Empirical Economics
https://doi.org/10.1007/s00181-020-01855-x

Using repeated cross-sectional data to examine the role of immigrant birth-country networks on unemployment duration: an application of Guell and Hu (2006) approach

Kusum Mundra[1,2] · **Fernando Rios-Avila**[3]

Received: 2 April 2019 / Accepted: 12 March 2020

Abstract
Guell and Hu (J Econom 133:307–341, 2006) propose a robust econometric estimator using repeated cross-sectional data in the context of duration/survival analysis where otherwise limited or no results are possible because reliable panel data at an individual level are missing. We present a detailed exposition of the Guell and Hu (GH) strategy and show a Monte Carlo simulation for unemployment duration where the GH method produces better estimates compared to panel data with individual matching errors. We further apply the GH model to examine the immigrant unemployment duration in the USA using the Current Population Survey data from 2001 to 2013 and focus on the role of the birth-country networks on the unemployment duration around the Great Recession. We find that birth-country networks measured at the state level significantly lower unemployment duration for all immigrants, and this effect is stronger during the pre- and post-recession periods than during the recession. We also find that networks are more effective in lowering duration for immigrants unemployed for 1–2 months than for immigrants who are unemployed for longer periods, and this effect is the strongest during the post-recession period. The findings are robust to different specifications and measures of networks including those measured at the local MSA level.

Keywords Repeated cross section · Unemployment duration · Immigrants · Birth-country social networks · Great Recession

JEL Classification J61 · J64 · C 5 · D10

Helpful comments were received from the participants at the Southern Economic Association Meeting, Population Association of America Meeting, and the RGS/RWI Workshop on the Economics of Migration.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00181-020-01855-x) contains supplementary material, which is available to authorized users.

Extended author information available on the last page of the article

Published online: 01 April 2020

1 Introduction

Many issues in economics are best answered in a panel data setting: for example, changes in gender wage gap over time, immigrants earning assimilation at an individual level in their destination country, estimating poverty dynamics in developing countries, and changes in unemployment characteristics and unemployment spells to name a few. However, panel data at the individual level are often not available; it may not be representative for some populations of interest; it may provide bias estimates in the presence of mismatched data; and it may follow individuals for short periods of time. In consequence, available micro-level panel data may not be effective in providing reliable information on changes and dynamics in the population. To overcome the problem of data unavailability, many studies have developed methodologies based on pseudo-/synthetic panels, combining information from pooled cross section data to create, follow and analyze the dynamics of cohorts across time (Dang et al. 2014), with less attention being paid to the analysis of individual-level dynamics.

Guell and Hu (GH) (2006) proposed one of the few methodologies aiming to analyze dynamics at the individual level, instead of at group levels (Sider 1985 and Baker 1992), and unlike other methods Nickell (1979) does not require inflows into unemployment to be constant over time. The GH-proposed estimation strategy uses repeated cross section data to elaborate duration/survival analysis, allowing for robust econometric analysis in scenarios where panel data are limited or not available. In particular, using the framework of unemployment duration analysis, they use an extension of the generalized method of moments, to identify latent risks of leaving unemployment, by comparing the characteristics of those who are unemployed across time.

While this estimation approach is promising for empirical studies focused on survival or duration type of analysis, when access to panel data is limited, there has been limited use of the GH method in the literature. Valletta (2013) is one of the few studies that use the GH method to analyze the relationship between the "house lock" effect around the Great Recession and the impact of unemployment duration.

In this paper, we aim to evaluate and facilitate the use of GH methodology providing a revision of the methodology and presenting a Monte Carlo simulation study that shows the effectiveness of the GH method in a duration analysis framework compared to a standard discrete survival model. Using an unemployment duration analysis framework, the Monte Carlo simulation shows that the GH method produces unbiased estimations, where otherwise no analysis could be done. While the estimates are less efficient compared to scenarios where panel data are fully available, we also show that the GH methodology produces better estimates compared to a scenario where available panel data suffer from random mismatching when following individuals across time.[1]

As an illustration, we use the GH method to analyze the determinants of unemployment duration among immigrants in the USA, with emphasis on the role of immigrant's

[1] This scenario is typical in surveys like the Current Population Survey (CPS) in the USA, where up to 75% of individuals interviewed each month can be followed from one month to the next. However, due to administrative reasons, there is certain degree of mismatch when linking individuals from one month to the next. See Rivera-Drew et al. (2014) for a recent discussion regarding linking records in the Current Population Survey.

social networks for the periods before, during and after the Great Recession. We argue that because the unemployed and immigrant population is relatively small and that the post-match process to identify individuals across time does not take into consideration immigration status, the estimation of standard duration models may provide bias and inconsistent results, compared to the estimations based on the GH methodology. Using monthly data on unemployed immigrants from the Current Population Survey (CPS) for the years 2001 to 2013, we identify potential immigrants' networks using the share of people born in the same country and living in the same state. We find that networks significantly lower unemployment duration and the effect is stronger during the pre- and post-recession periods than during the recession. We also find that networks significantly lower unemployment duration for short unemployment duration but not for longer periods of unemployment. Our findings still hold when immigrant's networks are identified using Census and American Community Survey (ACS) instead of the CPS data, and when networks are measured at the local Metropolitan Statistical Area (MSA) level instead of at the state level.

The rest of the paper is organized as follows. Section 2 discusses the Guell and Hu model in detail. Section 3 presents the Monte Carlo simulation study used to evaluate the GH model in the framework of unemployment duration. Section 4 discusses the application of the estimator to analyze the role of immigrant social networks on unemployment duration around the Great Recession. Section 5 concludes.

2 Econometric model: estimating the probability of leaving unemployment using uncompleted spells from repeated cross section data

Guell and Hu (2006) propose a method that exploits the use of repeated cross-sectional data, using a synthetic cohort analysis, to analyze determinants of unemployment duration when panel data are unavailable. Under the assumption that the multiple cross-sectional data collects information that represents the same population at different points in time, Guell and Hu (2006) methodology estimates the individual probability to remain unemployed for an additional period, by comparing the distribution of characteristics of individuals who were unemployed in the first and the second period, assuming that the only difference between them is the number of periods they remain unemployed.

As Guell and Hu (2006) describe, because the methodology assumes no access to individual panel data, the method has some limitations. First, it does not allow the inclusion of unobserved heterogeneity; second, it does not allow to easily incorporate time-varying covariates; and lastly, it cannot distinguish between transitions into employment or out of the labor force, when analyzing the end of an unemployment spell. Nevertheless, the methodology provides a powerful strategy for duration analysis studies, when panel data are not available. In this section, we provide a detailed explanation of the strategy proposed by Guell and Hu (2006) and describe a Monte Carlo simulation to compare the performance of this methodology, with a more standard, but infeasible, approach.

2.1 Single duration cohort with a single period

As described in Guell and Hu (2006), the estimation of discrete hazard models when panel data are available is straightforward. Consider the distribution $f(X|d,t)$ which characterizes the distribution of characteristics X of individuals who are unemployed for d periods at time t. Let y be an indicator that takes the value of 1 if an individual remains unemployed for an additional period and 0 otherwise. If panel data are available and y is observed for everyone in the population, the distribution of characteristics for those who remain unemployed for an additional period would be given by $f(X|y_i = 1, d, t)$, which is equal to $f(X|d+1, t+1)$. Using Bayes' rule, a nonparametric estimator of the conditional probability of an individual i remaining unemployed for an additional period, conditional on X, d and t, would be given by

$$P(y_i = 1|X,d,t) = P(y_i = 1|d,t)\frac{f(X|y_i = 1,d,t)}{f(X|d,t)}. \tag{1}$$

Under the simplifying assumption that we have a single cohort of interest and only 2 years of data, thus dropping the indices d and t, Eq. (1) can be estimated using a logit model:

$$P(y_i = 1|X) = \frac{e^{X_i\beta}}{1+e^{X_i\beta}} = \Lambda(X_i\beta). \tag{2}$$

Notice that Eqs. (1) and (2) can also be estimated using their sample equivalents, as long as y is observed for everyone. However, if y is unobserved, this becomes an infeasible estimator for the conditional probability of remaining unemployed $P(y_i = 1|X)$.[2]

Without loss of generality, consider a scenario where instead of having access to fully observed panel data (y is fully observed), one has access to two samples that collect data for two periods of time. The base sample, $\mathbb{S}_b = \left\{S_d^0\right\}$, collects data for individuals who were unemployed for d periods at $t = 0$, and the continuation sample, $\mathbb{S}_c = \left\{S_{d+1}^1\right\}$, collects data for individuals who were unemployed for $d+1$ periods at $t = 1$. Further, assume no individual is followed across time, implying that no panel data information exists.

