Estimating Gravity Models with High-Dimensional Fixed Effects On Privacy-Protected Data*

Jung Sakong† Alexander K. Zentefis‡

April 24, 2023

Abstract

We propose an econometric method to estimate gravity models with high-dimensional fixed effects on data protected by differential privacy. The method adapts the Method of Simulated Moments (MSM) to identify a large volume of fixed effects. We illustrate the method by estimating a gravity model of consumer flows to bank branches using privacy-protected geolocation data from mobile devices. We find significant differences between the estimates obtained from the proposed method and those obtained from standard gravity model estimation methods. The approach extends to a range of applications where high-dimensional fixed effects require estimation in MSM routines.

JEL classification: C15, C24, G21, R20
Keywords: differential privacy, simulation methods, location economics

*We give special thanks to Bo Honoré and Luojia Hu for suggesting parts of the econometric method we use in this paper and for their very helpful feedback. We are grateful to Tori Healey, Gen Li, Lizzie Tong, and Yi (Layla) Wang for their extraordinary research assistance. We thank Taha Ahsin, Treb Allen, Costas Arkolakis, Helen Banga, Asaf Bernstein, Mehdi Beyhaghi, Nick Buchholz, Lorenzo Caliendo, Claire Célérier, Judy Chevalier, Tony Cookson, Jess Cornaggia, Doug Diamond, Jonathan Dingel, Jane Dokko, João Granja, Cecilia Fieler, Raffi Garcia, Paul Goldsmith-Pinkham, Yaming Gong, Gary Gorton, Jessie Handbury, Xuan Hung Do, Stefan Jacewitz, Kristoph Kleiner, Noura Kone, Sam Kortum, Cameron LaPoint, Simone Lenzu, Xiang Li, Runjing Lu, Yuhei Miyauchi, Luciana Orozco, Piyush Panigrahi, Karen Pence, Matthieu Picault, Roberto Robatto, Claudia Robles-Garcia, Rosa Sanchis-Guarner, Shri Santosh, Katja Siem, Fiona Scott Morton, Brad Shapiro, Kelly Shue, Mike Sinkinson, Amir Sufi, Nancy Wallace, Brian Waters; and participants at various seminars and conferences for their very helpful comments. We also thank Jill Kelly, Miriam Olivares, Yichen Yang, the Marx Science and Social Science Library at Yale, Patricia Carbajales, Pat Claflin, Mazair Fooladi Mahani, and the Clemson Center for Geospatial Technologies for their valuable assistance in geoprocessing. The views expressed in this paper are those of the authors and do not reflect those of the Federal Reserve Bank of Chicago or the Federal Reserve System. All errors are our own.

†Federal Reserve Bank of Chicago; 230 La Salle St, Chicago, IL 60604 (email: jung.sakong@chi.frb.org)
‡Yale School of Management; 165 Whitney Ave, New Haven, CT 06510 (email: alexander.zentefis@yale.edu)

1 Introduction

Differential privacy is a system of methods protecting individual privacy in data sets while preserving the utility of the data for analysis (Dwork 2006). Differential privacy algorithms are masking more and more economic data sets over time. Just some examples are the 2020 Census tables and American Community Survey microdata (Ruggles, Fitch, Magnuson and Schroeder 2019), financial transactions data (Karger and Rajan 2020), and health records (Allen et al. 2020). To make it difficult to identify individuals in data sets, differential privacy algorithms commonly add noise to the data, while truncating and censoring low-valued observations. But these distortions introduce non-classical measurement error into the data, which hinder traditional econometric methods, such as OLS, from producing consistent estimates of economic model parameters.

This paper proposes an econometric method to estimate fixed-effects gravity models on data protected by differential privacy. Since Harrigan (1996) first introduced them, fixed-effects gravity models are now ubiquitous to the spatial economics and trade literature (Head and Mayer 2014). These models represent the flows of goods, people, or dollars between origins and destinations using a log-linear equation. Such an equation consists of origin and destination fixed effects, which are possibly time-varying, along with a measure of bilateral "accessibility" between locations. This term is regularly taken to be the physical distances between locations, but is often appended to include a vector of bilateral cost variables, such as tariffs, transport costs, or differences in culture, language, currency, or laws.

Methods to address econometric issues when estimating gravity models have been proposed before. One major issue was how to handle zero flows that are regularly observed in the data, despite a gravity model representing bilateral relations with strictly positive variables. In trade, Eaton and Tamura (1994); Helpman, Melitz and Rubinstein (2008); and Westerlund and Wilhelmsson (2011) suggest clever methods to account for zero flows, including Tobit procedures, two-step Heckman procedures, and Poisson fixed-effects estimators. Recognizing potential biases introduced when log-linearizing a gravity equation, Silva and Tenreyro (2006) suggest a Poisson pseudo-maximum-likelihood (PPML) procedure, which became seminal to the literature. Extensions of this estimator can handle a large number of fixed effects (Larch, Wanner, Yotov and Zylkin 2019), which becomes relevant when estimating gravity models with thousands or hundreds of thousands of origins and destinations. Such a high number of locations come into play when modeling very localized choices, say, of consumers selecting establishments to visit (Davis, Dingel, Monras and Morales 2019; Miyauchi, Nakajima and

Redding 2021), or of firms deciding the regions of foreign countries to export (Bricongne, Fontagné, Gaulier, Taglioni and Vicard 2012). When the data in a spatial gravity model's estimation is subject to differential privacy and many fixed effects require estimation, we hope this paper's econometric method can be of use.[1]

Our approach at estimating fixed-effects gravity models on privacy-protected data adapts Daniel McFadden's Method of Simulated Moments (McFadden 1989) to identify a large volume of fixed effects. The Method of Simulated Moments (MSM) estimates the parameters of a model by closely matching moments calculated from observed data with moments calculated from data that is simulated from the model. The parameters of the model iteratively update until the differences between observed and simulated moments are sufficiently small.

A key insight of our approach is to simulate data from a gravity model and then apply the same differential privacy algorithm to the simulated data that the data provider used to privacy-protect the real-world data. The parameters of the gravity model update until the computed moments from the simulated data (after being made "privacy-protected" per the algorithm) closely match the computed moments from the privacy-protected real-world data. The procedure requires researchers to know the differential privacy algorithm that the data provider adopted, but from our experience, data providers are generally open to sharing this information because differential-privacy algorithms are fairly standardized in the industry and revealing them does not undo the data's privacy protection.

Implementing MSM this way is straightforward in models with no or few fixed effects (Adda and Cooper 2003). But a fixed-effects gravity model with a very large number of origins and destinations, potentially generating billions of bilateral relations, severely complicates the procedure. We address this computational challenge by sampling origin-destination pairs using stratified sampling and applying probability weights to the sampled observations. The rest of the procedure involves simulating observations from the sampled pairs, applying the differential privacy algorithm to the simulated data, iterating the origin and destination fixed effects estimates until they converge in a Gauss-Siedel-style method (Guimaraes and Portugal 2010), and finally selecting estimates of the gravity model's remaining parameters that minimize the weighted sum of squared errors between the simulated model moments and the observed data moments.

We illustrate the econometric method by estimating a gravity model of consumer flows to bank branches using privacy-protected geolocation data from mobile devices. The observed

[1]For other articles proposing different methods to handle privacy protected data in estimations, see Agarwal and Singh (2023); Neunhoeffer, Sheldon and Smith (2023); Barrientos, Bowen, Snoke and Williams (2023).

data are the number of visitors from their home Census block groups to bank branches per month. We represent the consumer movements over time with a fixed-effects gravity model, consisting of (1) block group × time fixed effects, (2) bank branch × time fixed effects, and (3) the distances between pairs of block groups and branches multiplied by a time-varying gravity coefficient parameter β_t. To protect user privacy, the mobile device data provider adds noise to the number of visitors from a block group to a branch, and the provider truncates and censors these visitor counts if the number is too low.

We run the econometric procedure month-by-month to account for dynamic branch entry and exit. The monthly point estimates of the gravity coefficient range from -1.45 to -1.26, implying that, if a representative branch is located 1% farther away from a representative block group, the expected number of residents from that block group who travel to that branch will drop by around 1.26% to 1.45% per month. This range is in line with the gravity coefficient estimate of -1.05 that Agarwal, Jensen and Monte (2018) find for the average out-of-home purchase, where the authors evaluate how consumer expenditures in nonfinancial sectors vary with distances from merchants.

Differential privacy methods introduce non-classical measurement error into the data and bias the estimates from traditional econometric approaches. To assess the magnitude of the bias in our setting, we compare the β_t parameter estimates from our approach to estimates computed on the observed privacy-protected data using the two mainstream approaches in gravity model estimation: OLS regression and PPML estimation. The differences are stark. The traditional approaches deliver estimates that are roughly an order of magnitude smaller than the MSM estimates, ranging from -0.331 to -0.038 depending on the specification. The comparison reveals the downward bias that differential privacy methods introduce to traditional econometric approaches at estimating gravity models, and it stresses the need for the alternative MSM procedure.

Overall, we propose an econometric method to estimate gravity models with high-dimensional fixed effects on data protected by differential privacy, and we illustrate our method in an application involving privacy-protected geolocation data on consumer trips to bank branches. In our approach, we adapt the Method of Simulated Moments (MSM) to allow for the estimation of hundreds of thousands of fixed effects. While our focus is on estimating gravity models, the econometric approach in this paper extends to a range of applications where high-dimensional fixed effects require estimation in MSM routines.

Outline. The paper proceeds as follows. Section 2 presents a fixed-effects gravity model in the context of our application to banking. Section 3 describes the privacy-protected mobile device data we use to estimate a gravity model of consumer flows to bank branches. Section 4 details the econometric method for estimation. Section 5 presents the results. Section 6 concludes.

2 A Gravity Model of Branch Visits

In a companion paper (Sakong and Zentefis 2023), we describe how bank branches remain a vital source of bank participation in the United States, especially for low-income and Black households. But both types of households, whether banked or unbanked, are significantly less likely to visit bank branches compared to White and high-income households. In that paper, we estimate a gravity model of consumer flows to bank branches to better understand whether differences in access to branches or in demand for branch products and services explains the disparities. That paper uses the econometric method detailed in this paper.

The fixed-effects gravity model we consider is the log-linear equation:

$$\log\left(\text{No. of visitors}_{ijt}\right) = \gamma_{it} + \lambda_{jt} - \beta_t \log\left(\text{Distance}_{ij}\right) + \varepsilon_{ijt}. \tag{1}$$

The left-hand-side of Eq. (1) is the natural logarithm of the number of visitors from Census block group i to bank branch j in time period t. The right-hand-side of Eq. (1) includes four terms. The first term, γ_{it}, is a block group $\times$ time fixed effect that captures all characteristics of block group i's residents that contribute to them visiting any branch in the period. Informally, it represents factors that influence a block group's "demand" for branch products or services at any location (e.g., average wealth, income, financial sophistication, trust in banks, flexibility in time).

The second term, λ_{jt}, is a branch $\times$ time fixed effect that captures all characteristics of branch j that make it a destination for residents of any block group in the period. Informally, it represents factors that contribute to a branch's "quality" (e.g., the branch having attractive deposit or loan rates, higher staff attentiveness, or many ATMs that avoid long customer queues).

