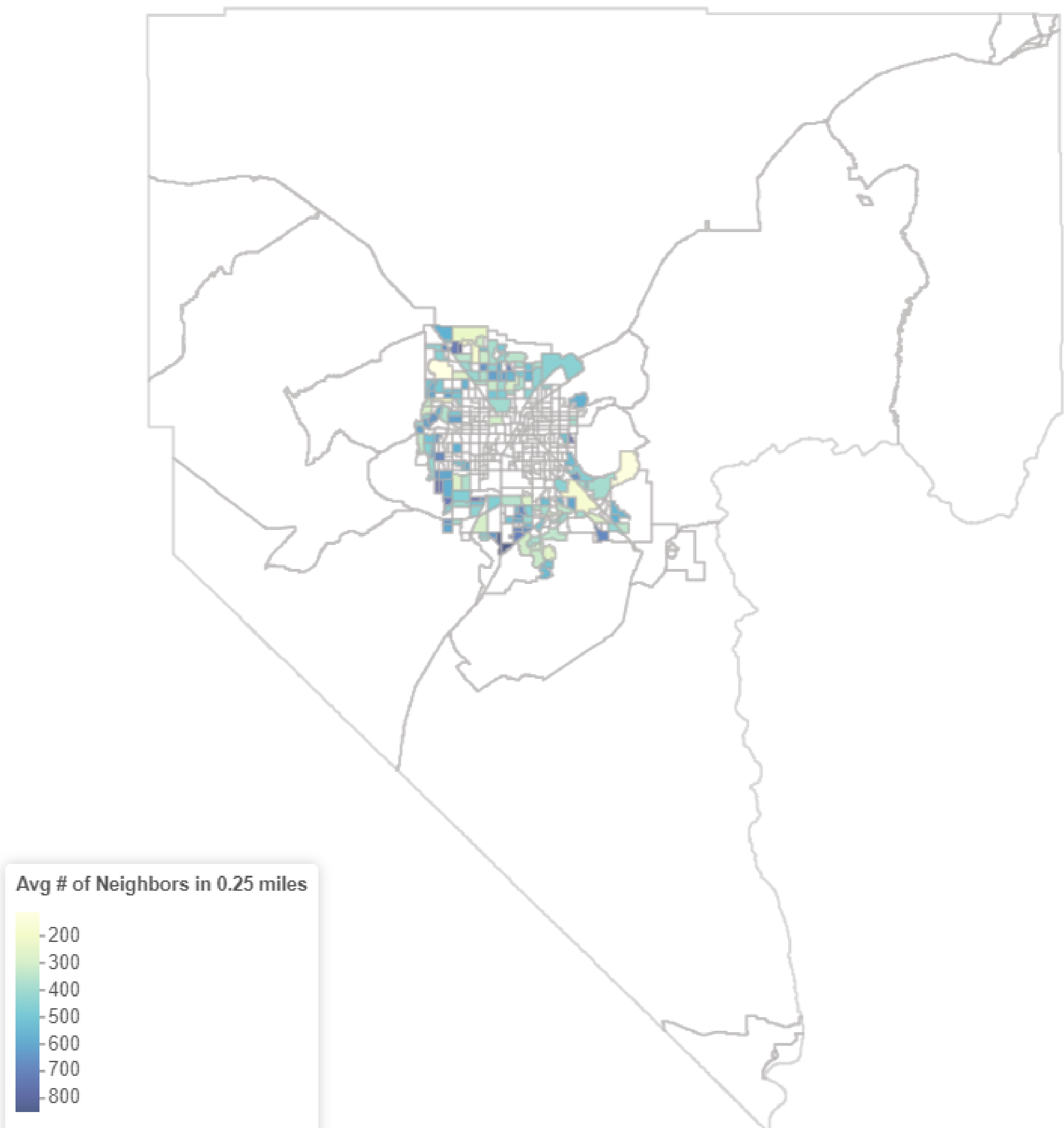


Figure 5: Average Number of Neighbors in a Census Tract with Cutoff Distance of 0.1 miles