Let $\tilde{y}$ be an indicator for whether an individual i belongs to the first sample $\mathbb{S}_b$ ($\tilde{y}_i = 0$) or the second sample $\mathbb{S}_c$ ($\tilde{y}_i = 1$), and $g(X, \tilde{y})$ denotes the joint distribution of characteristics observed when both samples are pooled together, so that $g(X|\tilde{y}_i = 0)$ and $g(X|\tilde{y}_i = 1)$ denote the distribution of characteristics X observed in samples $\mathbb{S}_b$ and $\mathbb{S}_c$, respectively. Because the base and continuation samples are random samples of the same population, the observed distributions of X can be used instead of their counterparts when y is fully observed:

$$f(X) \cong g(X|\tilde{y}_i = 0) \tag{3}$$

$$f(X|y_i = 1) \cong g(X|\tilde{y}_i = 1). \tag{4}$$

[2] We use $P(.)$ for probability distribution for a discrete variable and $f(.)$ and $g(.)$ for continuous variables.

Using Eqs. (3) and (4), and substituting them into (1), we can obtain an estimator of the conditional probability of remaining unemployed, relative to the unconditional probability of remaining unemployed, that does not depend on y:

$$\frac{P(y_i = 1|X)}{P(y_i = 1)} = \frac{g(X|\tilde{y}_i = 1)}{g(X|\tilde{y}_i = 0)}. \tag{5}$$

Applying Bayes' rule on the right-hand side of Eq. (5), we have

$$\frac{P(y_i = 1|X)}{P(y_i = 1)} = \frac{P_g(\tilde{y}_i = 1|X)}{P_g(\tilde{y}_i = 0|X)} \frac{P_g(\tilde{y}_i = 0)}{P_g(\tilde{y}_i = 1)} = \frac{P_g(\tilde{y}_i = 1|X)}{P_g(\tilde{y}_i = 0|X)} \frac{m_0}{m_1} \tag{6}$$

where $P_g(\tilde{y}_i = 1)$ and $P_g(\tilde{y}_i = 1|X)$ denote the unconditional and conditional probability of an observation belonging to the sample $\mathbb{S}_c$, $P_g(\tilde{y}_i = k) = \frac{m_k}{m_0+m_1}$ for $k = 0, 1$, and m_0 and m_1 are the sample sizes of $\mathbb{S}_b$ and $\mathbb{S}_c$.

Finally, noting that $P_g(\tilde{y}_i = 0|X) = 1 - P_g(\tilde{y}_i = 1|X)$, and reordering some terms, we have that

$$P_g(\tilde{y}_i = 1|X) = \frac{P(y_i = 1|X)}{\frac{m_0}{m_1} P(y_i = 1) + P(y_i = 1|X)}. \tag{7}$$

This is equivalent to the expression (5) provided in Guell and Hu (2006). Because $\tilde{y}$ is observed for all individuals, Eq. (7) can be used to indirectly identify $P(y_i = 1|X)$, choosing, for example, a logistic functional form as in Eq. (2). This is equivalent to identifying the latent conditional probability of an individual to remain unemployed for an additional period. It should be noted that because $\mathbb{S}_b$ and $\mathbb{S}_c$ are random samples of the same population at two points in time, $E\left(\frac{m_0}{m_1} P(y_i = 1)\right) = 1$. However, in finite samples, GH recommend to estimate $\frac{m_0}{m_1} P(y_i = 1)$ as an additional parameter in the model (e^{α}).[3] With these considerations, Eq. (7) can be rewritten as

$$P_g(\tilde{y}_i = 1|X) = \frac{\Lambda(X_i\beta)}{e^{\alpha} + \Lambda(X_i\beta)}. \tag{8}$$

Finally, the parameters of interest α and β are estimated by maximizing the following likelihood function:

$$L(\beta, \alpha) = \prod_{i \in \mathbb{S}_b} \left(\frac{e^{\alpha}}{e^{\alpha} + \Lambda(X_i\beta)} \right) \prod_{i \in \mathbb{S}_c} \left(\frac{\Lambda(X_i\beta)}{e^{\alpha} + \Lambda(X_i\beta)} \right). \tag{9}$$

2.2 Extensions: multiple cohorts, grouped duration data and multiple periods

The previous section describes the basic Guell and Hu (2006) model when there is a single cohort of interest who are analyzed across two periods of time: Individuals who have been unemployed for d periods at time t are unemployed for $d + 1$ periods at $t +$

[3] This functional form is chosen to guarantee that $\frac{m_0}{m_1} P(y = 1)$ is strictly positive.

1. As briefly described in Guell and Hu(2006) (Sect. 4.1), this model can be modified to accommodate pooling individuals with different unemployment spells at time t, pooling individuals from different points in time but similar unemployment spells, as well as allowing to analyze data when, due to data limitations, grouped duration data and longer duration cohorts (i.e., from d to $d+k$) are required for correct identification of the parameters.

2.2.1 Multiple cohorts

Consider two samples. The base sample $\mathbb{S}_b = \left\{S_1^0, S_2^0, \ldots, S_s^0\right\}$ contains information for people that have been unemployed for $d = 1, 2, 3, \ldots, s$ periods at time $t = 0$, and sample $\mathbb{S}_c = \left\{S_2^1, S_3^1, \ldots, S_{s+1}^1\right\}$ contains information for people who have been unemployed for $d = 2, 3, 4, \ldots, s+1$ periods at time $t = 1$. In principle, we could estimate s separate models for each pair of subsamples $\left\{S_d^0; S_{d+1}^1\right\}$ for $d = 1, \ldots, s$, to analyze the conditional probability of remaining unemployed for an additional period. This implies maximizing the following likelihood function:

$$L(B, \mathrm{A}) = \prod_{d=1}^{s}\left(\prod_{i\in S_d^0}\left(\frac{e^{\alpha_d}}{e^{\alpha_d}+\Lambda(X_i\beta_d)}\right)\prod_{i\in S_{d+1}^1}\left(\frac{\Lambda(X_i\beta_d)}{e^{\alpha_d}+\Lambda(X_i\beta_d)}\right)\right) \tag{10}$$

where $B = [\beta_1, \beta_2, \ldots, \beta_s]$ and $\mathrm{A} = [\alpha_1, \alpha_2, \ldots, \alpha_s]$ correspond to parameters that are assumed to be different for each unemployment spell cohort. Because this strategy rapidly increases the number of estimated parameters, one can impose the assumption that $\beta_1 = \beta_2 = \cdots = \beta_k = \beta$ for all variables in X, allowing only the intercept to vary by unemployment spell cohort. Under this assumption, the conditional probability of remaining unemployed, and the parameter α, can be rewritten as follows:

$$P(y_i = 1|X, d) = \Lambda(X_i\beta + \delta_d) \tag{11}$$

where α_d and δ_d are the cohort-specific coefficients that can be estimated using a set of dummies in the model specification. This is, in essence, a simple nonparametric alternative to model the time dependence of the risk of unemployment. With this modification, and considering $D = [\delta_1, \delta_2, \ldots, \delta_s]$, the Likelihood objective function becomes

$$L(\beta, D, A) = \prod_{d=1}^{s}\left(\prod_{i\in S_d^0}\left(\frac{e^{\alpha_d}}{e^{\alpha_d}+\Lambda(X_i\beta+\delta_d)}\right)\prod_{i\in S_{d+1}^1}\left(\frac{\Lambda(X_i\beta+\delta_d)}{e^{\alpha_d}+\Lambda(X_i\beta+\delta_d)}\right)\right). \tag{12}$$

2.2.2 Grouped duration data

Due to a combination of small cohorts and rare events, the identification of the parameters δ_d and α_d, for all unemployment spells cohorts, may be difficult. In such scenarios,

an alternative is to compare cohorts that are farther apart from each, as well as combining cohorts that are too small for identification. The combination of cohorts in the analysis requires imposing restrictions on the cohort-specific coefficients.[4] However, combining samples with different lengths of unemployment duration spells require additional work.