In the third term, the parameter β_t is the elasticity of visitor flows with respect to distance in the period. In many microfounded gravity models, the parameter can be interpreted as the product of residents' traveling costs and their elasticity of substitution between branches (Eaton and Kortum 2002; Ahlfeldt, Redding, Sturm and Wolf 2015). The term Distance_{ij} is

the geographic distance between block group i and branch j. In the estimation, we measure distance using the haversine formula, which accounts for the curvature of the Earth, and we compute the distances between branches and block groups' centers of population (see Footnote 8). The fourth term, ε_{ijt}, is a mean-zero disturbance.

Traditionally, OLS and PPML are used to estimate a fixed-effects gravity model like in Eq. (1). However, when the underlying data used in the estimation are distorted by differential privacy methods, the disturbance term ε_{ijt} may no longer satisfy the classical measurement error assumption. For instance, if low-valued branch visitor counts are dropped or bottom-coded to reduce the chances of identifying a particular visitor, the error term is not random, but systematically related to the true visitor count. The mobile device data we use in estimating Eq. (1) are subject to these kinds of distortions, and we describe that data next.

3 Geolocation Data on Branch Visitors

Branch visitors are based on geolocation data from mobile devices between January 2018 and December 2019. The data provider is the firm SafeGraph. The data are monthly and include both branch locations and information about branch visitors. We do not use the "raw" pings from individual mobile devices, but rather, we use SafeGraph's aggregated geolocation data that try to protect user privacy. Rather than reporting the physical whereabouts of an individual device through time, this aggregated data report the home Census block groups of branch visitors and the associated number of visitors from each block group per month. In essence, the data provide the network of consumer trips from home block groups to bank branches each month.

The aggregated data are benefited by elaborate algorithms that SafeGraph has developed to accurately estimate whether a mobile device visits a particular destination and to pinpoint a mobile device's home origin, using the device's reported pings over time. However, the data do not give the demographic attributes of the mobile device owners, nor their home addresses or starting points of their trips, nor their duration spent at a branch, nor what they do at the branch.

A visitor in the SafeGraph data is identified by a mobile device, one device is treated as one visitor, and a device must spend at least 4 minutes at an establishment to qualify as a visitor. Appendix A provides background information on the SafeGraph data and a detailed explanation of how we construct our primary sample. Here, we give a summary.[2]

[2]SafeGraph asks all researchers who use the company's data to include the disclaimer: "SafeGraph is a data

3.1 Primary Sample

Our primary (core) data set includes bank branches in all 50 states and the District of Columbia. SafeGraph categorizes businesses by their six-digit NAICS codes. To ensure that we only analyze depository institutions in the SafeGraph data, we take advantage of information from the FDIC's 2019 Summary of Deposits (SOD).

In our core sample, we include only businesses in SafeGraph with NAICS codes equal to 522110 (Commercial Banking), 522120 (Savings Institutions), or 551111 (Offices of Bank Holding Companies) whose brands are also listed in the SOD. For example, Wells Fargo & Company and SunTrust Banks, Inc. are two bank brands with branch locations in the SOD. We therefore include all Wells Fargo and SunTrust Bank branch locations in SafeGraph. We identify the physical locations of bank branches from SafeGraph's geographic coordinates, and not from the SOD's, as we found that SafeGraph's coordinates typically were more accurate.[3]

Our core sample is confined to bank branches for which SafeGraph has visitor data. Many bank locations recorded in SafeGraph lack such information, as it is often difficult to attribute mobile device visits to particular branches. There are two main reasons. First, in dense environments such as multi-story buildings or shopping malls, SafeGraph might not be confident about the geometric boundary of a place. Not knowing the boundary makes it awfully difficult to attribute visitors to a unique place that is part of a shared space. To reduce false attributions, SafeGraph instead allocates visitors to the larger "parent" space, such as the encompassing mall. Second, and related, a bank branch might be entirely enclosed indoors within a parent location (i.e., a customer must enter the parent's structure to reach the branch). Because mobile device GPS accuracy deteriorates severely within indoor structures, SafeGraph is reluctant to assign visitors to an enclosed branch. Instead, those visitors are aggregated to the level of the parent location. For example, many Woodforest National Bank branches are enclosed in Walmart Supercenters. (Walmart partners with Woodforest to provide the retail company's banking services.) Visitors to these enclosed branches cannot be separated from visitors to Walmart, and so, these branches are deprived of visitor data.[4]

company that aggregates anonymized location data from numerous applications in order to provide insights about physical places, via the Placekey Community. To enhance privacy, SafeGraph excludes census block group information if fewer than two devices visited an establishment in a month from a given census block group." The documentation to the SafeGraph data is here: SafeGraph Documentation.

[3]For most branches, the geographic coordinates in SafeGraph and the SOD matched. When the two sources disagreed, a Google Maps search of a branch address in the SOD often confirmed that no physical place existed at that address. (The place's absence was not due to a branch closing.)

[4]Regarding branch openings and closings, if a bank branch closed and SafeGraph were aware of its closure,

The SOD registers 86,374 bank branch locations as of 2019. While SafeGraph can account for 71,468 branches according to our core sample definition (83% coverage), only 51,369 of these places have visitor data and constitute our core sample. Our core sample thus covers around 60% of bank branches in the United States. Appendix Fig. A.1 presents a time-series of the number of branches per month in our core sample. Per month, the number of recorded branches is fairly stable and averages around 38,000.[5]

3.2 Sampling Bias

Our core sample experiences two types of sampling bias: (i) differential privacy and (ii) sample selection. We discuss each bias below and describe how we address it.

Differential Privacy. The first bias emerges from SafeGraph's efforts to preserve user privacy. The company applies differential privacy methods to avoid identifying people by their home locations. First, Safegraph adds Laplace noise to all positive counts of visitors to a branch from each home Census block group of the branch's visitors. Second, they round each of these block group $\times$ branch visitor counts down to the nearest integer. Third, they drop from the data all rounded visitor counts less than 2. Fourth, if a rounded visitor count equals 2 or 3, they raise it to 4. These last two data adjustments render our sample subject to both truncation from below and censoring from below, leading to non-classical measurement error. Fig. 2 presents the distribution of the observed (raw) visitor counts, which reveals both the truncation and censoring. Roughly 84% of the observed visitor counts equal 4, which implies a substantial amount of data distortion. The distortion also appears to vary by demographic attributes of residents. For example, in block groups with predominately Black residents (80%+), about 88% of visitor counts equal 4, whereas in the remaining block groups, about 83% equal 4.

Sample Selection. The second bias relates to sample selection, as our data on branch visitation patterns might not be representative of the true population behavior in the U.S. Potential sampling bias arises from two sources: our set of branches and our set of visitors.

any visitors to the building (say, if a new business opened there) would no longer be attributed to the branch. Likewise, if a branch opened and SafeGraph were aware of it, visitors would start being attributed to the branch. Nevertheless, if SafeGraph is unaware of a branch's opening or closing, visitors would be incorrectly attributed and count toward measurement error.

[5]We focus our analysis on depository institutions in this paper and leave for follow-up work the study of access to non-depository institutions, like credit unions, and non-traditional financial institutions, like check cashers and payday lenders.

To address potential sampling bias from missing around 40% of U.S. branches, in Section 3.3 we compare the representation of different demographic groups in the areas covered by our core sample of branches to the areas covered by all branches in the SOD. Overall, differences in demographic characteristics between the two sets of areas are precisely estimated, but small.

Regarding our sample of visitors, SafeGraph aggregates data from around 10% of all mobile devices in the country. We calculate about 30 million unique mobile devices per month on average visiting all businesses recorded in SafeGraph, and our core sample reports 1.6 million visitors to bank branches per month on average.[6] The 2010 U.S. Census records 217,740 Census block groups, and our core sample includes 215,686 unique visitor home block groups, implying close to complete coverage of U.S. local home areas.

Nevertheless, we cannot rule out non-random sampling of mobile devices based on unobserved characteristics of visitors. We do not know the precise demographic attributes of an individual bank branch visitor. The 2019 FDIC Survey reports smartphone ownership rates by household characteristics. Overall, 85.4% of respondents own smartphones, with Black respondents reporting slightly lower rates (81.5%) compared to White respondents (85.4%). Ownership rates decline to 66.4% among those aged 65+, 63.3% for those earning less than $15,000 per year, and 75.6% for residents living outside Metropolitan areas. Smartphone ownership rates are also lower among the unbanked (63.7%) compared to the banked (86.6%). We likely under sample these groups with lower mobile device ownership rates.

Looking at the entire SafeGraph sample, Squire (2019) quantifies the sampling bias in the company's mobility data. He documents that the number of devices from SafeGraph's identified home locations correlates highly at the county level with 2010 U.S. Census numbers in terms of population counts (97%), inferred educational attainment (99%), and inferred household income (99%).[7]

Despite this strong alignment between the Census and SafeGaph at the county level, Thaenraj (2021) identifies around 1,000 Census block groups in the SafeGraph data that register more devices residing there than the number of people living there according to the Census. Squire (2019) also discusses this feature of the SafeGraph panel, and he interprets

[6] Appendix Fig. A.1 presents a time-series of the number of branch visitors each month over the sample period. The number of visitors rises over the sample period, starting from around 900 thousand in January 2018 and ending with 1.85 million in December 2019. The change could reflect a combination of increasing bank visitation and improving visitor coverage over time.

[7] Couture, Dingel, Green, Handbury and Williams (2022) analyze mobile device data from the provider PlaceIQ, and the authors find that it too is broadly representative of the general population based on assigned household attributes and movement patterns.

these outlier Census block groups as most likely representing errors or technical limits in SafeGraph's attribution of devices to home block groups. Less extreme misattributions are also possible, but any misattribution is likely between neighboring block groups with similar demographics because the SafeGraph representation lines up well at the county level.

3.3 Descriptive Statistics

Table 1 reports descriptive statistics of our core sample. The typical branch has 40 unique visitors per month on average, but there is wide dispersion across branches, as the standard deviation of visitors is over twice as high at 94. For each branch, SafeGraph provides both the median distance visitors travel to get there and the median time they spend there. On average, the median distance traveled is 5 miles, but the standard deviation is 16 miles. The median dwell time is 49 minutes on average, but for half the branches in the sample, the median dwell time is 9 minutes or less. Finally, of the 36.5 million total mobile devices recorded in our core sample with information on the type of device, 52% are iOS and 46% are Android.

Table 2 compares demographic characteristics of residents living in the geographic areas covered by our core sample of bank branches with those in the areas covered by the full set of branches in the SOD. Demographic attributes in the table are taken from the 2019 5-year ACS and are averaged at the Census Bureau's zip code tabulation area (ZCTA). In ZCTAs having branches in the SOD, the fraction of White households is 80.5%, which aligns closely with the 79.9% share of White households in ZCTAs having branches in our core sample. The SOD and core sample are also similar according to the percentage of Black households (9.5% in SOD vs. 10.3% in our core sample) and the percentage of Hispanic households (10.6% vs. 10.9%). Median household income in areas covered by our sample is just over $500 (1%) higher on average than median household income in areas covered by the SOD. Urban areas in our core sample are over-represented by about 3% compared to the SOD, which coincides with greater smartphone ownership rates in urban over rural areas. The differences in demographic attributes between the two samples are precisely estimated, but overall, the economic magnitudes of the differences are small relative to the mean values across areas.