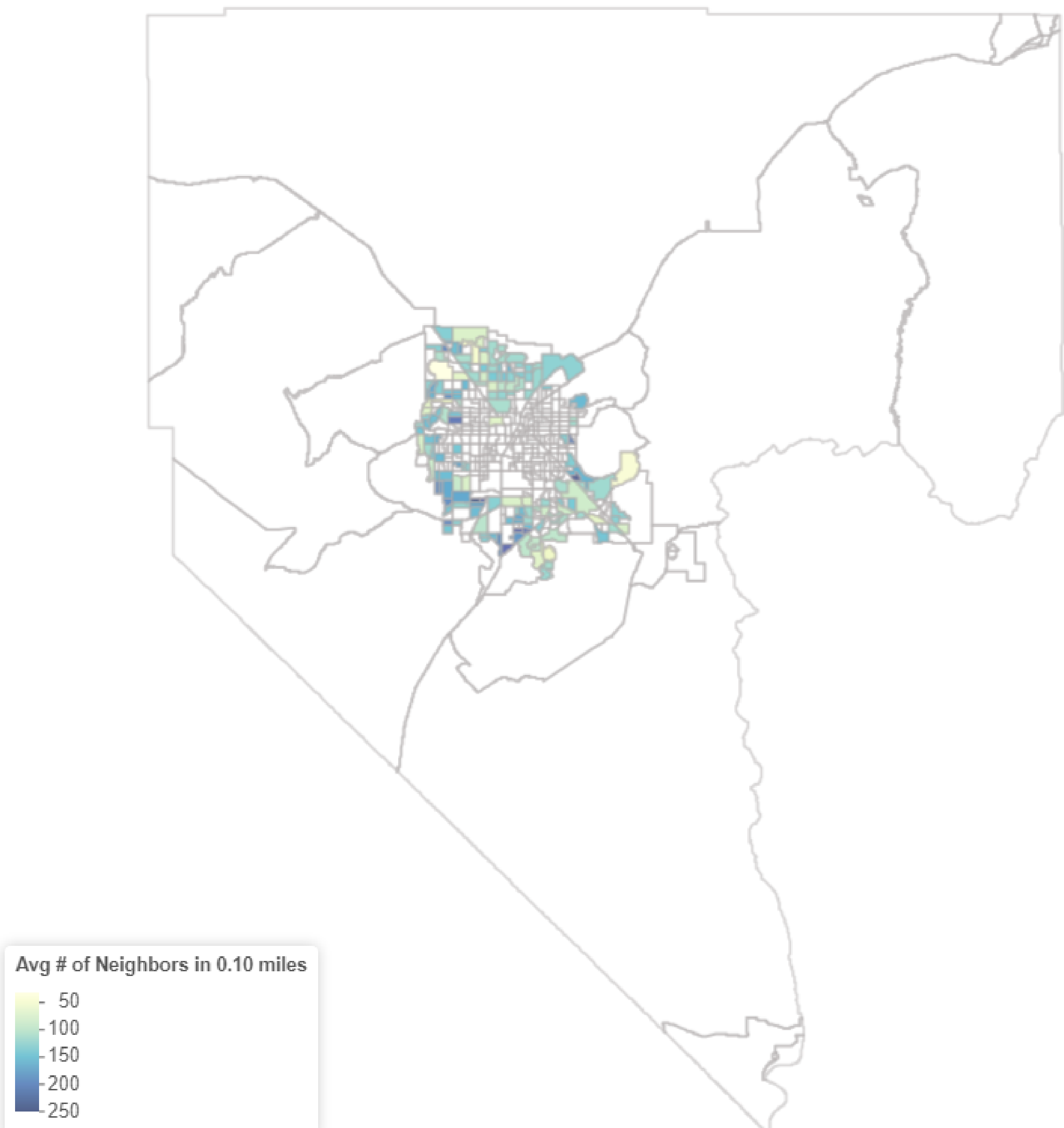