Consider two samples. The base sample $\mathbb{S}_b = \left\{S_d^0\right\}$ contains information on individuals who have been unemployed for d periods at time $t = 0$, and the continuation sample $\mathbb{S}_c = \left\{S_{d+k}^k\right\}$ contains information on individuals who have been unemployed for $d + k$ periods at time $t = k$. Let y^k be an indicator that takes the value of 1 if an individual who was unemployed at time $t = 0$ and remained unemployed for additional k periods, and 0 otherwise. Assume that we are interested in identifying the risks of remaining unemployed for *one* additional period. The probability of finding someone with characteristics X, who was unemployed for d periods at time 0, remains unemployed for additional k periods can be written as the product of the probability of remaining unemployed in all periods in between:

$$P\left(y_i^k = 1|X, d\right) = P(y_i = 1|X, d + k - 1) \ldots P(y_i = 1|X, d + 1)P(y_i = 1|X, d). \tag{13}$$

Because we do not observe individuals at any point between t and $t + k$, the conditional probabilities in Eq. (14) cannot be individually identified without imposing additional constraints. Under the simplifying assumption that all these probabilities are equal to each other, we can rewrite Eq. (13) as

$$P\left(y_i^k = 1|X, d\right) = P(y_i = 1|X, d)^k = \Lambda\left(X_i\beta + \delta_{d,d+k}\right)^k. \tag{14}$$

Substituting this into Eq. (8), we have

$$P_g(\tilde{y}_i = 1|X, d) = \frac{\Lambda\left(X_i\beta + \delta_{d,d+k}\right)^k}{e^{\alpha_{d,d+k}} + \Lambda\left(X_i\beta + \delta_{d,d+k}\right)^k} \tag{15}$$

where $\tilde{y}_i$ is an indicator for whether or not observation i belongs to the base or continuation sample; $e^{\alpha_{d,d+k}} = \frac{m_d}{m_{d+k}} P\left(y^k = 1|d\right)$; and m_d and m_{d+k} are the sample sizes of S_d^0 and S_{d+k}^k and do not require additional manipulation for its estimation. This expression can be substituted into the maximum likelihood objective function in Eq. (12) to estimate β, $\delta_{d,d+k}$ and $\alpha_{d,d+k}$. For example, if we analyze multiple cohorts, using samples that are k periods apart, the corresponding likelihood function becomes

[4] For example, this could imply using the constraints $\delta_7 = \delta_8 = \delta_9 = \delta_{7,8,9}$, assuming that individuals with characteristics $X = x$ who were unemployed for 7, 8 or 9 periods have the same risk of remaining unemployed for one additional period.

$$L(\beta, D, A) = \prod_{d=1}^{s}\left(\prod_{i\in S_d^0}\left(\frac{e^{\alpha_{d,d+k}}}{e^{\alpha_{d,d+k}} + \Lambda\left(X_i\beta + \delta_{d,d+k}\right)^k}\right)\right.$$
$$\left.\prod_{i\in S_{d+k}^k}\left(\frac{\Lambda\left(X_i\beta + \delta_{d,d+k}\right)^k}{e^{\alpha_{d,d+k}} + \Lambda\left(X_i\beta + \delta_{d,d+k}\right)^k}\right)\right). \tag{16}$$

2.2.3 Multiple periods

The last extension is the scenario where one is interested in analyzing unemployment exit risks for multiple cohorts, but using pooled data for multiple periods of time. Consider two samples. The base sample $\mathbb{S}_b = \left\{S_1^0, S_2^0, S_1^1, S_2^1, \ldots, S_1^\tau, S_2^\tau\right\}$ contains information on individuals who have been unemployed for $d = 1, 2$ periods at time $t = 0, 1, \ldots, \tau$, and the continuation sample $\mathbb{S}_c = \left\{S_2^1, S_3^1, S_2^2, S_3^2, \ldots, S_2^{\tau+1}, S_3^{\tau+1}\right\}$ contains information on individuals who have been unemployed for $d = 2, 3$ periods at time $t = 1, 2, \ldots, \tau + 1$.

Two aspects are important for this setup. First, every base subsample identified in $\mathbb{S}_b$ has a corresponding continuation subsample in $\mathbb{S}_c$. This structure is mandatory for the correct identification of the model and was implicitly assumed in the previous sections. For example, S_2^1 is the continuation subsample for S_1^0 and S_3^2 is the continuation subsample for S_2^1. Second, there are some individuals who are part of $\mathbb{S}_b$ and $\mathbb{S}_c$. Namely, individuals belonging to S_2^1 appear as the fourth element of the base sample, but also appear as the first element of the continuation sample. Because of this, and as will be described in the implementation section, all individuals in the pooled sample are potential members of both the base and continuation sample. However, depending on the application and data structure, some observations may not belong to either sample.

Similar to the case of multiple cohorts, it is possible to allow all coefficients of the model to vary across time. However, a simplifying assumption is that all estimated coefficients are constant across time when estimating the model. In this case, the corresponding likelihood function becomes

$$L(\beta, D, A) = \prod_{t=0}^{\tau}\prod_{d=1}^{s}\left(\prod_{i\in S_d^t}\left(\frac{e^{\alpha_d}}{e^{\alpha_d} + \Lambda(X_i\beta + \delta_d)}\right)\prod_{i\in S_{d+1}^{t+1}}\left(\frac{\Lambda(X_i\beta + \delta_d)}{e^{\alpha_d} + \Lambda(X_i\beta + \delta_d)}\right)\right). \tag{17}$$

3 Monte Carlo simulation

To show the performance of the GH method compared to a more standard hazard models using discrete data, we create a synthetic dataset that simulates unemployment duration for a population, taking into consideration five important demographic

variables that effects individuals labor market outcomes: education, age, sex and race based on the following structure:

$$\begin{aligned}
\text{educ} &= \min(x_2, 20), \text{ where } x_2 \sim \text{Poisson}(10) \\
\text{age}^* &= \text{educ} + \text{round}(x_1) + 18, \text{ where } x_1 \sim \chi^2(10) \\
\text{age} &= \begin{cases} = 25 \text{ if age}^* < 25 \\ = age^* \text{ if age}^* \in [25, 55] \\ = 55 \text{ if age}^* > 55 \end{cases} \\
\text{sex} &= \text{Bernoulli}(0.50) \\
\text{race} &= \text{Bernoulli}(0.60).
\end{aligned} \tag{18}$$

To simulate unemployment duration, instead of modeling the conditional probability of remaining unemployed $P(y_i = 1|X, d)$, we simulate its complement. In other words, the probability that unemployment ends after d periods is given by the following process:

$$\begin{aligned}
P(y_i = 0|X, d) = \Lambda\Big(&h(d) + 0.1 * (\text{age} - 35) - 0.01 * (\text{age} - 35)^2 \\
&+ 0.5 * (\text{education}) - .5\text{sex} + 0.4 * \text{race})
\end{aligned} \tag{19}$$

where $\Lambda(.)$ is the logit function and $h(d)$ is the time dependence component that is given by

$$h(d) = f_{\chi^2_{10}}(d) * 35 - 4 \tag{20}$$

where $f_{\chi^2_{10}}(d)$ is the density function of a chi-square distribution with 10 degrees of freedom.

The parameters of this process were chosen to reflect common findings with respect to the probability of exiting unemployment: an inverse U-shape with respect to age, education having a positive impact and large differences in the probability of exiting unemployment by gender and race. The process $h(d)$ was also chosen to allow for an inverse U-shape effect between length of unemployment spell and the probability of ending unemployment.

Using the data process given above, 1 million observations are generated for a population, and each observation is given a maximum of 35 periods over which they are followed. The condition of unemployment is evaluated at each period using a random draw from a Bernoulli distribution with $p = P(y_i = 0|X, d)$. Once an observation ends unemployment, it is dropped from the sample. To simulate a survey structure, each individual is randomly assigned to a period t_0 between 1 and 72. There is an average of 13,888 that "enter" unemployment at each period t_0. The current period t is defined as $t = t_0 + d - 1$, where d is the length of the current unemployment spell. This constitutes the simulated population data with a fully observed unemployment duration for up to 35 periods.

To implement the GH methodology, we draw a random sample of 2.5% of the data and restrict the sample to observations within the $t = 37$ to 72. This is done so that

Table 1 Construction of cohorts in pooled sample

Cohort		Base sample $\mathbb{S}_b$ or $\tilde{y} = 0$		Continuation sample $\mathbb{S}_c$ or $\tilde{y} = 1$
1	S_1^t	$d = 1$ at t	S_2^{t+1}	$d = 2$ at $t + 1$
2	S_2^t	$d = 2$ at t	S_3^{t+1}	$d = 3$ at t + 1
..		…		…
9	S_9^t	$d = 9$ at t	S_{10}^{t+1}	$d = 10$ at $t + 1$
10	$S_{10,11,12}^t$	$d = [10,11,12]$ at t	$S_{13,14,15}^{t+3}$	$d = [13,14,15]$ at $t + 3$
11	$S_{13,\ldots,18}^t$	$d = [13,\ldots,18]$ at t	$S_{19,\ldots,24}^{t+6}$	$d = [19,\ldots,24]$ at $t + 6$

d stands for the number of periods observation i has been unemployed at time t. $\tilde{y}$ is an indicator that takes the values of 0 and 1 to indicate they belong to the base ($\mathbb{S}_b$) or the continuation ($\mathbb{S}_c$) sample. Observations not classified in this table are excluded for the estimation

survey information is available only for a window of time; and to guarantee that at each period t, there is a probability larger than zero of sampling someone who has been unemployed for 1 to 35 periods. Each random sample has an average of 52,500 observations. $<16\%$ of the data correspond to individuals who appear more than once in the data, and just 2% correspond to individuals that are sampled for two or more consecutive periods. For all purposes of the simulation, each observation is treated as a different individual, ignoring the panel data information. For the estimation of discrete hazard models, the unemployment status at $t + 1$ is created for all individuals based on the fully observed data. The information regarding unemployment duration is censored to a maximum of 25 periods.