4 Econometric Method

Here, we layout the steps of the econometric method we use to estimate the parameters of the fixed effects gravity model of Eq. (1) using the privacy-protected geolocation data. The

approach adapts the Method of Simulated Moments (MSM) to identify high-dimensional fixed effects. A key insight of the approach is to simulate data from the gravity model and then apply the same differential privacy algorithm to the simulated data that the data provider used to privacy-protect the geolocation data. We run the method separately per year-month of our sample period (January 2018 - December 2019).

4.1 Specify the DGP for visitors

The data generating process (DGP) we simulate is the number of visitors from block groups to branches through time. We assume that the true number of visitors from block group i to branch j in year-month t, denoted V^*_{ijt}, is Poisson distributed. Using the gravity model from Eq. (1), we express the true visitor count as obeying

$$V^*_{ijt} \sim \text{Pois}\left(\exp\left(\gamma_{it} + \lambda_{jt} - \beta_t \log \text{Distance}_{ij}\right)\right). \tag{2}$$

We measure distance in miles between branches and the population-weighted center of visitors' home block groups. We use the haversine formula to calculate distance, which accounts for the curvature of the Earth.[8]

To account for the differential privacy algorithm in the simulation, we let L_{ijt} denote the Laplace noise that SafeGraph adds to V^*_{ijt} to protect user privacy. Noise is added only if SafeGraph observes a visitor (i.e., $V^*_{ijt} > 0$). The noise $L_{ijt} \sim \text{Laplace}(0, b)$, where b is the scale of the distribution, and SafeGraph informed us that $b = \frac{10}{9}$. Let V^+_{ijt} denote the number of visitors after the noise is added, giving:

$$V^+_{ijt} = V^*_{ijt} + L_{ijt}. \tag{3}$$

[8] The centers of population are computed using population counts from the 2010 Census and are found here: 2010 Census Centers of Population. The haversine distance between two latitude-longitude coordinates $(lat_1, long_1)$ and $(lat_2, long_2)$ is $2r \arcsin\left(\sqrt{h}\right)$, where r is the Earth's radius and $h = \text{hav}(lat_1 - lat_2) + \cos(lat_1)\cos(lat_2)\,\text{hav}(long_2 - long_2)$. The haversine function $\text{hav}(\theta) = \sin^2\left(\frac{\theta}{2}\right)$. We take the Earth's radius to be 3,956.5 miles, which is midway between the polar minimum of 3,950 miles and the equatorial maximum of 3,963 miles. The haversine formula treats the Earth as a sphere and is less precise than other measures that consider the Earth's ellipticity, such as Vincenty's formula. Yet another alternative that is more representative of actual travel is the road driving time between locations. Even so, the haversine formula is simple, fairly accurate, and convenient to compute. In Online Appendix Table A.1, we regress the driving times between about 1 million random block groups and bank branches onto the corresponding haversine distances. Driving times are computed using the Origin-Destination Cost Matrix of ArcGIS Pro under the default settings. Regressions are run across the entire 1 million sample and over parts of the sample associated with various demographic attributes, such as including only block groups with Black population shares exceeding 80% from the 2019 5-yr. American Community Survey. Across the samples, the regressions produce very high R^2, ranging from 0.972 to 0.993. Haversine distance is computationally easier to calculate, and these regression results suggest that it correlates highly with driving time.

Let $\lfloor V^+_{ijt} \rfloor$ denote the integer floor to which SafeGraph rounds the noisy visitor count. To accommodate SafeGraph's truncation and censoring, we denote z_{ijt} as an indicator for whether a block group × branch visitor count is present in the sample. The selection equation is

$$z_{ijt} = \begin{cases} 1 & \text{if } \lfloor V^+_{ijt} \rfloor \geq 2, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Let V_{ijt} denote the visitor count observed in the sample, subject to SafeGraph's censoring. The observation equation is

$$V_{ijt} = \max \left\{ 4, \lfloor V^+_{ijt} \rfloor \right\}, \tag{5}$$

In the simulation, we implement Eqs. (2) to (5).

4.2 Sample block group × branch pairs

Technically speaking, every possible block group i and branch j pair should enter Eq. (2). But our data of over fifty-thousand branches, over two-hundred-thousand block groups, altogether spanning twenty-four months, makes it computationally impractical to have the billions of possible block group × branch pairs enter the MSM estimation. Instead, we sample pairs using stratified sampling.

In each year-month, block group × branch pairs in the SafeGraph data register either positive (and ≥ 4) or missing observed visitor counts. If a block group × branch pair has a positive visitor count, then we know that residents of the block group visited the branch in the period, and we sample this block group × branch pair in our simulation with probability 1. If a block group × branch pair has a missing visitor count in the year-month, then either residents of the block group did not visit the branch in the period, or the visitor count was left out of the data from SafeGraph's differential privacy methods. In each year-month, we sample from this *alternative* set of missing block group × branch pairs such that (i) every pair in the alternative set has the same probability of being sampled, and (ii) each block group and each branch is part of at least one block group × branch pair in the alternative set. The second condition ensures that each block group and branch is represented in the stratified sampling. We set the sampling probability to $1/2000$, which implies that, on average, the randomly sampled alternative set of block group × branch pairs represents slightly higher than a 0.05% sample size of all possible block group × branch pairs with missing visitor counts.

To establish notation for the stratified sampling of block group × branch pairs, we let n_t denote the stratified sample of block group × branch pairs in year-month t. This set is

the union of the set of pairs with positive observed visitor counts that are sampled with probability 1, denoted n_t^1, and the alternative set of pairs with missing observed visitor counts that are sampled with probability $1/2000$, denoted n_t^0. Let N_{it}^0 denote the *population* of the block group $\times$ branch pairs with missing observed visitor counts that are associated with block group i. (We use lowercase notation for the stratified sample of pairs and uppercase notation for the population of pairs.)

We implement the stratified sampling in the following manner to satisfy conditions (i) and (ii) above. To satisfy (ii), we pick 1 pair randomly from N_{it}^0 for each block group i. Notice that we could have chosen more than one pair per block group i to satisfy (ii), but choosing just one reduces the estimation time. Next, to satisfy (i), we draw a uniform random variable $u_{ijt} \sim U[0,1]$ for each pair in N_{it}^0. We include the pair in the sample if $u_{ijt} \leq \dfrac{p - \frac{m}{|N_{it}^0|}}{1 - \frac{m}{|N_{it}^0|}}$, where $p = 1 - \sqrt{1 - 1/M}$, and $1/M$ is our target sampling probability of $1/2000$, and $|\cdot|$ is cardinality of a set. We loop this procedure through each block group i. We then repeat the process for each branch j (i.e., draw a uniform random variable for each pair again in N_{it}^0, but looping through all branches).

Notice that we rely on the uniform random variable draw falling short of a threshold to determine whether a block group $\times$ branch pair is sampled because the number of pairs in each set N_{it}^0 is discrete, but we want the sampling probability to be the same across all block group $\times$ branch pairs with missing visitor counts. The probability of a pair being sampled when looping through block groups is the union of the initial 1 random pair choice satisfying condition (ii) and the threshold condition on the uniform draw. That probability is the following:

$$\frac{1}{|N_{it}^0|} + \frac{p - \frac{1}{|N_{it}^0|}}{1 - \frac{1}{|N_{it}^0|}} - \frac{1}{|N_{it}^0|}\frac{p - \frac{1}{|N_{it}^0|}}{1 - \frac{1}{|N_{it}^0|}}, \tag{6}$$

which is simply the probability of the union of the two independent events, where we have used the relation $P(A \bigcup B) = P(A) + P(B) - P(A \bigcap B)$ for independent events A and B. Some algebra reveals that Eq. (6) equals p. Because we repeat the process across all branches, the probability of a block group $\times$ branch pair being sampled either from the loop through block groups or the loop through branches is

$$p + p - p^2 = \frac{1}{M},$$

which matches our target sampling probability, as desired.

The stratified sampling requires that we apply probability weights to any variable

measured at the block group × branch level, such as visitor counts or pairwise distances, so as to rebalance the data and make it represent the target population as closely as possible. We assign probability weights equaling 1 to the sampled pairs in the set n_t^1 because these pairs were sampled with probability 1. We assign probability weights denoted ω_t to the sampled pairs in the set n_t^0. These probability weights satisfy:

$$\omega_t|n_t^0|+1|n_t^1| = \text{Total no. of block groups in year-month } t \times \text{Total number of branches in year-month } t. \tag{7}$$

Rearranging Eq. (7) shows that the probability weight ω_t per year-month is the number of *population* pairs with missing observed visitor counts divided by the number of *sampled* pairs with missing observed visitor counts. Following standard practice, we have the probability weights equal the reciprocal of the likelihood of being sampled ($M = 2000$), but they can deviate slightly from M by chance because of the random sampling.

4.3 Initialize the fixed effects routine

In each year-month of the sample period, the MSM uses the visitor data v and the model parameters $\psi \equiv \left\{\beta_t, \gamma_{it}, \lambda_{jt}\right\}$ to minimize the distance between simulated model moments and data moments. With the very large number of block groups and branches in our sample, the model of visitor counts in Eq. (2) requires hundreds of thousands of fixed effects to be estimated. Estimating all these parameters from the MSM minimization problem alone would be computationally impractical. Instead, we adopt an iterative routine to identify the fixed effects $\left\{\gamma_{it}, \lambda_{jt}\right\}$ and let the minimization problem identify β_t. Holding fixed an estimate of β_t and given initial estimates of the fixed effects, the routine updates the fixed effects estimates until they converge. After the fixed effects converge per estimate of β_t in the year-month, the MSM minimization problem then chooses the optimal β_t estimate that satisfies the moment conditions in the year-month. We initialize the fixed effects routine with guessed estimates $\hat{\gamma}_{it}^0 = \hat{\lambda}_{jt}^0 = 1$ for all i and j and t.

4.4 Simulate visitor counts

We run $S = 10$ simulations of the visitor counts per block group × branch pair. The S simulations are run per year-month of the sample. We differentially simulate visitor counts from the two sets of sampled block group × pairs, n_t^0 and n_t^1, because of their different probability weights.

Consider first the set n_t^1 of pairs with positive observed visitor counts that were sampled with probability 1. Per year-month, we begin the simulation by drawing $|n_t^1| \times S$ Laplace random variables having mean zero and scale 10/9, and we draw $|n_t^1| \times S$ independent Uniform random variables over the unit interval. We draw these random variables only once at the beginning of each year-month's run so that the MSM does not have the underlying sample change for every guess of the model parameters. Given an estimate $\hat{\beta}_t$ of the gravity coefficient and the initial guessed estimates $\left\{\hat{\gamma}_{it}^0, \hat{\lambda}_{jt}^0\right\}$ of the fixed effects, we then apply the inverse Poisson CDF to transform the Uniform random variables into Poisson random variables with distinct means given in Eq. (2).