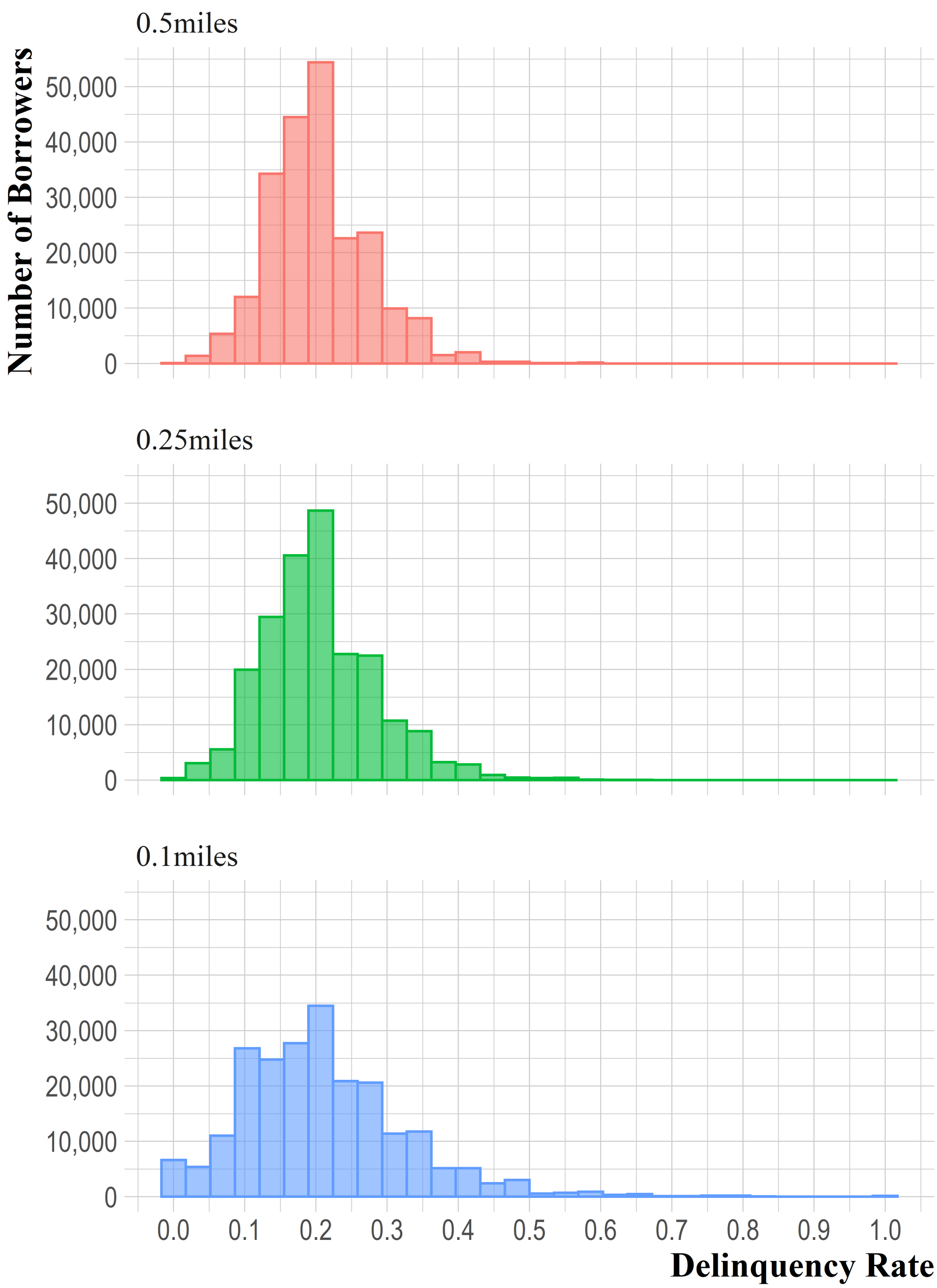