To implement the GH model, the data are prepared as follows. Following the discussion on multiple periods analysis, all observations from the random sample are duplicated, so that each observation and its duplicate are assigned to the base sample $\mathbb{S}_b$, and the duplicates are assigned to the continuation sample $\mathbb{S}_c$, respectively. Because the sample is not large enough to use all values of d as unemployment spell cohorts, we use the rules described in Table 1 to construct duration cohorts that will be identified in the data. Observations that do not have an appropriate base or continuation sample are excluded from the data. For example, individuals from $t = 37$ are excluded from the continuation sample, because the observations from the corresponding base sample, from $t = 36$, are not be observed. Similarly, observations from $t = 72$ are dropped from the base sample, because the corresponding continuation sample ($t = 73$) is not observed.

All individuals with unemployment duration d less than or equal to 9 periods at time t, who belong to $\mathbb{S}_b$, are allocated to their own cohort, with their corresponding continuation samples in $\mathbb{S}_c$. Individuals in $\mathbb{S}_b$ with 10–12 periods of unemployment duration are assigned to cohort 10, and their counterparts are identified as those who are unemployed for 13–15 periods at time $t + 3$ in $\mathbb{S}_c$. Individuals in $\mathbb{S}_b$ with 13–18 periods of unemployment at time t are assigned to cohort 11, and their counterparts are set to those who are unemployed for 19–24 periods at $t + 6$ in $\mathbb{S}_c$. All other individuals

not identified to a unique cohort are excluded from the sample because their cohort counterparts are not identified.[5]

Using this structure, a Monte Carlo simulation is implemented drawing 1000 samples for which GH estimator is implemented. To compare the results to a more standard methodology, we estimate logit models using the fully observed unemployment state on the next period as the dependent variable, but restricting the data to only 75% of the random sample. This is similar to the proportion of observations that are followed for two periods in the CPS.

To simulate the possibility of records mismatch, the next period unemployment status for 10% of the data is swapped to a different observation in the sample, conditional on having the same characteristics in terms of age, years of education, sex and race. On average, this induces a misclassification of $<3\%$ on the next period unemployment status. This exercise is to show the advantages of using the GH estimator compared to using panel data with potentially mismatched records of individuals across time, as it occurs in the CPS.[6]

The first column in Table 2 provides the coefficients from the logit model for the full simulated data that estimates the likelihood of exiting unemployment in the next period, which for the purposes of the simulation represents the true population coefficients. The second set of columns provides the results corresponding to the logit model for the random panel samples, assuming that unemployment status in the next period is fully observed. We report the average bias, simulation-based standard errors and the logarithm of the mean squared error (LMSE) using the squared difference between the estimated and population coefficients for all variables in the model.[7] As expected, the results show negligible bias.

The third set of columns summarizes the results from the GH model. The summary shows the estimations also have a negligible bias for most coefficients, although they are somewhat larger than the logit panel counterpart. The coefficients corresponding to the 10–12 and 13–18 cohorts show the largest biases mostly because they cannot be directly compared to the logit results.[8] In terms of overall model performance, based on the LMSE, we observe that while the logit model based on fully observed panel data outperforms GH estimator, GH model still performs reasonably well, but with standard errors that are larger than those of the fully observed panel data. It is also

[5] For example, because unemployment duration spells are censored to 25 periods, observations in $\mathbb{S}_c$ with unemployment spells longer than 24 periods cannot be unambiguously assigned to a specific cohort in $\mathbb{S}_b$. Also, individuals with 11 and 12 periods of unemployment duration in $\mathbb{S}_c$ are also excluded from the estimation.

[6] See Rivera-Drew et al. (2014) for a recent discussion regarding linking records in the Current Population Survey.

[7] There are other measures typically used to evaluate the overall performance of estimators, including the root-mean-square error (RMSE) and the absolute mean squared error (AMSE). LMSE is a simple monotonic transformation of the RMSE (LMSE $= 0.5 * \text{Log}(\text{RMSE})$) that was chosen to avoid problems with the scale of the statistic, but has no impact on its interpretation.

[8] For the logit model, the coefficients for the grouped cohorts can be used to estimate the average relative risk of exiting unemployment between t and $t+1$, for all individuals in that cohort. For the GH model, the coefficients for the grouped cohorts capture the relative risk of exiting unemployment between t & $t+1$, $t+1$ & $t+2$ and $t+2$ & $t+3$.

Table 2 Monte Carlo simulation summary

Variable	Population coefficients	Panel sample			GH: Pseudo-panel data				Panel sample with noise	
		Bias	LMSE	Sim. SE	Bias	LMSE	Sim. SE	Avg. SE	Bias	LMSE
Age	0.802	0.007	− 7.028	0.029	0.006	− 6.429	0.040	0.041	− 0.038	− 5.845
Age^2	− 0.010	0.000	− 15.811	0.000	0.000	− 15.202	0.000	0.001	0.000	− 14.607
Years of education	0.500	− 0.001	− 9.736	0.008	− 0.002	− 8.438	0.015	0.015	− 0.030	− 7.010
Sex	− 0.504	− 0.001	− 7.184	0.028	0.002	− 6.920	0.031	0.033	0.028	− 6.356
Race	0.401	− 0.001	− 7.151	0.028	− 0.001	− 6.986	0.030	0.032	− 0.022	− 6.652
Duration cohort										
2 Periods	0.245	− 0.004	− 6.248	0.044	− 0.007	− 3.952	0.139	0.144	− 0.043	− 5.664
3 Periods	0.803	− 0.003	− 6.080	0.048	− 0.017	− 3.992	0.135	0.137	− 0.112	− 4.257
4 Periods	1.564	0.000	− 5.898	0.052	− 0.005	− 3.984	0.136	0.147	− 0.200	− 3.159
5 Periods	2.319	− 0.002	− 5.714	0.057	− 0.004	− 3.722	0.156	0.167	− 0.278	− 2.526
6 Periods	2.918	− 0.013	− 5.374	0.067	− 0.019	− 3.252	0.196	0.199	− 0.344	− 2.166
7 Periods	3.287	0.007	− 4.965	0.083	− 0.012	− 2.867	0.238	0.249	− 0.360	− 1.956
8 Periods	3.391	− 0.018	− 4.543	0.102	− 0.048	-2.242	0.323	0.325	− 0.393	− 1.893
9 Periods	3.308	− 0.017	− 4.180	0.123	− 0.026	− 1.681	0.431	0.427	− 0.385	− 1.898
10–12 Periods*	2.781	− 0.027	− 4.312	0.113	− 0.338	− 1.720	0.254	0.253	− 0.361	− 2.080
13–18 Periods**	1.362	0.008	− 3.571	0.168	− 0.581	− 0.534	0.499	0.519	− 0.215	− 2.572
Constant	− 22.263	− 0.126	− 1.036	0.582	− 0.125	− 0.158	0.916	0.945	1.161	0.699

*,**For the GH pseudo-panel estimation we use a $k = 3$ for the cohort 10–12 periods and $k = 6$ for the cohort 13–18. Bias is the average difference between the population coefficients and the sample estimates. LMSE is the log mean squared error. For the GH pseudo-panel estimator, we also provide the simulation-based standard errors, and the average standard errors obtained from each simulation using clusters at the individual level. The results are obtained from 1000 random samples

worth mentioning that the simulation-based standard errors are almost identical to the average standard errors estimated using clusters at the individual level.

The fourth set of columns in Table 2 provides the summary statistics for the logit model results when data with 3% induced error are used. This can be considered as a scenario where one uses panel data, but some of the records are mismatched due to errors during data collection or record linking. This is similar to what one would face when analyzing the short CPS panel data. In this case, the panel data estimator performs poorly, exhibiting large bias for all estimated coefficient, greatly increasing the LMSE of the model. It is worth mentioning that this induced error does not affect the GH estimator, because it does not use information on future unemployment status, but instead makes use of observed unemployment spell duration. In this case, GH estimator outperforms the panel data model with matching errors.

While not exhaustive, this Monte Carlo simulation shows that the GH estimator is a feasible strategy to estimate duration/survival type of models, when panel data are not available. When panel data are available, the simulations suggest that making use of the panel data information provides better estimates than the GH estimator, mostly due to smaller standard errors. However, if the panel data are affected by mismatch records, the simulation suggests that the GH estimator is superior to the logit model because of the increased bias on the estimations.

In the next section, we apply the GH methodology to explore the determinants of unemployment duration and the role of immigrants' social networks on their unemployment duration around the Great Recession in the USA.