Each Poisson draw is a "true" block group × branch visitor count. To replicate SafeGraph's differential privacy methods in the simulations, we (i) add a Laplace draw to all non-zero true visitor counts to form a "noisy" block group × branch visitor count, (ii) round each noisy visitor count down to the nearest integer, (iii) set to 0 all noisy visitor counts below 2, and (iv) replace all noisy visitor counts that equal 2 or 3 with 4 (see Eqs. (2) to (5)). Simulated visitor counts are 0 if either the true visitor count (from the Poisson draw) is 0 or the noisy visitor count (from the Poisson draw plus the Laplace draw) falls below 2. This way, simulated visitor counts that equal 0 arise in the same two ways as would 0 visitor counts in the observed SafeGraph data. Let $\tilde{v} = \{\tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_S\}$ be the S simulated visitor counts in year-month t, where we have excluded a t subscript to simplify notation.

Consider next the set n_t^0 of block group × branch pairs with missing visitor counts that were sampled with probability $1/2000$. If an extra $|n_t^0| \times S$ pairs of visitor counts were simulated in the same manner described in the previous two paragraphs, those simulated visitor counts would have disproportionate impact on any computed moments because of the high probability weights that would multiply them. Noise from the simulation would be amplified and make the estimation unstable. Rather than simulating visitor counts for the block group × branch pairs in n_t^0, we construct their implied empirical probability distribution according to the parameter estimate of ψ in each iteration. If an infinite number of visitor counts from the pairs in n_t^0 were in fact simulated, their distribution would coincide with this constructed empirical distribution. Notice that we cannot apply this approach to the set n_t^1 of sampled block group × branch pairs because each pair in that set is drawn from a distinct distribution, due, in part, to the block group- and branch-specific fixed effects. For those pairs, we simulate draws. However, the sampled pairs in the set n_t^0 are meant to represent the remaining block group × branch pairs in the population with missing observed visitor counts, which are very high in number. One stratified sampled observation is meant to represent 2,000 observations from the

same distribution. We construct the empirical distribution that these sampled pairs represent.

Because the Laplace noise is added after the Poisson draw, this empirical distribution is a truncated and censored Laplace distribution whose mean is the realization of the Poission draw. With this in mind, let $G(y, k)$ be the CDF of a Laplace distribution with mean k and scale $10/9$. And let $\hat{\mu}_{ijt}$ denote the estimated mean of the Poisson distribution of visitor counts in Eq. (2). Namely,

$$\hat{\mu}_{ijt} \equiv \exp\left(\hat{\gamma}_{it} + \hat{\lambda}_{jt} - \hat{\beta}_t \log \text{Distance}_{ij}\right). \tag{8}$$

Finally, let the probability that the Poisson distribution draws a visitor count of k, given its estimated mean $\hat{\mu}_{ijt}$ be denoted $p\left(k, \hat{\mu}_{ijt}\right)$. Notice that the parameters of the empirical distribution update with every iteration of the estimated fixed effects and guess of β_t.

We construct 7 components of the empirical distribution that we use in the moments of the estimation. Because both the Laplace and Poisson distributions have infinite support, we must insert an upper bound to both supports when constructing the empirical distribution. We bound the Poisson support at $K = 20$ and the Laplace support at $L = 30$. The upper bounds imply that the 7 components of the empirical distribution hold approximately. As $K \to \infty$ and $L \to \infty$, they would hold exactly. The 7 components of the empirical distribution we compute are:

1. **Probability that the visitor count equals 0:**

$$\Pr\left(\tilde{V}_{ijt} = 0 | \hat{\mu}_{ijt}\right) \approx p\left(0, \hat{\mu}_{ijt}\right) + \sum_{k=1}^{K} p\left(k, \hat{\mu}_{ijt}\right) \times G(2, k). \tag{9}$$

   The probability that a simulated visitor count is zero equals the probability that the Poisson draw equals zero, represented by the first term in Eq. (9), plus the cumulative probability that the Poisson draw has a positive value but the Laplace draw reduces that positive value to the lower bound of 0. That cumulative probability is represented by the second term in Eq. (9). In that term, the Laplace draw has mean k to adjust for different possible positive draws of the Poisson. Moreover, the CDF value of the Laplace distribution given that mean, $G(2, k)$, is positioned at 2 because SafeGraph truncates any visitor count below 2. Thus, the second term is the cumulative probability that the simulated visitor count falls below 2 after the Laplace noise is added to a positive Poisson draw. The Laplace probability multiplies the Poisson probability because the two draws are independent. Notice that no Laplace piece enters the first term because SafeGraph adds Laplace noise only to positive observed visitor counts.

2. **Probability that the visitor count exceeds 0:**

$$\Pr\left(\tilde{V}_{ijt} > 0 | \hat{\mu}_{ijt}\right) \approx \sum_{k=1}^{K} p\left(k, \hat{\mu}_{ijt}\right) \times \left(1 - G\left(2, k\right)\right). \tag{10}$$

This probability is simply the complement of the previous one. Because the visitor count exceeds 0 in this scenario, Laplace noise is always added to the Poisson draw, and hence, the "survival function" of the Laplace, given by $1 - G\left(2, k\right)$, multiplies each Poisson probability. The survival value is the probability that the visitor count avoids truncation.

3. **Probability that the visitor count equals 4:**

$$\Pr\left(\tilde{V}_{ijt} = 4 | \hat{\mu}_{ijt}\right) \approx \sum_{k=1}^{K} p\left(k, \hat{\mu}_{ijt}\right) \times \left(G\left(5, k\right) - G\left(2, k\right)\right). \tag{11}$$

The probability that the visitor count equals 4 is the probability that the Poisson draw lands at or above 1 visitor count times the probability that the Lapalace draw pushes the visitor count to a value in the interval between 2 and 4 inclusive (i.e., the censoring region). Because SafeGraph rounds visitor counts down to the nearest integer, the probability that the Laplace draw carries the visitor count into the censored region is $G\left(5, k\right) - G\left(2, k\right)$. For example, a Poisson draw plus a Laplace draw that equaled $4.\overrightarrow{9}$ would round down to 4.

4. **Probability that the visitor count exceeds 4:**

$$\Pr\left(\tilde{V}_{ijt} > 4 | \hat{\mu}_{ijt}\right) \approx \sum_{k=1}^{K} p\left(k, \hat{\mu}_{ijt}\right) \times \left(1 - G\left(5, k\right)\right). \tag{12}$$

This probability is simply the complement of the previous one. The survival function of the Laplace above 4, given by $1 - G\left(5, k\right)$, multiplies each Poisson probability. The survival value is the probability that the visitor count avoids censoring.

5. **Expected visitor count:**

$$\mathbb{E}\left(\tilde{V}_{ijt} | \hat{\mu}_{ijt}\right) \approx \sum_{k=1}^{K} p\left(k, \hat{\mu}_{ijt}\right) \left[4 \times \left\{G\left(5, k\right) - G\left(2, k\right)\right\} + \sum_{l=5}^{L} l \times \left\{G\left(l+1, k\right) - G\left(l, k\right)\right\}\right]. \tag{13}$$

The formula for the mean visitor count is broken up into two parts. Both parts are multiplied by the probability, $p\left(k, \hat{\mu}_{ijt}\right)$, that the Poisson draw lands at or above 1 visitor count so that the observation enters the support of the empirical distribution. The first part is the probability that the Laplace draw pushes the visitor count to a value in

the interval between 2 and 4 inclusive (the censoring region) multiplied by 4 visitors. The second part is the probability that the Laplace draw pushes the visitor count to a value of 5 or higher, multiplied by that value. Because SafeGraph rounds visitor counts down to the nearest integer, the probability of each value in this second part is the CDF of the Laplace distribution at 1 above that value less the CDF at the value, given by $G(l+1,k) - G(l,k)$.

6. **Expected log visitor count, conditional on the visitor count exceeding 0:**

$$\mathbb{E}\left(\log \tilde{V}_{ijt} | \tilde{V}_{ijt} > 0, \hat{\mu}_{ijt}\right) \approx \frac{\sum_{k=1}^{K} p\left(k, \hat{\mu}_{ijt}\right)\left[\log 4 \times \{G(5,k) - G(2,k)\} + \sum_{l=5}^{L} \{\log l \times (G(l+1,k) - G(l,k))\}\right]}{\Pr\left(\tilde{V}_{ijt} > 0 | \hat{\mu}_{ijt}\right)}. \tag{14}$$

The formula for the mean of the natural logarithm of the visitor count is very similar to that of the mean of the visitor count from Eq. (13). The only adjustments are that the natural logarithm is taken as needed and that the mean is re-weighted to account for the positive visitor count requirement. That re-weighting is exhibited via the division by $\Pr\left(\tilde{V}_{ijt} > 0 | \hat{\mu}_{ijt}\right)$, defined in Eq. (10), which is the way to compute the mean of a truncated random variable.

7. **Expected log visitor count, conditional on the visitor count exceeding 4:**

$$\mathbb{E}\left(\log \tilde{V}_{ijt} | \tilde{V}_{ijt} > 4, \hat{\mu}_{ijt}\right) \approx \frac{\sum_{k=1}^{K} p\left(k, \hat{\mu}_{ijt}\right)\left[\sum_{l=5}^{L} \{\log l \times (G(l+1,k) - G(l,k))\}\right]}{\Pr\left(\tilde{V}_{ijt} > 4 | \hat{\mu}_{ijt}\right)}. \tag{15}$$

This conditional mean is even simpler to compute than the one in Eq. (14). The formula consists of just the second component in the numerator of Eq. (14), and the re-weighting in the denominator is the probability of the visitor count exceeding 4, given in Eq. (12).

4.5 Iterate the fixed effects until convergence

Under a fixed estimate $\hat{\beta}_t$, the next step is to iterate the estimated fixed effects until they converge. Because the fixed effects are measured at the block group or branch level, and not the block group × branch level like the visitor counts, we need two other sets of probability weights for the fixed effects estimation due to the stratified sampling. The block group and branch weights are defined similarly as the block group × branch weights in Eq. (7), but they are measured from the perspective of a block group or branch.