Figure 6: Distribution of Delinquency Rates with Different Cutoff Distances



lying network represented by the adjacency matrix W is exogenously predetermined. In the empirical study, to alleviate the concern of network endogeneity caused by self-selection of residential locations, we include block group fixed effects in the estimation.[8] As argued by Bayer et al. (2008) and Grinblatt et al. (2008), housing markets are thin, which limits people's ability to pick the exact residential location in their desired neighborhood. Hence, people's immediate neighbors can be considered as quasi-random conditional on the broad-neighborhood fixed effect. The block group fixed effect also control for regional random shocks (e.g., regional layoffs[9]) that could be confounded with the delinquency spillover effect.

For the logit model, all the coefficient estimates are statistically significant at the 5% level with the expected signs (except that of *bedroom* is statistically insignificant). In particular, a borrower's delinquency risk increases with more neighboring foreclosures in the previous time period, giving evidence to the contagion effect (Towe & Lawley 2013). The delinquency risk also increases with a higher property value and LTV ratio. After controlling for the other covariates (including the property value), larger houses have lower delinquency risks. The positive sign of the coefficient estimate of *owner* is not surprising since occupancy fraud is found to be common in various mortgage markets, including government-sponsored-enterprise-guaranteed, private-securitized, and portfolio-held mortgage markets (Haughwout et al. 2011, Elul & Tilson 2015, Piskorski et al. 2015, Griffin & Maturana 2016). Both Haughwout et al. (2011) and Griffin & Maturana (2016) suggest loans with fraud occupancy status perform much worse than otherwise comparable loans.

In both the NPL-1 and NPL-2 models, we find a positive and significant delinquency spillover effect, while the coefficient estimates for other covariates remain largely the same. The NPL-2 estimate of the delinquency spillover effect is more than twice the NPL-1 estimate. This is consistent with our findings in the Monte Carlo simulations that ignoring the missing

[8]A block group is a subdivision of a census tract or block numbering area. It is the smallest geographic entity for which the decennial census tabulates and publishes sample data.

[9]People who work together tend to live very close to one another. This means that when a company has layoffs, this tends to affect particular neighborhoods — the regional layoffs.

data issue in the MM data leads to a substantial downward bias of the estimated spillover effect.

In our specification of the spatial weight w_{ij}, we assume $w_{ij} = 0$ if the geographical distance between homeowners i and j is greater than a cutoff distance. For the estimation results reported in Table 3, the cutoff distance is set to 0.5 miles. As a robustness check, Table 4 reports the NPL estimates with a cutoff distance of 0.1, 0.25, and 0.5 miles respectively. The results from this sensitivity analysis are reasonable and consistent with our main findings. For most covariates, the estimated coefficients are very similar with different cutoff distances. For the delinquency spillover effect, as the cutoff distance decreases, the NPL-1 estimate decreases significantly (e.g., 1.12 for 0.5 miles and 0.38 for 0.1 miles, with a drop of 66%) whereas the NPL-2 estimate is considerably stable across different cutoffs (2.32 for 0.5 miles and 2.12 for 0.1 mile, with a drop of less than 10%). The high sensitivity of the NPL-1 estimate to the cutoff distance makes intuitive sense. As the spatial network becomes more sparse with a shorter cutoff distance, the bias of the NPL-1 estimate due to the missing data issue becomes more severe since the neighborhood delinquencies in the MM sample become less representative of the population level. This is consistent with our findings in the Monte Carlo simulations that the bias of the estimated spillover effect is larger with a more sparse network (under the rook contiguity).

Table 3: Estimation Results

	Logit	NPL-1	NPL-2
delinquency spillover effect		1.1210*	2.3226***
		(0.6459)	(0.4629)
foreclosure contagion effect	0.0063***	0.0058***	0.0048***
	(0.0013)	(0.0012)	(0.0011)
owner	0.0586**	0.0585**	0.0593**
	(0.0274)	(0.0270)	(0.0270)
square footage	-0.3104***	-0.3084***	-0.3061***
	(0.0260)	(0.0266)	(0.0257)
bedrooms	0.0268	0.0261	0.0246
	(0.0197)	(0.0194)	(0.0189)
log property value	0.8200***	0.8149***	0.8062***
	(0.0289)	(0.0325)	(0.0325)
LTV_60to80	0.4499***	0.4499***	0.4487***
	(0.0496)	(0.0495)	(0.0494)
LTV_80to100	0.8105***	0.8103***	0.8080***
	(0.0479)	(0.0479)	(0.0479)
LTV_gt100	0.6670***	0.6659***	0.6630***
	(0.0613)	(0.0617)	(0.0617)
block group dummies	included	included	included
log-likelihood	-27352.78	-27351.48	-27345.05

Standard errors in parentheses. Statistical significance: ***p<0.01; **p<0.05; *p<0.1.