4 Role of social networks on immigrant unemployment duration around Great Recession: an application of GH model

4.1 Background

There is substantial evidence indicating that social networks improve ethnic minorities and migrants' labor market outcomes mostly through job referrals, facilitating their transition into employment (Granovetter 1995; Ionnides and Loury (2004): Munshi 2003; Mouw 2003; Patel and Vella 2013 to name a few), but little is known regarding how social networks affect immigrant unemployment duration. Nevertheless, a few studies have focused on this particular link between networks and unemployment duration for the population in general and mostly for immigrants in Europe (for example Cingano and Rosolia 2012; Patacchini and Zenou 2012; Uhlendorff and Zimmermann 2014).[9] The heterogeneity of unemployment duration for immigrants has not been explored for the USA, and particularly the role of social networks on unemployment duration for the USA is missing in the literature. Because social networks play an important role in immigrants' job searches, it raises an important question of

[9] Focused on Germany, Uhlendorff and Zimmermann (2014) find that migrants are more likely to experience longer unemployment duration despite staying at their jobs for similar lengths of time when compared to natives with similar observable and unobservable characteristics. Diop-Christensen and Pavlopoulus (2016) find similar results for 12 European countries, but conclude that immigrants benefit more from increase in demand for low-skilled workers.

whether or not networks also lower their unemployment duration, particularly around an economic crisis such as the Great Recession. Around the Great Recession, the USA lost over 7.5 million jobs, with an unemployment rate that surpassed the 10% mark and a rapid increase in unemployment duration (Farber and Valletta 2015; Grusky et al. 2011). These increases in unemployment duration in combination with extended unemployment benefits generated a burden on the economy. In this framework, it is important to understand the role of social networks as a mechanism to reduce the effects that unemployment may have on immigrants around a period of economic crisis such as the Great Recession.[10]

While one may consider immigrant's social networks to facilitate information during periods of economic growth, during periods of economic crisis such as the Great Recession, social networks may have a negative impact on unemployment duration because of increasing competition among immigrants, or have no effect because the quality of the information of the networks declines as the share of unemployed immigrants increases. Moreover, the effect of networks on unemployment duration depends on how long immigrants have been unemployed. Calvo-Armengol and Jackson (2004) present a theoretical model showing that the effect of networks on employment outcomes depends on the initial state of the networks and on the length of time the agent has been unemployed. The longer an individual is unemployed, the lower are her chances of finding a job due to duration dependence but also because the quality of networks worsens and her networks consist of more unemployed migrants, which are less helpful in job searches. A larger number of unemployed individuals have less information on vacancies and fewer contacts for finding potential good job leads. Moreover, with a larger share of unemployed migrants in her network there is a larger competition for the same jobs.

The literature also suggests that immigrants and natives have had different experiences regarding their unemployment outcomes through the Great Recession in the USA. As a result of the housing, bust immigrants' labor outcomes deteriorated faster than for natives mainly because less educated Hispanic immigrants are often employed as independent contractors and temporary workers making their jobs very sensitive to the business cycle (Orrenius and Zavodny 2009; Mundra 2019). However, some research has shown that higher mobility among certain immigrant groups has helped them in their labor market outcomes around the Great Recession (Zhu et al. 2014; Cadena and Kovak 2016). Under financial stress, as experienced by many households in the recent financial meltdown and sub-prime crisis, immigrants may have relied on their social networks for income and financial support.[11] On the other hand, as described in Liu and Edwards (2015), in contrast to social networks arguments, immigrants may have suffered negative externalities due to the presence of a larger concentration of immigrants in the local labor market by facing tougher competition and having worse employment prospects.

[10] According to the Congressional Budget Office, federal budget spending on unemployment insurance benefits increased almost five times from 33 billion in 2004 to 155 billion in 2011. For households, increased unemployment not only lowered their income and hence their standard of living, but also reduced their chance of reintegrating back into the labor market.

[11] See Mundra and Uwaifo-Oyeler (2018) for more details.

In summary, the larger the size of the immigrant network the greater is the potential pool of job information and the stronger are immigrant's chances of finding employment in the labor market, reducing their unemployment duration. However, networks might not be effective in job searches during the time of a national slowdown on the scale of the Great Recession. As indicated before, in periods of long economic stress with high unemployment and increasing unemployment duration, the quality of a network might decline rapidly and the likelihood of competition might increase, thereby reducing its effectiveness in job searches and in lowering unemployment duration.

4.2 Data, Specification and Model Implementation

The data used in this empirical application are constructed from the monthly CPS obtained from Integrated Public Use Microdata Series (IPUMS) for the years 2001–2013.[12] We restrict the sample to unemployed people between 20 and 64 years of age, who were born in a foreign country, excluding individuals born to American parents.[13]

Because networks are generally difficult to identify, we measure networks using a concentration index that captures the population share of immigrants at the state level who originated from the same birth country. This measure, which is common in the literature, is a proxy for potential network in the area where they live because the larger is the concentration of immigrants, the higher is the probability of the individual connecting with members of her network.[14] Specifically, for each immigrant, networks are measured as the average share of the population who migrated from the immigrant's birth country and lived in the same state as the immigrant during the previous calendar year. Thus, for an immigrant surveyed in February of 2005 her birth-country network is measured using information from January 2004 to January 2005. The network measures are estimated using survey weights and concentrate on the total employed and non-employed population 15 years or older, excluding individuals born abroad to American parents.

While intuitively one might prefer to measure networks based on small geographical areas, capturing impacts in the local labor markets, there are many arguments that justify the use of a state-level network variable. First, immigrants, who are disproportionately low-skilled workers, are more mobile across MSAs (Cadena and Kovak 2016), but less likely to move across states (Kritz and Nogle 1994; Gurak and Kritz 2000), making a statewide measure less sensitive to within-state migration of the immigrants. Second, the record of high unemployment rates and a weak labor market observed through the Great Recession might have increased competition for jobs among immigrants in small geographical areas. Because of this, measuring networks

[12] According to the National Bureau of Economic Research, the Great Recession is defined as the period between December 2007 to June 2009, the pre-recession covers the period from January 2001 to November 2007, and the post-recession covers the period from July 2009 to December 2013.

[13] Unemployed immigrant workers are identified using self-reported unemployment status based on their activities during the week previous to the interview. For this individuals, the current length of unemployment spell is measured using the reported number of consecutive weeks that individual has been looking for work.

[14] Similar measures have been used in the literature by McConnell and Akresh (2008), Munshi (2003). Rauch and Trindade (2002), and Mundra (2005), to name a few.

at the local labor market level may be more likely to capture a competition effect, rather than the information-spreading effect of networks. Third, using small geographical areas may introduce an attenuation bias on the measured effects of immigration because of the larger noise-to-signal ratio that is introduced when estimating variables like the network measure used in this paper (Aydemir and Borjas 2011). State-level measures are a compromise between measuring immigrant networks with a smaller error and obtaining more precise estimates of immigrant networks at the local labor markets.[15]

For the implementation of the GH estimator, data are arranged in a base ($\mathbb{S}_b$) and a continuation ($\mathbb{S}_c$) samples across the full range of duration intervals or classes, which are pooled together for analysis. This data preparation is similar to the one described previously for the Monte Carlo simulation exercise and follows Valletta (2013). Individuals unemployed for 5–8 weeks in months' t are paired with those unemployed to <5 weeks in $t-1$; those unemployed for 9–12 weeks in months t to those unemployed for 5–8 weeks in $t-1$; third, 13–16 weeks in month t to 9–12 weeks in $t-1$; fourth, 27–39 weeks in t to 13–26 weeks in $t-3$; fifth, 53–78 weeks in month t to 27–52 weeks in $t-6$; and sixth, 105 + weeks in month t to 53–104 weeks in $t-12$.

In order to analyze the impact of networks on unemployment duration, we implement the GH estimator as described in Sect. 2. In addition to the standard individual demographic controls (sex, race, age and education), we control for a comprehensive set of individual and labor market characteristics that could influence immigrants' unemployment duration status, including whether the immigrant has been in the USA < 10 years (*Recent Migrant*) and whether the immigrant is a naturalized citizen.

To account for the effect of extended unemployment benefits, we control for the maximum number of weeks people can potentially benefit from unemployment insurance at the state level (Farber and Valletta 2015). We also control for homeownership, as homeowners may be less mobile and more attached to the local labor market and hence face higher unemployment duration during an economic downturn (Blanchflower and Oswald 2013). In the last decade, some states have passed employment verification laws to protect native and legal immigrant employment against undocumented immigrants, particularly for the unskilled group.[16] To control for the differences and changes across time regarding employment of immigrants and market regulations, we introduce as control an E-Verify variable that takes the value of 1 if there is a partial implementation of the initiative and 2 if there is full implementation. We also control for whether the state implemented a policy like E-Verify

To measure the health of the labor market we use the share of employed migrants and non-migrants between the ages of 20 to 64 for each state, year and month. State fixed effects are included to account for unobserved time-invariant factors that we are not

[15] For more discussion on how networks operate at the state level, see Mundra and Uwaifo-Oyeler (2018).