Notice that the stratified sample of block group × branch pairs also creates a stratified sample of block groups and branches *separately*. With this in mind, we let b_{it} denote the

stratified sample of branches for block group i in year-month t. This set is the union of the set of branches from the pairs sampled with probability 1, denoted b_{it}^1, and the set of branches from the pairs sampled with probability $1/2000$, denoted b_{it}^0. Likewise, let h_{jt} denote the stratified sample of home block groups for branch j in year-month t. This set is the union of the set of block groups from the pairs sampled with probability 1, denoted h_{jt}^1, and the set of block groups from the pairs sampled with probability $1/2000$, denoted h_{jt}^0. The block groups in h_{jt}^1 and branches in b_{it}^1 have probability weights equal to 1. The block groups in h_{jt}^0 have probability weights denoted ω_t^j, and the branches in b_{it}^0 have probability weights denoted ω_t^i. These probability weights are defined as:

$$\omega_t^j |h_{jt}^0| + 1|h_{jt}^1| = \text{Total no. of block groups in year-month } t, \qquad \forall\,(i,j) \in n_t, \tag{16}$$

$$\omega_t^i |b_{it}^0| + 1|b_{it}^1| = \text{Total no. of branches in year-month } t, \qquad \forall\,(i,j) \in n_t. \tag{17}$$

We use the block group- and branch-specific probability weights from Eqs. (16) to (17) only in the fixed effects iteration routine. We iterate the estimated fixed effects sequentially. We begin with the estimated branch fixed effects $\left\{\hat{\lambda}_{jt}\right\}$, while holding constant the estimated block group fixed effects $\{\hat{\gamma}_{it}\}$ at $\hat{\gamma}_{it}^0 = 1$, $\forall i$ and $\forall t$.

To estimate the branch fixed effects, we take advantage of another data field in SafeGraph: a branch's total number of visitors. The SafeGraph name for this field is `RAW_VISITOR_COUNT`. Unlike the number of visitors from a block group to the branch, a branch's total number of visitors is unaffected by SafeGraph's differential privacy methods. Because we presume that block group residents can visit any branch cross-country in the year-month, we can take advantage of a branch's total visitors to uniquely pin down the estimate of the branch's fixed effect. Let V_{jt}^T denote branch j's total visitors in year-month t.

The iteration process for estimating the branch fixed effects is as follows. Suppose we are on the k-th iteration. From Eq. (2), the expected number of visitors to branch j from block group i in year-month t based on the k-th iteration estimates of the fixed effects is

$$\hat{V}_{ijt}^k = \exp\left(\hat{\lambda}_{jt}^k\right) \exp\left(\hat{\gamma}_{it}^k\right) d_{ij}^{-\hat{\beta}_t}. \tag{18}$$

Summing across block groups, and adjusting for the probability weights defined in Eq. (16), we obtain a branch's expected total visitor count:

$$\hat{V}_{jt}^k = \exp\left(\hat{\lambda}_{jt}^k\right) \left(\sum_{i \in h_{jt}^1} \exp\left(\hat{\gamma}_{it}^k\right) d_{ij}^{-\hat{\beta}_t} + \sum_{i \in h_{jt}^0} \omega_t^j \exp\left(\hat{\gamma}_{it}^k\right) d_{ij}^{-\hat{\beta}_t} \right) \tag{19}$$

Given $\hat{\beta}_t$ and the k-th iteration of the estimated block group fixed effects, $\left\{\hat{\gamma}_{it}^k\right\}$, we determine the k-th iteration of each branch's estimated fixed effect, $\hat{\lambda}_{jt}^k$, by solving for the value that equates the branch's expected total visitor count, $\hat{V}_{jt}^k$ from Eq. (19), with the branch's observed total visitor count, V_{jt}^T. Mathematically speaking, the branch's fixed effect estimate satisfies:

$$\hat{\lambda}_{jt}^k = \log V_{jt}^T - \log\left(\sum_{i \in h_{jt}^1} \exp\left(\hat{\gamma}_{it}^k\right) d_{ij}^{-\hat{\beta}_t} + \sum_{i \in h_{jt}^0} \omega_t^j \exp\left(\hat{\gamma}_{it}^k\right) d_{ij}^{-\hat{\beta}}\right). \tag{20}$$

Per iteration, Eq. (20) pins down each branch's estimated fixed effect as a function of the estimated block group fixed effects (and the estimate of β_t). The estimated block group fixed effects will iterate until they converge, and by Eq. (20), once the estimated block group fixed effects converge, so too do the estimated branch fixed effects, given an estimate of β_t.

The iteration process for estimating the block group fixed effects is as follows. Suppose we are on the k-th iteration. For each block group i in the year-month, we divide the average observed visitor counts V_{ijt} across the branches in set b_{it}, by the average simulated visitor counts across all branches in set b_{it} and all simulations S. With this in mind, we let the average observed visitor count of block group i be

$$\overline{V}_{it} = \frac{1}{|b_{it}|} \sum_{j \in b_{it}} V_{ijt}. \tag{21}$$

Let the simulated visitor counts from simulation s in iteration k be denoted $\tilde{V}_{ijt}^k(s)$. The average simulated visitor count of block group i in simulation s is

$$\overline{\tilde{V}}_{it}^k(s) = \frac{\sum_{j \in b_{it}^1} \tilde{V}_{ijt}^k(s) + \sum_{j \in b_{it}^0} \omega_t^i \mathbb{E}\left(\tilde{V}_{ijt}^k | \hat{\mu}_{ijt}\right)}{\sum_{j \in b_{it}^1} 1 + \sum_{j \in b_{it}^0} \omega_t^i}, \tag{22}$$

where $\mathbb{E}\left(\tilde{V}_{ijt}^k | \hat{\mu}_{ijt}\right)$ is provided in Eq. (13). Because the calculation is at the block-group level, the probability weights we use are from the block-group perspective, and they either equal 1 or satisfy Eq. (17). Averaging across simulations delivers the mean simulated visitor count of block group i as

$$\overline{\tilde{V}}_{it}^k = \frac{1}{S} \sum_s \overline{\tilde{V}}_{it}^k(s). \tag{23}$$

The ratio of block group i's average observed visitor count to average simulated visitor count is thus:

$$\chi_{it}^k = \frac{\overline{V}_{it}}{\overline{\tilde{V}}_{it}^k} \tag{24}$$

We take ratios of averages rather than differences of averages because the fixed effects in the visitor count model in Eq. (2) are exponentiated. These block group-level ratios then multiplicatively update each block group's estimated fixed effect:

$$\hat{\gamma}_{it}^{k+1} = \hat{\gamma}_{it}^{k} \times \left(\chi_{it}^{k}\right)^{g}, \tag{25}$$

where g is a modifying term to avoid oscillating estimates, and we set its value to 0.5. Notice that if block group i's average simulated visitor count is higher than its average observed visitor count in the data, then $\chi_{it}^{k} < 1$, and the block group's estimated fixed effect is revised downward.

After each update of the estimated block group fixed effects, we re-transform the $|n_t^1| \times S$ Uniform random variables into Poisson random variables using (i) the estimate $\hat{\beta}_t$; (ii) the updated block group fixed effect estimates, $\left\{\hat{\gamma}_{it}^{k+1}\right\}$; and (iii) the updated branch fixed effect estimates, $\left\{\hat{\lambda}_{jt}^{k+1}\right\}$, based on Eq. (20). We then apply differential privacy methods to the "updated" simulated data. The process iterates until the estimated block group fixed effects converge.[9]

While the estimated fixed effects are updated using *ratios* of the averages between observed and simulated values, we found that the estimates converged faster under a convergence criterion that uses *differences* in the averages instead. We define convergence as the squared change between iterations in the mean squared difference between average observed and simulated visitor counts of a block group being sufficiently small. The criterion is similar in spirit to a GMM minimization problem in which the moments are the difference in means between the observed and simulated visitor counts of each block group i, using an identity weighting matrix. Minimization is reached when the change in the GMM objective function becomes sufficiently small. In the calculation of the average squared difference, we assign more weight to block groups with branch goers to more branches (higher $|b_{it}|$). Mathematically, the convergence condition is

$$\left[\frac{1}{|n_t|}\sum_i |b_{it}|\left(\overline{\tilde{V}}_{it}^{k+1} - \overline{V}_{it}\right)^2 - \frac{1}{|n_t|}\sum_i |b_{it}|\left(\overline{\tilde{V}}_{it}^{k} - \overline{V}_{it}\right)^2\right]^2 < \varepsilon \tag{26}$$

for small ε, which we set to $1e^{-9}$.

After the condition in Eq. (26) is met, we have converged fixed effects estimates, denoted $\left\{\hat{\gamma}_{it}^{\infty}\right\}$ and $\left\{\hat{\lambda}_{jt}^{\infty}\right\}$, for a given estimated $\hat{\beta}_t$. The final piece of the estimation is to select the optimal $\hat{\beta}_t$ that minimizes the distance between simulated and data moments in the year-month.

[9]The iterative process we use to identify the fixed effects is similar in spirit to the "zig-zag" algorithm, or Gauss-Seidel method, that is commonly used to identify high-dimensional fixed effects in linear models (Guimaraes and Portugal 2010).

4.6 Select the moments

To identify β_t, we choose 6 unconditional moments of the distribution of visitor counts. We select moments that describe important parts of the distribution. The moments are computed per year-month across all block groups and branches. Denote the vector of the data moments in the year-month as $m(v)$, and denote as $m(\tilde{v}_s|\psi)$ the analogous vector of simulated moments from simulation s.

Recall that n_t is the set of stratified sampled block group × branch pairs in year-month t. The set is the union of the set of pairs in n_t^1 that were sampled with probability 1 and the set of pairs in n_t^0 that were sampled with probability $1/2000$. Recall also that ω_t are the probability weights assigned to the pairs in the set n_t^0, given in Eq. (7). Both the data and simulated moments only include block group × branch pairs from the stratified sample. The 6 data and simulated moments are:

1. **Percent of visitor counts equal to 0:**

$$m_1(v) \equiv \frac{\sum_{(i,j)\in n_t^1} \mathbb{I}\left(V_{ijt}=0\right) + \sum_{(i,j)\in n_t^0} \mathbb{I}\left(V_{ijt}=0\right)\omega_t}{\sum_{(i,j)\in n_t^1} 1 + \sum_{(i,j)\in n_t^0} \omega_t}, \tag{27}$$

$$m_1(\tilde{v}|\psi) \equiv \frac{\sum_{(i,j)\in n_t^1} \mathbb{I}\left(\tilde{V}_{ijt}=0\right) + \sum_{(i,j)_t^0} \Pr\left(\tilde{V}_{ijt}=0|\hat{\mu}_{ijt}\right)\omega_t}{\sum_{(i,j)\in n_t^1} 1 + \sum_{(i,j)\in n_t^0} \omega_t}, \tag{28}$$

   where $\mathbb{I}(\cdot)$ stands for the indicator function and $\Pr\left(\tilde{V}_{ijt}=0|\hat{\mu}_{ijt}\right)$ is from Eq. (9). The data moment $m_1(v)$ is straightforward, separating pairs in the two sampled sets, n_t^0 and n_t^1, and applying the different probability weights. The simulated moment $m_1(\tilde{v}|\psi)$ adds the fraction of the simulated visitor counts from the sampled set n_t^1 equaling 0 to the probability of the visitor counts from the sampled set n_t^0 equaling 0, adjusted by the probability weights.