Table 4: NPL Estimation with Different Cutoff Distances

	Cutoff dist. = 0.5 mi		Cutoff dist. = 0.25 mi		Cutoff dist. = 0.1 mi	
	NPL-1	NPL-2	NPL-1	NPL-2	NPL-1	NPL-2
delinquency spillover effect	1.1210*	2.3226***	0.7677	2.1307***	0.3796	2.1194***
	(0.6459)	(0.4629)	(0.5698)	(0.3942)	(0.4732)	(0.3217)
foreclosure contagion effect	0.0058***	0.0048***	0.0058***	0.0044***	0.0059***	0.0037***
	(0.0012)	(0.0011)	(0.0013)	(0.0011)	(0.0013)	(0.0010)
owner	0.0585**	0.0593**	0.0585**	0.0594**	0.0584**	0.0591**
	(0.0270)	(0.0270)	(0.0270)	(0.0269)	(0.0270)	(0.0268)
square footage	-0.3084***	-0.3061***	-0.3088***	-0.3054***	-0.3088***	-0.2998***
	(0.0266)	(0.0257)	(0.0266)	(0.0253)	(0.0268)	(0.0241)
bedrooms	0.0261	0.0246	0.0263	0.0245	0.0263	0.0233
	(0.0194)	(0.0189)	(0.0194)	(0.0187)	(0.0195)	(0.0180)
log property value	0.8149***	0.8062***	0.8157***	0.8033***	0.8169***	0.7930***
	(0.0325)	(0.0325)	(0.0326)	(0.0325)	(0.0327)	(0.0328)
LTV_60to80	0.4499***	0.4487***	0.4499***	0.4485***	0.4498***	0.4472***
	(0.0495)	(0.0494)	(0.0495)	(0.0494)	(0.0495)	(0.0494)
LTV_80to100	0.8103***	0.8080***	0.8104***	0.8078***	0.8104***	0.8058***
	(0.0479)	(0.0479)	(0.0479)	(0.0479)	(0.0479)	(0.0478)
LTV_gt100	0.6659***	0.6630***	0.6661***	0.6626***	0.6664***	0.6605***
	(0.0617)	(0.0617)	(0.0617)	(0.0617)	(0.0617)	(0.0616)
block group dummies	included	included	included	included	included	included
log-likelihood	-27351.48	-27345.05	-27351.97	-27344.11	-27352.47	-27341.00

Standard errors in parentheses. Statistical significance: ***p<0.01; **p<0.05; *p<0.1.