[16] E-Verify is a Govt. policy which allows enrolled employers to verify the eligibility of their employees to work in the USA. Studying the impact of 2007 Legal Arizona Workers Act (LAWA), the first E-Verify law to be passed, Bohn et al. (2014) show that in response to this Government policy there was a substantial decrease in the state's unauthorized population and that LAWA failed to improve the labor market outcome of legal low-skilled workers who compete with undocumented immigrants in the state.

able to control otherwise. Since immigrants with different origins and backgrounds might behave differently, or be treated differently, we include a set of dummies to capture the general region from which the immigrant originates. To account for the time dependence factor, we also include dummies to indicate how long the immigrant has been unemployed, based on the unemployment duration class described previously.

In addition to these characteristics, we include a recession dummy for months December 2007–June 2009 and post-recession dummy for July 2009–December 2013, to control for the recession effect in the model. Our key variable of interest *Network* is included in the model alone and interacted with the two recession dummies. The coefficient on the two interaction terms, *Network*Recession* and *Network*PostRecession*, will help to identify the difference in the conditional unemployment probability of immigrants during a recession and post-recession period compared to the pre-recession period. If networks are relied upon more during an economic crisis, particularly at the state level, these coefficients will be positive and statistically significant. We report the marginal effects based on the identified latent logit model (Eq. 17).[17]

4.3 Findings

Table 3 reports the coefficients for all explanatory variables of the benchmark model and different specifications for robustness using different measures of networks, and Table 4 provides the marginal effects of the main variable of interest, networks, using benchmark model and various specifications for robustness. A positive coefficient indicates that larger magnitudes of the variables are associated with a higher probability of unemployment continuation rates and hence longer unemployment duration, whereas a negative coefficient indicates the opposite. Because networks are measured using information from the previous 12 months, all results are clustered at the state and birth-country levels. We find evidence of a U-shape relationship between unemployment duration and age. Women and immigrants with less than high school education are more likely to remain unemployed for longer periods. We also find that household size, number of children, being the head of household (or spouse) and race are not related to the length of unemployment duration. Similar to findings in the earlier literature (Valletta 2013; Blanchflower and Oswald 2013), we find no evidence that being a homeowner affects unemployment duration among immigrants. The controls for assimilation factors indicate that while being a recent immigrant (ten years or less since moving to the USA) is not related to longer unemployment spells, being a US citizen increases the probability of an immigrant remaining unemployed.

Regarding employment regulations, the variables identifying the level of implementation of E-Verify have the expected sign but are not statistically significant. Living in a healthier local labor market with high employment shares reduces the probability of staying unemployed and thus reduces unemployment duration. We find that even after taking account of regional origin heterogeneity and with detailed labor market controls our baseline results hold; immigrants with larger networks have a higher probability of transitioning out of unemployment.

[17] Data descriptive is given in the Supplementary Material.

Table 3 Baseline and robustness of networks model: GH-ML coefficient estimates

	Baseline	ACS	Census 1990	Census 2000	Using MSA level network	CPS measure MSA consistent sample
		(1)	(2)	(3)	(4)	(5)
Networks	− 1.536**	− 1.572***	− 1.262***	− 1.494***	− 0.666*	− 1.325*
	(0.673)	(0.579)	(0.452)	(0.559)	(0.344)	(0.765)
Networks × Recession	− 0.0870	− 0.329	0.271	− 0.0626	0.107	− 0.501
	(1.087)	(1.040)	(0.705)	(1.013)	(0.646)	(1.115)
Networks × Post-recession	− 0.772	− 0.611	− 0.0507	− 0.287	0.0284	− 1.087
	(0.646)	(0.653)	(0.512)	(0.654)	(0.813)	(0.907)
Recession	0.801***	0.801***	0.790***	0.788***	0.886***	0.922***
	(0.167)	(0.166)	(0.165)	(0.160)	(0.267)	(0.270)
Post-recession	0.666***	0.647***	0.659***	0.632***	0.685**	0.724***
	(0.192)	(0.186)	(0.188)	(0.181)	(0.301)	(0.270)
Household head or spouse	0.0464	0.0529	0.0631	0.0557	− 0.00217	− 0.00755
	(0.053)	(0.052)	(0.053)	(0.052)	(0.060)	(0.057)
Married	0.0330	0.0352	0.0155	0.0335	0.0502	0.051
	(0.042)	(0.043)	(0.042)	(0.043)	(0.051)	(0.050)
Age	− 0.00928	− 0.00915	− 0.0106	− 0.00804	− 0.0133	− 0.0136
	(0.013)	(0.013)	(0.013)	(0.012)	(0.016)	(0.016)
$Age^2/100$	0.0378**	0.0369**	0.0399**	0.0360**	0.0470**	0.0471**
	(0.016)	(0.016)	(0.017)	(0.016)	(0.021)	(0.021)
Women	0.226***	0.223***	0.235***	0.223***	0.248***	0.253***
	(0.040)	(0.040)	(0.039)	(0.039)	(0.056)	(0.058)
HS education + some college	0.0721	0.0639	0.0814*	0.075	0.0625	0.0514
	(0.047)	(0.045)	(0.048)	(0.046)	(0.054)	(0.050)

Table 3 continued

	Baseline	ACS	Census 1990	Census 2000	Using MSA level network	CPS measure MSA consistent sample
		(1)	(2)	(3)	(4)	(5)
College or grad school	0.0942*	0.104*	0.127**	0.105*	0.102	0.0882
	(0.055)	(0.055)	(0.058)	(0.054)	(0.065)	(0.068)
Household size	0.0160	0.0203	0.0179	0.0194	0.0109	0.0104
	(0.016)	(0.015)	(0.016)	(0.016)	(0.018)	(0.019)
Number of children	− 0.0174	− 0.0264	− 0.0239	− 0.0262	− 0.0136	− 0.0107
	(0.019)	(0.019)	(0.019)	(0.019)	(0.023)	(0.025)
White	− 0.0634	− 0.0378	− 0.0476	− 0.0418	− 0.0571	− 0.0798
	(0.086)	(0.085)	(0.089)	(0.086)	(0.098)	(0.089)
House owner	− 0.0180	− 0.0215	− 0.017	− 0.0213	− 0.034	− 0.0277
	(0.042)	(0.042)	(0.043)	(0.042)	(0.051)	(0.059)
Recent migrant (10 years or less)	0.0269	0.00635	0.0252	0.016	0.016	0.013
	(0.068)	(0.064)	(0.069)	(0.064)	(0.083)	(0.077)
US citizen	0.242***	0.246***	0.245***	0.245***	0.260***	0.257***
	(0.052)	(0.050)	(0.053)	(0.051)	(0.063)	(0.051)
Partial E-Verify	− 0.0744	− 0.0532	− 0.0821	− 0.0565	− 0.126	− 0.147
	(0.130)	(0.123)	(0.126)	(0.120)	(0.154)	(0.173)
Full E-Verify	0.160	0.19	0.181	0.176	0.000595	− 0.0116
	(0.386)	(0.388)	(0.418)	(0.382)	(0.451)	(0.486)
Similar to E-Verify Policy	0.134*	0.127	0.121	0.144*	0.192	0.192*
	(0.081)	(0.077)	(0.085)	(0.079)	(0.124)	(0.099)
ln(Unemployment weeks benefits)	− 0.504**	− 0.502**	− 0.537**	− 0.506**	− 0.609*	− 0.594
	(0.210)	(0.206)	(0.210)	(0.205)	(0.345)	(0.372)

Table 3 continued

	Baseline	ACS	Census 1990	Census 2000	Using MSA level network	CPS measure MSA consistent sample
		(1)	(2)	(3)	(4)	(5)
State-level employment share	− 0.3302***	− 32.55***	− 33.75***	− 32.53***	− 40.41***	− 40.31***
All workers 20− 64	(0.0644)	(6.294)	(6.315)	(6.209)	(11.377)	(11.450)
Time dependence						
5–8 weeks unemp	1.075***	1.079***	1.124***	1.084***	1.139***	1.131***
	(0.175)	(0.177)	(0.189)	(0.180)	(0.262)	(0.243)
9–12 weeks unemp	0.920***	0.920***	0.917***	0.918***	0.987***	0.989***
	(0.147)	(0.147)	(0.145)	(0.146)	(0.218)	(0.216)
13–26 weeks unemp	− 0.857***	− 0.851***	− 0.869***	− 0.849***	− 0.848***	− 0.829***
	(0.103)	(0.103)	(0.109)	(0.103)	(0.135)	(0.126)
27–52 weeks unemp	− 0.526***	− 0.504***	− 0.536***	− 0.507***	− 0.582***	− 0.559***
	(0.135)	(0.135)	(0.141)	(0.135)	(0.183)	(0.158)
53–104 weeks unemp	− 1.840***	− 1.815***	− 1.864***	− 1.816***	− 1.778***	− 1.752***
	(0.221)	(0.219)	(0.234)	(0.223)	(0.290)	(0.236)
_cons	25.58***	25.09***	26.07***	25.00***	31.24***	31.18***
	(5.477)	(5.280)	(5.292)	(5.207)	(9.474)	(9.577)
alpha_cons	1.137***	1.131***	1.147***	1.132***	1.190***	1.184***
	(0.054)	(0.054)	(0.053)	(0.054)	(0.072)	(0.065)
N	125778	122742	120975	123415	102437	101749