2. **Percent of visitor counts equal to 4:**

$$m_2(v) \equiv \frac{\sum_{(i,j)\in n_t^1} \mathbb{I}\left(V_{ijt}=4\right) + \sum_{(i,j)\in n_t^0} \mathbb{I}\left(V_{ijt}=4\right)\omega_t}{\sum_{(i,j)\in n_t^1} 1 + \sum_{(i,j)\in n_t^0} \omega_t}, \tag{29}$$

$$m_2(\tilde{v}|\psi) \equiv \frac{\sum_{(i,j)\in n_t^1} \mathbb{I}\left(\tilde{V}_{ijt}=4\right) + \sum_{(i,j)\in n_t^0} \Pr\left(\tilde{V}_{ijt}=4|\hat{\mu}_{ijt}\right)\omega_t}{\sum_{(i,j)\in n_t^1} 1 + \sum_{(i,j)\in n_t^0} \omega_t}, \tag{30}$$

   where $\Pr\left(\tilde{V}_{ijt}=4|\hat{\mu}_{ijt}\right)$ is from Eq. (11).

3. **Average log distance, in cases where $V_{ijt}, \tilde{V}_{ijt} = 0$:**

$$m_3(v) \equiv \frac{\sum_{(i,j)\in n_t^1} \mathbb{I}\left(V_{ijt}=0\right)\log d_{ij} + \sum_{(i,j)\in n_t^0} \mathbb{I}\left(V_{ijt}=0\right)\omega_t \log d_{ij}}{\sum_{(i,j)\in n_t^1} 1 + \sum_{(i,j)\in n_t^0}\omega_t}, \tag{31}$$

$$m_3(\tilde{v}|\psi) \equiv \frac{\sum_{(i,j)\in n_t^1} \mathbb{I}\left(\tilde{V}_{ijt}=0\right)\log d_{ij} + \sum_{(i,j)\in n_t^0} \Pr\left(\tilde{V}_{ijt}=0|\hat{\mu}_{ijt}\right)\omega_t \log d_{ij}}{\sum_{(i,j)\in n_t^1} 1 + \sum_{(i,j)\in n_t^0}\omega_t}. \tag{32}$$

4. **Average log distance, in cases where $V_{ijt}, \tilde{V}_{ijt} = 4$:**

$$m_4(v) \equiv \frac{\sum_{(i,j)\in n_t^1} \mathbb{I}\left(V_{ijt}=4\right)\log d_{ij} + \sum_{(i,j)\in n_t^0} \mathbb{I}\left(V_{ijt}=4\right)\omega_t \log d_{ij}}{\sum_{(i,j)\in n_t^1} 1 + \sum_{(i,j)\in n_t^0}\omega_t}, \tag{33}$$

$$m_4(\tilde{v}|\psi) \equiv \frac{\sum_{(i,j)\in n_t^1} \mathbb{I}\left(\tilde{V}_{ijt}=4\right)\log d_{ij} + \sum_{(i,j)\in n_t^0} \Pr\left(\tilde{V}_{ijt}=4|\hat{\mu}_{ijt}\right)\omega_t \log d_{ij}}{\sum_{(i,j)\in n_t^1} 1 + \sum_{(i,j)\in n_t^0}\omega_t}. \tag{34}$$

5. **OLS coefficient from regressing log visitor counts onto their associated log distances, in cases where $V_{ijt}, \tilde{V}_{ijt} > 0$:**

   First, using the observed data, we define the regression's dependent and independent variables, respectively, as

$$y_{ijt} = \left\langle \log V_{ijt} \right\rangle_{(i,j)\in n_t^1}, \tag{35}$$

$$X_{ijt} = \left[\ \langle 1 \rangle_{(i,j)\in n_t^1}, \ \left\langle \log d_{ij} \right\rangle_{(i,j)\in n_t^1} \ \right]. \tag{36}$$

   Here, $\langle \cdot \rangle_{(i,j)\in n_t^1}$ denotes a vector with length equaling the number of elements in the set n_t^1. The dependent variable y_{ijt} consists of a vector of log visitor counts, whereas the independent variables are a vector of ones and a vector of log distances. With these variables established, the data moment is

$$m_5(v) \equiv \text{Second element of } \left(X_{ijt}'X_{ijt}\right)^{-1}\left(X_{ijt}'y_{ijt}\right) \tag{37}$$

   Notice that, because the data moment reflects only positive observed visitor counts from the set n_t^1 of sampled block group × branch pairs, the probability weights all equal 1 and do not appear in the data moment.

   The corresponding simulated moment uses a weighted least squares (WLS) coefficient because the probability weights do not all equal 1. With this in mind, we define

the observation weights of the WLS as

$$\tilde{\eta}_{ijt} \equiv \begin{bmatrix} \langle 1 \rangle_{(i,j) \in n_t^1 : \tilde{V}_{ijt} > 0} \\ \left\langle \omega_t \Pr\left(\tilde{V}_{ijt} > 0 | \hat{\mu}_{ijt}\right) \right\rangle_{(i,j) \in n_t^0} \end{bmatrix}, \tag{38}$$

where $\Pr\left(\tilde{V}_{ijt} > 0 | \hat{\mu}_{ijt}\right)$ is from Eq. (10). The observation weights consist of (1) a vector of ones with length equaling the number of block group × branch pairs in n_t^1 *that also* have positive simulated visitor counts, and (2) a vector of weighted probabilities that the simulated visitor counts from the pairs in the sampled set n_t^0 exceed 0.

The dependent variable in the WLS is defined as

$$\tilde{y}_{ijt} \equiv \sqrt{\tilde{\eta}_{ijt}} \odot \begin{bmatrix} \left\langle \log \tilde{V}_{ijt} \right\rangle_{(i,j) \in n_t^1 : \tilde{V}_{ijt} > 0} \\ \left\langle \mathbb{E}\left(\log \tilde{V}_{ijt} | \tilde{V}_{ijt} > 0, \hat{\mu}_{ijt}\right) \right\rangle_{(i,j) \in n_t^0} \end{bmatrix}, \tag{39}$$

where $\odot$ is the element-wise product and $\mathbb{E}\left(\log \tilde{V}_{ijt} | \tilde{V}_{ijt} > 0, \hat{\mu}_{ijt}\right)$ is from Eq. (14). The dependent variable consists of (1) a weighted vector of log simulated visitor counts with length equaling the number of block group × branch pairs in n_t^1 *that also* have positive simulated visitor counts, and (2) a weighted vector of mean log simulated visitor counts from the pairs in the sampled set n_t^0, conditional on the simulated visitor counts exceeding 0.

The independent variable in the WLS is defined as

$$\tilde{X}_{ijt} \equiv \begin{bmatrix} \sqrt{\tilde{\eta}_{ijt}}, & \sqrt{\tilde{\eta}_{ijt}} \odot \begin{pmatrix} \left\langle \log d_{ij} \right\rangle_{(i,j) \in n_t^1 : \tilde{V}_{ijt} > 0} \\ \left\langle \log d_{ij} \right\rangle_{(i,j) \in n_t^0} \end{pmatrix} \end{bmatrix}. \tag{40}$$

The independent variable consists of (1) the square root of the weights from Eq. (38), and (2) the element-wise product of the square root of the weights and log distances.

With these terms established, we set the simulated moment as

$$m_5(\tilde{v}|\psi) \equiv \text{Second element of } \left(\tilde{X}'_{ijt}\tilde{X}_{ijt}\right)^{-1}\left(\tilde{X}'_{ijt}\tilde{y}_{ijt}\right). \tag{41}$$

6. **OLS coefficient from regressing log visitor counts onto their associated log distances, where $V_{ijt}, \tilde{V}_{ijt} > 4$:**

The sixth data moment is similar to the fifth data moment, except that it conditions on the visitor count exceeding 4 rather than 0. Specifically, let

$$q_{ijt} = \left\langle \log V_{ijt} \right\rangle_{(i,j) \in n_t^1 : V_{ijt} > 4} \tag{42}$$

$$Z_{ijt} = \left[\ \langle 1 \rangle_{(i,j) \in n_t^1 : V_{ijt} > 4}, \quad \left\langle \log d_{ij} \right\rangle_{(i,j) \in n_t^1 : V_{ijt} > 4} \ \right]. \tag{43}$$

The data moment is then

$$m_6(v) \equiv \text{Second element of } \left(Z_{ijt}' Z_{ijt} \right)^{-1} \left(Z_{ijt}' q_{ijt} \right). \tag{44}$$

The sixth simulated moment is also similar to the fifth simulated moment, just now conditioning on $\tilde{V}_{ijt} > 4$. Thus, let the WLS observation weights be

$$\tilde{\xi}_{ijt} \equiv \left[\begin{array}{c} \langle 1 \rangle_{(i,j) \in n_t^1 : \tilde{V}_{ijt} > 4} \\ \left\langle \omega_t \Pr\left(\tilde{V}_{ijt} > 4 | \hat{\mu}_{ijt} \right) \right\rangle_{(i,j) \in n_t^0} \end{array} \right], \tag{45}$$

where $\Pr\left(\tilde{V}_{ijt} > 4 | \hat{\mu}_{ijt} \right)$ is from Eq. (12). The dependent variable in the WLS is defined as

$$\tilde{q}_{ijt} \equiv \sqrt{\tilde{\xi}_{ijt}} \odot \left[\begin{array}{c} \left\langle \log \tilde{V}_{ijt} \right\rangle_{(i,j) \in n_t^1 : \tilde{V}_{ijt} > 4} \\ \left\langle \mathbb{E}\left(\log \tilde{V}_{ijt} | \tilde{V}_{ijt} > 4, \hat{\mu}_{ijt} \right) \right\rangle_{(i,j) \in n_t^0} \end{array} \right], \tag{46}$$

where $\mathbb{E}\left(\log \tilde{V}_{ijt} | \tilde{V}_{ijt} > 4, \hat{\mu}_{ijt} \right)$ is from Eq. (15). Likewise, the independent variable in the WLS is defined as

$$\tilde{Z}_{ijt} \equiv \left[\ \sqrt{\tilde{\xi}_{ijt}}, \quad \sqrt{\tilde{\xi}_{ijt}} \odot \left(\begin{array}{c} \left\langle \log d_{ij} \right\rangle_{(i,j) \in n_t^1 : \tilde{V}_{ijt} > 4} \\ \left\langle \log d_{ij} \right\rangle_{(i,j) \in n_t^0} \end{array} \right) \right]. \tag{47}$$

With these terms established, we set the simulated moment as

$$m_6(\tilde{v} | \psi) \equiv \text{Second element of } \left(\tilde{Z}_{ijt}' \tilde{Z}_{ijt} \right)^{-1} \left(\tilde{Z}_{ijt}' \tilde{q}_{ijt} \right). \tag{48}$$

In the procedure, we take the mean of the simulated moments by averaging values across the S simulations. Let $\hat{m}(\tilde{v} | \psi)$ be the estimate of the model moments from the S simulations:

$$\overline{m}(\tilde{v} | \psi) = \frac{1}{S} \sum_S m(\tilde{v}_s | \psi). \tag{49}$$

The final step of the MSM procedure is to find the estimated $\hat{\beta}_t$ that minimizes the distance between the data moments and simulated model moments.