3.3 Counterfactual studies

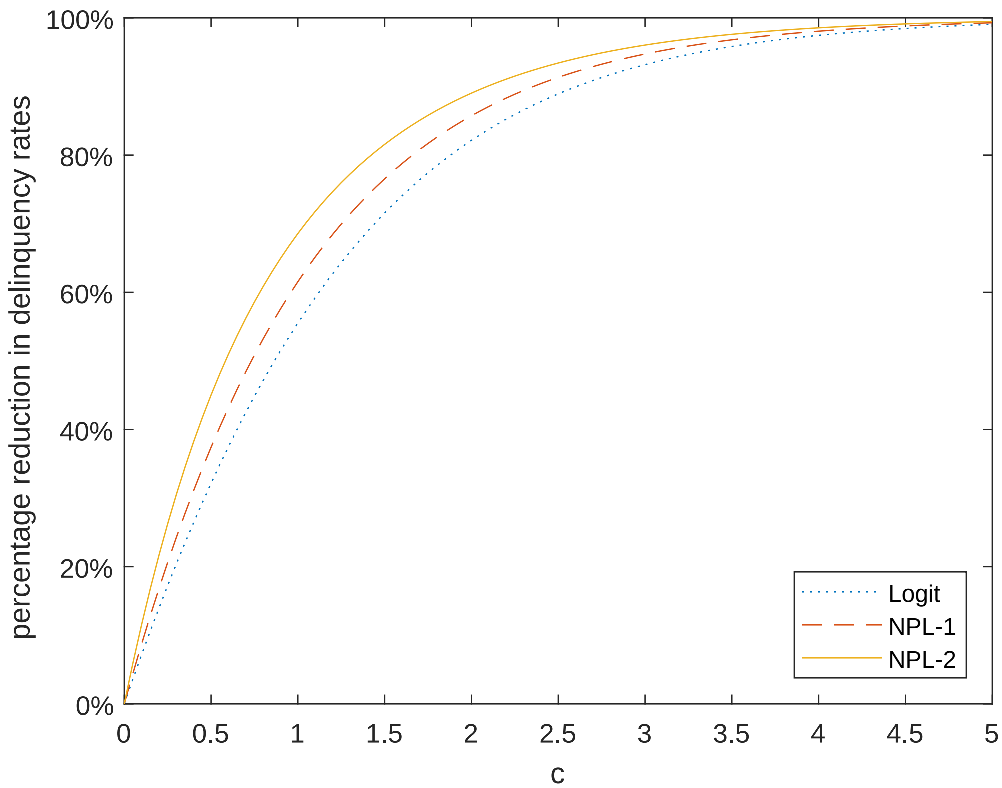
To illustrate the policy relevance of our empirical model and estimation strategy, we carry out two counterfactual studies. In the first study, we hypothetically remove properties in foreclosure, one at a time, from the data, and calculate the corresponding reduction in the aggregate delinquency level. More specifically, we first calculate the predicted delinquency probability for every borrower in the study region and add the probabilities up to obtain the initial aggregate delinquency level. Then, we remove a foreclosure from the study region, re-calculate the predicted delinquency probability for every borrower, and then add them up to get the new aggregate delinquency level. Taking the difference between the two aggregate delinquency levels (before and after the removal of a foreclosure) gives the reduction in the aggregate delinquency level from removing that foreclosure. We then repeat this exercise for every foreclosure in the study region to obtain the corresponding reduction in the aggregate delinquency level. Table 5 reports the summary statistics of the reductions based on the logit, NPL-1 and NPL-2 estimates in Table 3. From the table, we can see that the marginal effect of removing a property in foreclosure tends to be understated when the spillover effect is ignored (logit) or inconsistently estimated due to the missing data problem (NPL-1). This exercise sheds light on the importance of correctly estimating the delinquency spillover effect in evaluating the effectiveness of a foreclosure prevention program.

Table 5: Aggregate Delinquency Reduction from the Removal of a Neighboring Foreclosure

	Mean	SD	Min	Max
Logit	0.16	0.09	0	0.63
NPL-1	0.19	0.12	0	0.81
NPL-2	0.22	0.15	0	1.14

In the second study, we add a constant c, which can be interpreted as a mortgage payment reduction, to the utility function (2) of all mortgage borrowers in the study region. The dotted, dashed, and solid lines in Figure 7 represent, respectively, the predicted percentage reduction in delinquency rates as c increases, based on the logit, NPL-1 and NPL-2 estimates

Figure 7: Reduction in Delinquency Rates with a Utility Shock c



in Table 3. Similar to the first study, we can see that the marginal effect of loan payment reduction is understated when the spillover effect is ignore (logit) or inconsistently estimated due to the missing data problem (NPL-1).

4 Conclusion

This paper proposes a modified NPL algorithm for the missing data problem in the dependent variable of a discrete choice network model. We carry out Monte Carlo simulations to show that the proposed estimator works well in finite samples and ignoring this missing data issue leads to a downward bias of the estimated spillover effect. We provide an empirical illustration of our method and conduct some counterfactual experiments to demonstrate the importance of consistently estimating the spillover effect in policy analysis.