All models include the region of origin and state fixed effect
*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

Table 4 GH-ML estimate: average marginal effects

	Baseline model	ACS	Census 1990	Census 2000	Using MSA level network	CPS data MSA consistent sample
	(1)	(2)	(3)	(4)	(5)	(6)
Variables of interest						
Overall network	− 0.349	− 0.351	− 0.227	− 0.302	− 0.109	− 0.325
	(0.089)	*(0.081)*	*(0.072)*	*(0.079)*	*(0.095)*	*(0.100)*
	[0.000]	**[0.000]**	**[0.002]**	**[0.000]**	**[0.251]**	**[0.001]**
Network pre-recession	− 0.307	− 0.315	− 0.25	− 0.3	− 0.128	− 0.256
	(0.134)	*(0.116)*	*(0.090)*	*(0.111)*	*(0.060)*	*(0.145)*
	[0.022]	**[0.006]**	**[0.005]**	**[0.007]**	**[0.034]**	**[0.078]**
Network recession	− 0.264	− 0.313	− 0.16	− 0.256	− 0.081	− 0.266
	(0.125)	*(0.141)*	*(0.114)*	*(0.147)*	*(0.105)*	*(0.100)*
	[0.035]	**[0.026]**	**[0.162]**	**[0.080]**	**[0.442]**	**[0.008]**
Network post-recession	− 0.397	− 0.38	− 0.223	− 0.31	− 0.101	− 0.384
	(0.098)	*(0.094)*	*(0.081)*	*(0.089)*	*(0.144)*	*(0.127)*
	[0.000]	**[0.000]**	**[0.006]**	**[0.001]**	**[0.486]**	**[0.002]**
Time dependence dummies	x	x	x	x	x	x
Recession × network interaction	x	x	x	x	x	x
Demographic characteristics	x	x	x	x	x	x
State-level characteristics	x	x	x	x	x	x
Region of origin FE	x	x	x	x	x	x
State fixed effects	x	x	x	x	x	x

Standard errors in parentheses in italics; *p* values in square brackets in bold. Standard errors were clustered at the state-birth country level. Networks measured at the state level unless mentioned at the MSA level

Table 4 reports the overall average marginal effect of networks as well as the marginal effects of networks for the pre-recession, recession and post-recession period. For our preferred benchmark model from col 1 in Table 3, we find that a 1 pp increase in birth-country share lowers the unemployment continuation probability of immigrants by 0.35 pp over the entire period of analysis. During the pre-recession period, this reduction is 0.31 pp; during the recession period, this reduction drops to 0.26 pp; but during the post-recession period, this reduction rises to 0.40 pp. All these effects are statistically significant. In summary, after controlling for all individual- and state-level characteristics in various specifications, we find that immigrants benefit from larger networks as this variable remains negatively related to longer unemployment spells. After taking into account the impact of the recession timing factors, the marginal effect of networks in reducing the average likelihood of remaining unemployed after the recession is almost 50% larger than their effect during the recession period (Table 4—column 1).

As described in the data section, the analysis until now uses monthly CPS data to measure the size of immigrant networks, based on the average share of people living in any given state during the previous 12 months. We expect that using this lagged information would reduce potential endogeneity problems caused by using contemporaneous data, while allowing us to pool a larger body of data to obtain a more accurate measure of networks compared to data from a single month. Nevertheless, there is a possibility that this network measure is not exogenous and the estimates may be inconsistent. We explore various alternative measures of birth-country networks other than CPS in our benchmark model. We present three estimates where the network variable is measured using annual data from ACS and using data drawn from the 5% Census data for the years 1990 and 2000, all obtained from the IPUMS. The coefficient of the models is given in (cols 2–4) Table 3. Our estimation indicates that using both of these network measures provides results consistent with the preferred model, with somewhat smaller marginal effects when data from the 1990 Census are used (cols 3 and 4) in Table 4.

There is also a possibility that using state-level data is not appropriate to capture network effects on unemployment duration. Column 5 in Table 4 gives the estimates of the average marginal effect from the MSA level network, and this effect is almost three times smaller than the benchmark model and is insignificant, except for the pre-recession period. It is also possible that the estimates are smaller because we are using a sample that excludes individuals whose MSA cannot be identified. To account for the sample change, we re-estimate the benchmark model using the state-level network measure and the same sample as in col 5. Results are given in col 6. We find that the network effect for those living in Metro areas is smaller compared to the benchmark, but statistically significant and larger than the estimate in col 5, and the marginal effect of networks during and post-recession is significant.

To further examine the sensitivity of using survey data such as CPS we restrict the sample to immigrants for whom our estimated network size is at least 0.1%, 0.25% and 0.5%, which imply a reduction in the sample by 6.6%, 18.1% and 29.6%, respectively, compared to the benchmark sample. In general, while some results are less statistically

significant, the core results of our model remain consistent.[18] As a final robustness check, to address concerns of endogeneity, we use Bartik-type instrument (Bartel 1989, Bartik 1991) using the tendency of migrants to move to areas with already high concentrations of migrants as an instrument for our networks measure. Based upon these estimates, the Bartik consistent sample and the Bartik IV estimates are consistent with the benchmark results.[19]

To test if there are any differential impacts of networks based on unemployment duration, we estimate an extension of our baseline model (Table 4—column 1) by including interactions of *Network* and *Recession* dummies with the unemployment duration categories. In Table 5, we present the marginal effect of networks estimated at different points of unemployment duration and for the pre-recession, recession and post-recession periods using unrestricted interaction between duration dummies with the network and recession dummies.[20] We find that after including interactions between network size and duration, there is a differential effect before and after the recession. Across all specifications, larger networks significantly lower the unemployment continuation probability for immigrants with short unemployment spells (1–4 weeks), with the strongest effect in the post-recession period. We also find that larger networks reduce the unemployment continuation probability of immigrants with long unemployment spells before the recession (27–104 weeks). This effect decreased during and after the recession.

To put our findings into context, we estimate the average probability of continuation of unemployment at different network sizes for immigrants who have been unemployed between 1 and 4 weeks and for those who have been unemployed between 27 and 52 weeks for each of the three periods. In this case, we use the estimates of the preferred model and make predictions about the probability of unemployment continuation for various simulated changes in network size. Figures 1 and 2 display these effects for immigrant groups who have been unemployed for 1–4 weeks and 27–52 weeks, respectively, using the coefficients for pre-recession, recession and post-recession. We see that for immigrants who have been unemployed for 1–4 weeks larger networks help to lower unemployment duration for all time periods and for the post-recession period immigrants with larger birthplace networks have a lower risk of continuing to be unemployed versus immigrants with smaller networks (Fig. 1). However, this is not the case for immigrants with 27–52 weeks of unemployment, as the predicted post-recession continuation probability is almost the same regardless of network size (Fig. 2).[21]

Overall, examining whether immigrant social networks have a significant differential effect over various duration categories we find that networks are significantly more effective in lowering unemployment duration over the first 2 months of unemployment duration, and this effect further increased after the recession. However, networks are

[18] The marginal effects are given in the Supplementary Material.

[19] Bartik results are given in the Supplementary Material

[20] Bartik method and the results are given in the Supplementary Material.

[21] Figures 1 and 2 provide the predicted probability that an average person in the sample would remain unemployed for any given network size. Standard errors and confidence intervals are available upon request.

Table 5 Networks and duration dependence: average marginal effects of networks across various unemployment duration categories

	Pre-recession	Recession	Post-recession
1–4 weeks	− 0.464	− 0.428	− 1.382
	(0.234)	*(0.149)*	*(0.262)*
	[0.047]	**[0.004]**	**[0.000]**
5–8 weeks	− 0.588	− 0.379	0.035
	(0.177)	*(0.130)*	*(0.190)*
	[0.001]	**[0.004]**	**[0.855]**
9–12 weeks	0.120	− 0.156	− 0.429
	(0.226)	*(0.127)*	*(0.175)*
	[0.595]	**[0.218]**	**[0.014]**
13–26 weeks	0.063	− 0.200	− 1.013
	(0.190)	*(0.171)*	*(0.209)*
	[0.742]	**[0.241]**	**[0.000]**
27–52 weeks	− 0.954	− 0.373	0.040
	(0.217)	*(0.374)*	*(0.319)*
	[0.000]	**[0.319]**	**[0.901]**
53–104 weeks	− 0.344	− 0.058	0.181
	(0.570)	*(0.422)*	*(0.314)*
	[0.546]	**[0.891]**	**[0.564]**

Standard errors in parentheses in italics; *p* values in square brackets in bold. Standard errors were clustered at the state-birth country level

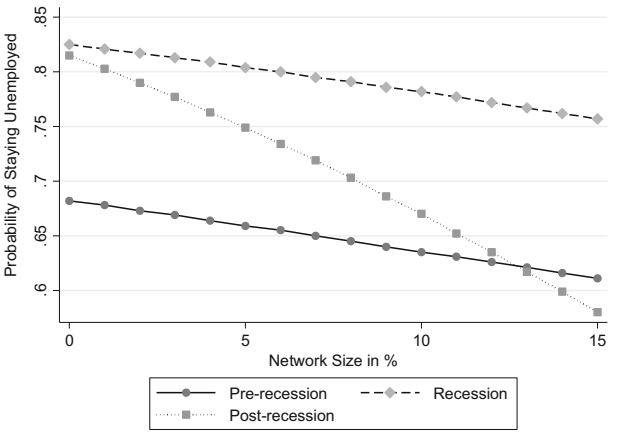

Fig. 1 Predicted probability of staying unemployed: effect of network size for immigrants who are unemployed < 4 weeks

ineffective when immigrants have been unemployed for a longer period of unemployment.