4.7 Construct the MSM estimator

The MSM estimator $\hat{\beta}_{t,\text{MSM}}$ minimizes the weighted sum of squared errors between the simulated model moments and data moments. So that all errors are expressed in the same units and the minimization problem is scaled properly, we compute the error $e_r(\tilde{v}, v|\psi)$ per moment r, which is the percent difference between a data moment and its corresponding model moment:

$$e_r(\tilde{v}, v|\psi) \equiv \frac{\overline{m}_r(\tilde{v}|\psi) - m_r(v)}{m_r(v)}, \quad \forall r. \tag{50}$$

Let $e(\tilde{v}, v|\psi)$ denote the vector of moment errors. The MSM estimator is then

$$\hat{\beta}_{t,\text{MSM}} = \underset{\beta_t}{\text{argmin}} \; e(\tilde{v}, v|\psi)' \, W e(\tilde{v}, v|\psi), \tag{51}$$

where W is a 6×6 weighting matrix that controls how each moment is weighted in the minimization problem. Notice that each candidate β_t in Eq. (51) is associated with a different set of converged fixed effects estimates $\left\{\hat{\gamma}_{it}^{\infty}, \hat{\lambda}_{jt}^{\infty}\right\}$.

We use the identity matrix I for the weighting matrix W. We also implemented a two-step procedure to select an optimal weighting matrix W, but that approach produced unstable estimates. This is not surprising, given evidence in the literature of the underperformance of the two-step procedure when there is uncertainty in the estimation of the weighting matrix (Arellano and Bond 1991; Hwang and Sun 2018).

Under this identity weighting matrix, one can derive the variance-covariance matrix of the MSM estimator $\hat{\beta}_{t,\text{MSM}}$ as

$$\widehat{\text{Var}}\left(\hat{\beta}_{t,\text{MSM}}\right) = \left(1 + \frac{1}{S}\right)\left[\frac{\partial \overline{m}(\tilde{v}|\psi)}{\partial \beta_t}' \frac{\partial \overline{m}(\tilde{v}|\psi)}{\partial \beta_t}\right]^{-1}, \tag{52}$$

where $\frac{\partial \overline{m}(\tilde{v}|\psi)}{\partial \beta_t}$ is the derivative of the vector of simulated moments, evaluated at $\hat{\beta}_{t,\text{MSM}}$. We calculate the derivatives numerically by taking a central difference around $\hat{\beta}_{t,\text{MSM}}$.

5 Estimation Results

Fig. 2 compares the distribution of observed "raw" visitor counts (in black) to simulated "true" visitor counts (in blue) according to the month-by-month MSM estimation of the Poisson model in Eq. (2). The simulated visitor counts include all positive draws from all simulations across every year-month in the sample period. The numbers of 0 visitor counts in both distributions are very large and are omitted for clarity. The black distribution reveals

the effects of the differential privacy on the raw visitor counts, having a large mass at 4. The MSM does a reasonable job spreading out the mass of visitors into the lower portion of the distribution that is lost in the observed data. The two distributions line up fairly well at the right tail, where we have censored the visitor counts at 10 in the figure for clarity. The "true" visitor distribution in blue obeys our assumed Poisson structure, which may not coincide with the true data generating process of visitor counts known only to SafeGraph. Nevertheless, as with standard MSM, potential misspecification of the simulating distribution does not interfere with the consistency of the estimates (McFadden 1989). Also displayed in the figure is the distribution of simulated "manipulated" visitor counts in red, which is the distribution of the "true" visitor counts after they are manipulated by the differential privacy methods in Eqs. (3) to (5).

Fig. 3 compares the observed number of visitors from each Census block group to their expected (i.e., predicted) counterparts from the simulation. It presents a binned scatter plot of the log observed number of branch goers from each block group versus the log expected number of branch goers from the block group based on the MSM estimates. If SafeGraph applied no differential privacy methods to their geolocation data, all dots in the figure would line up neatly on the red 45° line. The single caveat is that the expected number of visitors might not be whole numbers, whereas the observed number of visitors must be. The censoring levels off the log observed visitor counts at 1.4, which corresponds to 4 visitors. The truncation causes the observed visitor counts to enter below the expected visitor counts, and the gap between observed and expected counts is largest for block groups with few branch goers, which are areas where the truncation has the largest impact. The gap shrinks as the number of branch goers from a block group increases. In block groups with many branch goers, the observed and expected number of visitors nearly match. This implies that the MSM generates estimates that fit the geolocation data well in regions least affected by the differential privacy distortions, which one would hope for.

Fig. 4, Panel A presents the gravity coefficient estimates through time, along with 95% confidence intervals. The monthly point estimates of the gravity coefficient range from about -1.45 to -1.26, and they are fairly stable month-to-month. Thus, across the country, if a representative branch were located 1% farther away from a representative block group, the number of residents from that block group who travel to that branch would drop by around 1.26-1.45% per month. As for comparison, Agarwal et al. (2018) estimate a gravity model of consumer expenditures in nonfinancial sectors. They find a gravity coefficient of -1.05 for the average out-of-home purchase, but they document significant heterogeneity across sectors,

with Food Stores, for example, observing an estimate of -0.85; and Health Services, -0.33.

Fig. 4, Panels B and C present histograms of the estimated Census block group and bank branch fixed effects across all months of the sample period. A block group's fixed effect can be interpreted as the average log number of residents from that block group who visit any branch in the year-month, controlling for branch fixed effects and transportation costs. The bulk of the distribution of block group fixed effects range from exponentiated values around 0.01 to 20. Similarly, a branch's fixed effect can be interpreted as the branch's average log number of visitors in the year-month, controlling for visitors' block group fixed effects and their transportation costs. Most of the mass is within a range of exponentated values between 0.01 and 30. In an unreported regression, roughly 77% of the variation in a branch's fixed effect over time can be explained by the branch itself, suggesting that branch quality is fairly stable over time.

SafeGraph's differential privacy methods bias traditional methods of estimating the gravity model and prompts an alternative econometric method like the MSM. But computing estimates from the traditional methods is still useful to informally assess the magnitude of the bias. To this end, Online Table 3 presents gravity coefficient estimates and standard errors from the MSM estimation, along with estimates and standard errors from OLS and PPML estimations. The PPML and OLS estimates are computed on the observed, "raw" visitor counts. We run each estimation approach per year-month of the sample. PPML and OLS estimations are also run over the full sample panel period (January 2018 - December 2019). In addition, we run the OLS estimation on block group $\times$ branch pairs with more than 4 visitors (which avoid SafeGraph's censoring). PPML and OLS standard errors are two-way clustered by both Census block groups and bank branches.

The gravity coefficient estimates from the MSM range from -1.45 to -1.26. The estimates from OLS range from -0.062 to -0.038, roughly twenty to thirty times smaller in magnitude. The OLS estimate over the full panel is -0.053, still an order of magnitude below the MSM estimates. When the sample is limited to block group $\times$ branch pairs with greater than 4 visitors, the OLS estimates rise in magnitude, ranging from -0.33 to -0.27, which is still roughly four to five times smaller in magnitude than the MSM estimates. Computed over all block group $\times$ branch pairs, the PPML estimates register higher magnitudes than the OLS ones, ranging in values from -0.108 to -0.066. But they still are roughly ten to twenty times smaller in magnitude than the MSM estimates. The PPML gravity coefficient estimate over the full panel is -0.091.

Overall, Online Table 3 reveals the downward bias that SafeGraph's differential privacy

methods introduce to traditional methods of estimating the gravity equation, and it stresses the need for the alternative MSM procedure.

6 Conclusion

We propose an econometric method to estimate fixed-effects gravity models on data protected by differential privacy. The method adapts the Method of Simulated Moments to identify high-dimensional fixed effects. We illustrate the method by estimating a gravity model of consumer flows to bank branches using privacy-protected geolocation data from mobile devices. We find significant differences between the estimates obtained from the proposed method and those obtained from standard gravity model estimation methods. We hope the method can be useful in a range of other applications relying on big data that are affected by differential privacy methods.

References

Adda, J. and R. Cooper (2003) *Dynamic Economics: Quantitative Methods and Applications*: MIT Press.

Agarwal, A. and R. Singh (2023) "Causal Inference with Corrupted Data: Measurement Error, Missing Values, Discretization, and Differential Privacy," April, Working paper.

Agarwal, S., J. B. Jensen, and F. Monte (2018) "The geography of consumption," May, Working paper. Georgetown University, Washington, D.C.

Ahlfeldt, G. M., S. J. Redding, D. M. Sturm, and N. Wolf (2015) "The economics of density: Evidence from the Berlin Wall," *Econometrica*, 83 (6), 2127–2189.

Allen, J., C. Bavitz, M. Crosas et al. (2020) "The OpenDP White Paper," May, https://projects.iq.harvard.edu/files/opendifferentialprivacy/files/opendp_white_paper_11may2020.pdf, Working Paper.

Arellano, M. and S. Bond (1991) "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations," *Review of Economic Studies*, 58 (2), 277–297.

Barrientos, A. F., C. Bowen, J. Snoke, and A. Williams (2023) "Disclosing Economists' Privacy Perspectives: A Survey of American Economic Association Members on Differential Privacy and Data Fitness for Use Standards," April, Working paper.

Bricongne, J.-C., L. Fontagné, G. Gaulier, D. Taglioni, and V. Vicard (2012) "Firms and the global crisis: French exports in the turmoil," *Journal of international Economics*, 87 (1), 134–146.

Couture, V., J. I. Dingel, A. Green, J. Handbury, and K. R. Williams (2022) "JUE Insight: Measuring movement and social contact with smartphone data: a real-time application to COVID-19," *Journal of Urban Economics*, 127, 1–9, https://doi.org/10.1016/j.jue.2021.103328.

Davis, D. R., J. I. Dingel, J. Monras, and E. Morales (2019) "How segregated is urban consumption?" *Journal of Political Economy*, 127 (4), 1684–1738.

Dwork, C. (2006) "Differential privacy," in *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33*, 1–12, Springer.

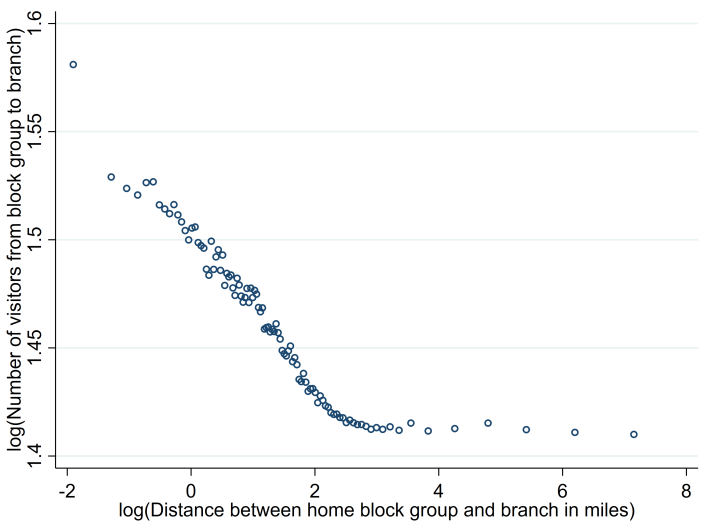
Eaton, J. and S. Kortum (2002) "Technology, geography, and trade," *Econometrica*, 70 (5), 1741–1779.