Although the motivation of this paper comes from the missing data issue in home mortgage delinquencies, the applicability of the proposed method is not limited to this specific setting. As the econometric model described in Section 2.1 is very general, this method can be applied to many other data sets. For example, in the Add Health survey, every student attending the sampled schools on the interview day was asked to identify their friends from the school roster and complete a questionnaire (in-school survey) on basic socio-demographic characteristics. Then, a subset of students selected from the rosters of the sampled schools was asked to complete a longer questionnaire containing more sensitive individual and household information (in-home survey). Using the notations in Section 2.2, the students that participated in the in-school survey can be considered as $\mathcal{N}$, and those that participated in the in-home survey can be considered as $\mathcal{N}^*$. Suppose the outcome variable of a study is from the in-home survey, while the covariates are from the in-school survey. Then, the researcher would encounter the same missing data problem as in this paper. When the outcome variable is continuous, Liu et al. (2017) find that the spillover effect in a linear network model is also likely to be underestimated neglecting this missing data problem. They propose a nonlinear least squares estimator to address this missing data problem and provide an empirical illustration using the Add Health data. On the other hand, when the outcome variable is binary, the modified NPL method in this paper can be adopted to consistently estimate the spillover effect.

References

Aguirregabiria, V. & Mira, P. (2007), ‘Sequential estimation of dynamic discrete games’, *Econometrica* **75**, 1–53.

Bayer, P., Ross, S. L. & Topa, G. (2008), ‘Place of work and place of residence: Informal hiring networks and labor market outcomes', *Journal of Political Economy* **116**(6), 1150–1196.

Bhutta, N., Dokko, J. & Shan, H. (2010), The depth of negative equity and mortgage default decisions. FEDS working paper.

Boucher, V., Bramoullé, Y., Djebbari, H. & Fortin, B. (2014), 'Do peers affect student achievement? Evidence from Canada using group size variation', *Journal of Applied Econometrics* **29**, 91–109.

Boucher, V. & Houndetoungan, A. (2020), Estimating peer effects using partial network data. Working paper, Université Laval.

Breza, E., Chandrasekhar, A. G., McCormick, T. H. & Pan, M. (2020), 'Using aggregated relational data to feasibly identify network structure without network data', *American Economic Review* **110**(8), 2454–2484.

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

Brock, W. A. & Durlauf, S. N. (2001*b*), Interactions-based models, *in* J. J. Heckman & E. Leamer, eds, 'Handbook of Econometrics', Vol. 5, North-Holland, pp. 3297–3380.

Calomiris, C. W., Longhofer, S. D. & Miles, W. R. (2013), 'The foreclosure–house price nexus: a panel var model for us states, 1981–2009', *Real Estate Economics* **41**(4), 709–746.

Campbell, J. Y., Giglio, S. & Pathak, P. (2011), 'Forced sales and house prices', *American Economic Review* **101**(5), 2108–2131.

Chandrasekhar, A. & Lewis, R. (2016), Econometrics of sampled networks. Working paper, Stanford University.

Chomsisengphet, S., Kiefer, H. & Liu, X. (2018), 'Spillover effects in home mortgage defaults: Identifying the power neighbor', *Regional Science and Urban Economics* **73**, 68–82.

Cohen, J. P., Coughlin, C. C. & Yao, V. W. (2016), 'Sales of distressed residential property: What have we learned from recent research?', *Federal Reserve Bank of St. Louis Review* **98**(3), 159–188.

de Paula, A., Rasul, I. & Souza, P. C. (2019), Identifying network ties from panel data: Theory and an application to tax competition. CeMMAP working paper: CWP55/19.

Deng, Y., Quigley, J. M. & Van Order, R. (2000), 'Mortgage terminations, heterogeneity and the exercise of mortgage options', *Econometrica* **68**(2), 275–307.

Elul, R., Souleles, N. S., Chomsisengphet, S., Glennon, D. & Hunt, R. (2010), 'What "triggers"mortgage default?', *American Economic Review* **100**(2), 490–94.

Elul, R. & Tilson, S. (2015), Owner-occupancy fraud and mortgage performance. FRB of Philadelphia Working Paper No. 15-45.

Foote, C. L., Gerardi, K. & Willen, P. S. (2008), 'Negative equity and foreclosure: Theory and evidence', *Journal of Urban Economics* **64**(2), 234–245.

Gerardi, K., Herkenhoff, K. F., Ohanian, L. E. & Willen, P. S. (2018), 'Can't pay or won't pay? unemployment, negative equity, and strategic default', *The Review of Financial Studies* **31**(3), 1098–1131.

Gerardi, K., Rosenblatt, E., Willen, P. S. & Yao, V. (2015), 'Foreclosure externalities: New evidence', *Journal of Urban Economics* **87**(C), 42–56.