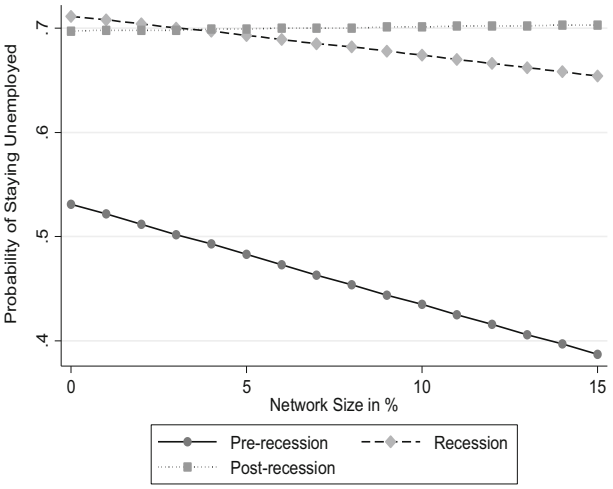

Fig. 2 Predicted probabilities of staying unemployed: effect of network size for immigrants who are unemployed between 27 and 54 weeks

5 Conclusion

This paper presents a detailed exposition of the methodology proposed by Guell and Hu (2006) for the implementation of duration/survival analysis when panel data are unavailable. Using a Monte Carlo simulation, we show that a logit model using panel data performs better than the GH model. However, GH model using cross-sectional survey data with uncompleted unemployment spells does better than the inadequate panel data. This suggests that in the absence of panel data or limited availability of panel data, the GH estimator can be used for duration analysis to obtain robust econometric estimates at the individual level using repeated cross-sectional survey data.

We apply the GH estimator to explore the role of immigrant birth-country networks on immigrant unemployment duration, particularly around the Great Recession, using CPS monthly data on unemployed immigrants over the years 2001–2013. Using the share of immigrants from the same country of origin as a measure of network size and based on our preferred results, we find that birth-country networks have a significant effect on lowering the duration of unemployment and this is stronger during the pre- and post-recession period than during the recession. This finding persists if we use network measures calculated from other data sources, use Bartik Type IV estimation method, or measure networks at the MSA level instead of at the state level.

Interestingly, when we allow the role of networks to vary with unemployment duration we find some evidence that networks are more effective in lowering unemployment duration for immigrants who have been unemployed for a shorter period. We find that the risk of continuing to be unemployed is significantly lower for immigrants with 1–4 weeks of unemployment duration and this effect is the strongest during the post-recession period.

This paper provides an important insight into the novel method proposed by Guell and Hu (2006), and using a simple Monte Carlo demonstrates the strength of this

method in duration analysis when the continuous duration data at an individual level are not available. Furthermore, this paper gives a detailed implementation of these methods using CPS survey data to examine the unemployment duration and the role of social networks around the Great Recession and makes an important contribution toward the GH method that becomes more accessible to analyze similar issues in the absence of adequate micro-level panel data.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Aydemir A, Borjas GJ (2011) Attenuation bias in measuring the wage impact of immigration. J Labor Econ 29(1):69–112

Baker M (1992) Digit preference in CPS data. Econ Lett 19:117–121

Bartel A (1989) Where do the new US immigrants live. J Labor Econ 7:371–391

Bartik TJ (1991) Who benefits from state and local economic development policies?. W. E. Upjohn Institute for Employment Research, Kalamazoo

Blanchflower DG, Oswald AJ (2013) Does high home-ownership impair the labor market. In: NBER working paper series working paper 19079

Bohn S, Lofstrom M, Raphael S (2014) Did the 2007 Legal Arizona workers act reduce the state's unauthorized immigrant population? Rev Econ Stat 96(2):258–269

Cadena BC, Kovak BK (2016) Immigrants equilibrate local labor markets: evidence from the great recession. Am Econ J: Appl Econ 8(1):257–290

Calvo-Armengol A, Jackson M (2004) The effects of social networks on employment and inequality. Am Econ Rev 94(3):426–454

Cingano F, Rosolia A (2012) People I know: job search and social networks. J Labor Econ 30(2):291–332

Dang H-A, Lanjouw P, Mckenzie D (2014) Using repeated cross-sections to explore movements into and out of poverty. J Dev Econ 107(C):112–128

Diop-Christensen A, Pavlopoulos D (2016) Institutions and structures as barriers? A comparison of native-born and immigrant unemployment durations across 12 European countries. Int J Social Welf 25(4):347–360

Farber H, Valletta R (2015) Do extended unemployment benefits lengthen unemployment spells? Evidence from recent cycles in the US labor market. J Hum Resour 50(4):873–909

Granovetter MS (1995) Getting a job: a study of contacts and careers, 2nd edn. University of Chicago Press, Chicago

Grusky DM, Western B, Wimer C (eds) (2011) The great recession. Russell Sage Foundation, New York

Guell M, Hu L (2006) Estimating the probability of leaving unemployment using uncompleted spells from repeated cross-section data. J Econom 133:307–341

Gurak T, Kritz MM (2000) The interstate migration of U.S. immigrants: individual and contextual determinants. Soc Forces 78(3):1017–1039

Ionnides YM, Loury LD (2004) Job information networks, neighborhood effects, and inequality. J Econ Lit 42:1056–1093

Kritz MM, Nogle JM (1994) Nativity concentration and internal migration among the foreign born. Demography 31(3):509–524

Liu CY, Edwards J (2015) Immigrant employment through the great recession: individual characteristics and metropolitan contexts. Soc Sci J 94(1):137–151

Mouw T (2003) Social capital and finding a job: do contact matter? Am Sociol Rev 68(6):868–870

Mundra K (2005) Immigration and international trade: a semiparametric empirical investigation. J Int Trade Econ Dev 14(1):65–91

Mundra K (2019) Minority and immigrant experience in the recent housing market: evidence from the 2009 American housing survey, immigration symposium. East Econ J 46(1):53–81
Mundra K, Uwaifo-Oyeler R (2018) Determinants of immigrant homeownership: examining their changing role during the great recession and beyond. Int Migrat Rev 53(2):648–694
Munshi K (2003) Networks in the modern economy: mexican migrants in the U. S. labor market. Q J Econ 118(2):549–599
Nickell SJ (1979) Estimating the probability of leaving unemployment. Econometrica 47(5):1249–1266
Orrenius PM, Zavodny M (2009) Tied to the business cycle: how immigrants fare in good and bad economic times. Migration Policy Institute
Patacchini E, Zenou Y (2012) Ethnic networks and employment outcomes. Reg Sci Urban Econ 42:938–949
Patel K, Vella F (2013) Immigrant networks and their implications for occupational choice and wages. Rev Econ Stat 95(4):1249–1277
Rauch J, Trindade V (2002) Ethnic Chinese networks in international trade. Rev Econ Stat 84(1):116–130
Rivera-Drew J, Flood S, Warren JR (2014) Making full use of the longitudinal design of the current population survey: methods for linking records across 16 months. J Econ Soc Meas 39(3):121–144
Sider H (1985) Unemployment duration and incidence: 1968–1982. Am Econ Rev 75(3):461–472
Uhlendorff A, Zimmermann KF (2014) Unemployment dynamics among migrants and natives. Economica 81(322):348–367
Valletta RG (2013) House lock and structural unemployment. Labour Econ 25:86–97
Zhu P, Liu CY, Painter G (2014) Does residence in the ethnic community help immigrants in a recession? Reg Sci Urban Econ 47:112–127

Affiliations

Kusum Mundra[1,2] · Fernando Rios-Avila[3]

✉ Kusum Mundra
kmundra@newark.rutgers.edu

Fernando Rios-Avila
friosavi@levy.org

[1] Department of Economics, Rutgers University, Newark, USA

[2] IZA, Bonn, Germany

[3] Levy Economics Institute, Bard College, Annandale-on-Hudson, USA