Griffin, J. M. & Maturana, G. (2016), 'Who facilitated misreporting in securitized loans?', *The Review of Financial Studies* **29**(2), 384–419.

Griffith, A. (2020), Name your friends, but only five? the importance of censoring in peer effects estimates using social network data. Working paper, University of Washington.

Grinblatt, M., Keloharju, M. & Ikäheimo, S. (2008), 'Social influence and consumption: Evidence from the automobile purchases of neighbors', *The Review of Economics and Statistics* **90**(4), 735–753.

Gupta, A. (2019), 'Foreclosure contagion and the neighborhood spillover effects of mortgage defaults', *The Journal of Finance* **74**(5), 2249–2301.

Harding, J., Rosenblatt, E. & Yao, V. (2009), 'The contagion effect of foreclosed properties', *Journal of Urban Economics* **66**(3), 164–178.

Hardy, M., Heath, R. M., Lee, W. & McCormick, T. H. (2019), Estimating spillovers using imprecisely measured networks. arXiv preprint: arXiv:1904.00136.

Hartley, D. (2014), 'The effect of foreclosures on nearby housing prices: Supply or dis-amenity?', *Regional Science and Urban Economics* **49**, 108–117.

Haughwout, A., Lee, D., Tracy, J. S. & Van der Klaauw, W. (2011), Real estate investors, the leverage cycle, and the housing market crisis. FRB of New York Staff Report No. 514.

Huang, W., Nelson, A. & Ross, S. (2021), Foreclosure spillovers within broad neighborhoods. NBER Working Paper 28851.

Immergluck, D. & Smith, G. (2006), 'The external costs of foreclosure: The impact of single family mortgage foreclosures on property values', *Housing Policy Debate* **17**(1), 57–79.

Kasahara, H. & Shimotsu, K. (2012), 'Sequential estimation of structural models with a fixed point constraint', *Econometrica* **80**, 2303–2319.

Lee, L. F., Li, J. & Lin, X. (2014), 'Binary choice models with social network under heterogeneous rational expectations', *The Review of Economics and Statistics* **96**(3), 402–417.

Lewbel, A., Qu, X. & Tang, X. (2019), Social networks with misclassified or unobserved links. Working paper, Boston College.

Lin, Z. & Xu, H. (2017), ‘Estimation of social-influence-dependent peer pressures in a large network game’, *The Econometrics Journal* **20**, 86–102.

Liu, X. (2013), ‘Estimation of a local-aggregate network model with sampled networks’, *Economics Letters* **118**, 243–246.

Liu, X., Patacchini, E. & Rainone, E. (2017), ‘Peer effects in bed time decisions among adolescents: a social network model with sampled data’, *Econometrics Journal* **20**, S103–S125.

Munroe, D. J. & Wilse-Samson, L. (2013), Foreclosure contagion: Measurement and mechanisms. Working Paper, Columbia University.

Piskorski, T., Seru, A. & Witkin, J. (2015), ‘Asset quality misrepresentation by financial intermediaries: Evidence from the RMBS market’, *The Journal of Finance* **70**(6), 2635–2678.

Rust, J. (1987), ‘Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher’, *Econometrica* **55**(5), 999–1033.

Schuetz, J., Been, V. & Ellen, I. G. (2008), ‘Neighborhood effects of concentrated mortgage foreclosures’, *Journal of Housing Economics* **17**(4), 306–319.

Sojourner, A. (2013), ‘Identification of peer effects with missing peer data: Evidence from project star’, *The Economic Journal* **123**, 574–605.

Towe, C. & Lawley, C. (2013), ‘The contagion effect of neighboring foreclosures’, *American Economic Journal: Economic Policy* **5**(2), 313–335.

Wang, W. & Lee, L. F. (2013*a*), ‘Estimation of spatial autoregressive models with randomly missing data in the dependent variable’, *Econometrics Journal* **16**, 73–102.

Wang, W. & Lee, L. F. (2013*b*), 'Estimation of spatial panel data models with randomly missing data in the dependent variable', *Regional Science and Urban Economics* **43**, 521–538.