Eaton, J. and A. Tamura (1994) "Bilateralism and regionalism in Japanese and U.S. trade and direct foreign investment patterns," *Journal of the Japanese and International Economies*, 8 (4), 478–510.

Guimaraes, P. and P. Portugal (2010) "A simple feasible procedure to fit models with high-dimensional fixed effects," *The Stata Journal*, 10 (4), 628–649.

Harrigan, J. (1996) "Openness to trade in manufactures in the OECD," *Journal of International Economics*, 40 (1-2), 23–39.
Head, K. and T. Mayer (2014) "Gravity equations: Workhorse, toolkit, and cookbook," in *Handbook of International Economics*, 4, 131–195.
Helpman, E., M. Melitz, and Y. Rubinstein (2008) "Estimating trade flows: Trading partners and trading volumes," *Quarterly Journal of Economics*, 123 (2), 441–487.
Hwang, J. and Y. Sun (2018) "Should we go one step further? An accurate comparison of one-step and two-step procedures in a generalized method of moments framework," *Journal of Econometrics*, 207 (2), 381–405.
Karger, E. and A. Rajan (2020) "Heterogeneity in the marginal propensity to consume: Evidence from Covid-19 stimulus payments," Working paper. Federal Reserve Bank of Chicago, Chicago, IL.
Larch, M., J. Wanner, Y. V. Yotov, and T. Zylkin (2019) "Currency unions and trade: A PPML reassessment with high-dimensional fixed effects," *Oxford Bulletin of Economics and Statistics*, 81 (3), 487–510.
McFadden, D. (1989) "A method of simulated moments for estimation of discrete response models without numerical integration," *Econometrica*, 995–1026.
Miyauchi, Y., K. Nakajima, and S. J. Redding (2021) "The economics of spatial mobility: Theory and evidence using smartphone data," February, Working paper no. 28497. National Bureau of Economic Research, Cambridge, MA.
Neunhoeffer, M., D. Sheldon, and A. D. Smith (2023) "A Bootstrap-based General-purpose Approach for Statistical Inference with Differential Privacy," April, Working paper.
Ruggles, S., C. Fitch, D. Magnuson, and J. Schroeder (2019) "Differential privacy and census data: Implications for social and economic research," in *AEA Papers and Proceedings*, 109, 403–08.
Sakong, J. and A. K. Zentefis (2023) "Bank Branch Access: Evidence from Mobile Device Data," January, Working Paper. Yale University, New Haven, CT.
Silva, J. S. and S. Tenreyro (2006) "The log of gravity," *Review of Economics and Statistics*, 88 (4), 641–658.
Squire, R. F. (2019) "Quantifying sampling bias in SafeGraph Patterns," *SafeGraph Blog*, https://colab.research.google.com/drive/1u15afRytJMsizySFqA2EPlXSh3KTmNTQ#sandboxMode=true&scrollTo=xsNNli6GTN6s.
Thaenraj, P. (2021) "Identifying CBG Sinks," *SafeGraph Blog*, https://colab.research.google.com/drive/17-cpOxXN7PFUjdEHf329fmbyirmeuza6?usp=sharing#scrollTo=RRqtUyPBg40H.
Westerlund, J. and F. Wilhelmsson (2011) "Estimating the gravity model without gravity using panel data," *Applied Economics*, 43 (6), 641–649.

(A) All Block Group × Branch Pairs



(B) Block Group × Branch Pairs with >4 Visitors, with Fixed Effects

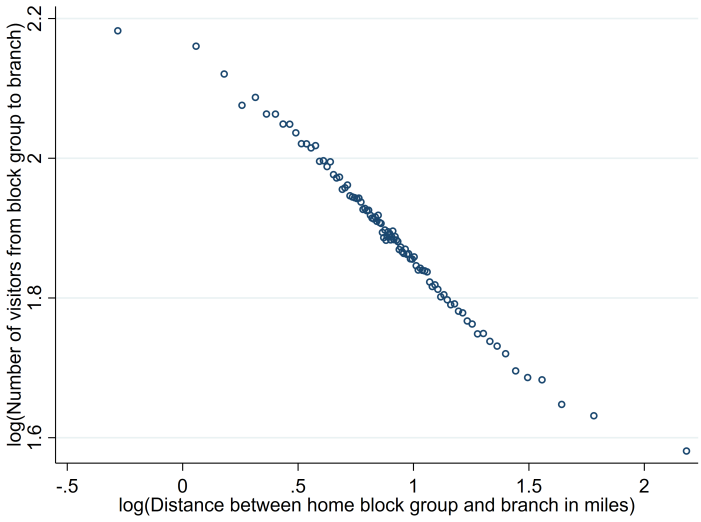


FIGURE 1
NUMBER OF VISITORS FROM BLOCK GROUPS TO BANK BRANCHES BY DISTANCE

The figure presents binned scatter plots of the log number of visitors from home Census block groups to bank branches according to the log mile distance between the block groups and branches. Visitor information is from our core SafeGraph sample ranging from January 2018 to December 2019. The core sample includes only businesses in SafeGraph with NAICS codes equal to 522110 (Commercial Banking), 522120 (Savings Institutions), or 551111 (Offices of Bank Holding Companies) for which we have visitor data and whose brands are also listed in the FDIC's 2019 Summary of Deposits. Distance is computed from the population-weighted center of a block group to a branch. Centers of population are from the 2010 Census, and we use the haversine formula to compute distance (see Footnote 8). Panel A presents the observed (raw) geolocation data and includes all block group × branch pairs, including those with visitor counts of 2 or 3 that SafeGraph rounds up to 4. Panel B only includes block group × bank branch pairs with greater than 4 visitors. In that panel, the log numbers of visitors are residualized by block group × year-month fixed effects and branch × year-month fixed effects. The log distances are residualized by the same set of fixed effects. To construct the binned scatter plots, we divide the x-axis values into 100 equal-sized (percentile) bins. We then calculate the mean of the y-axis values and the mean of the x-axis values within each bin. In addition, for Panel B we add back the unconditional mean of the log numbers of visitors and the unconditional mean of the log distances to re-scale values.

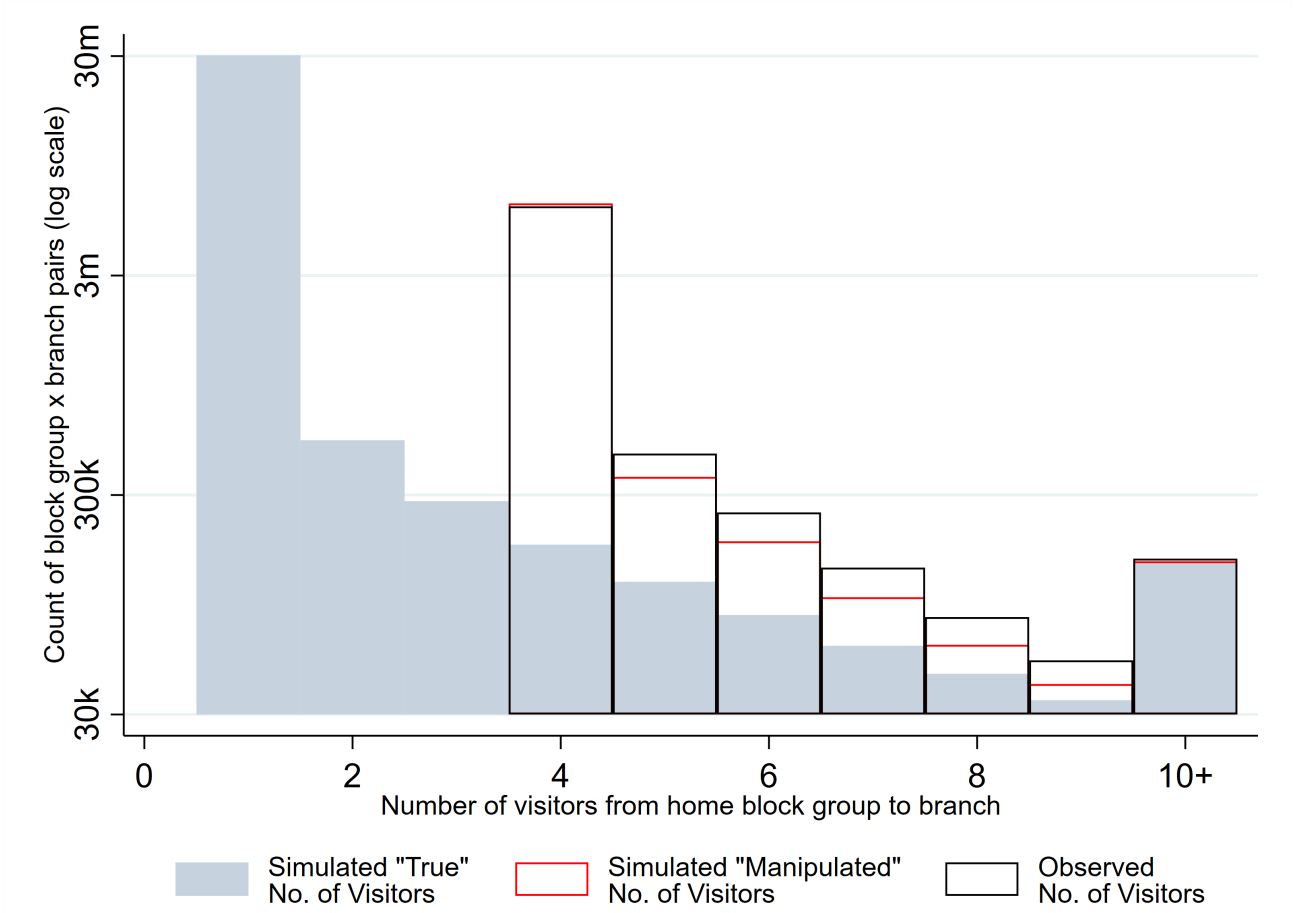


FIGURE 2
DISTRIBUTIONS OF VISITOR COUNTS

The figure presents distributions of observed visitor counts, simulated "true" visitor counts, and simulated "manipulated" visitor counts from visitors' home Census block groups to bank branches. Observed visitor counts, denoted V_{ijt} from Eq. (5), are the raw geolocation data from our core SafeGraph sample ranging from January 2018 to December 2019. The core sample includes only businesses in SafeGraph with NAICS codes equal to 522110 (Commercial Banking), 522120 (Savings Institutions), or 551111 (Offices of Bank Holding Companies) for which we have visitor data and whose brands are also listed in the FDIC's 2019 Summary of Deposits. Simulated "true" visitor counts, denoted V^*_{ijt} from Eq. (2), are draws from the underlying "true" distribution of visitors, which we assume to be Poisson. Simulated "maniputed" visitor counts are the "true" visitor counts after being manipulated via differential privacy methods presented in Eqs. (3) to (5). The simulated values are computed from the month-by-month Method of Simulated Moments estimation described in Section 4. The distribution of simulated visitor counts includes all positive draws from all simulations across every year-month in the sample period. To enhance the depictions of the distributions, we censor them at 10 visitors. That is, the number of block group $\times$ branch pairs with visitor counts exceeding 10 is assigned to 10+ visitors in the figure.

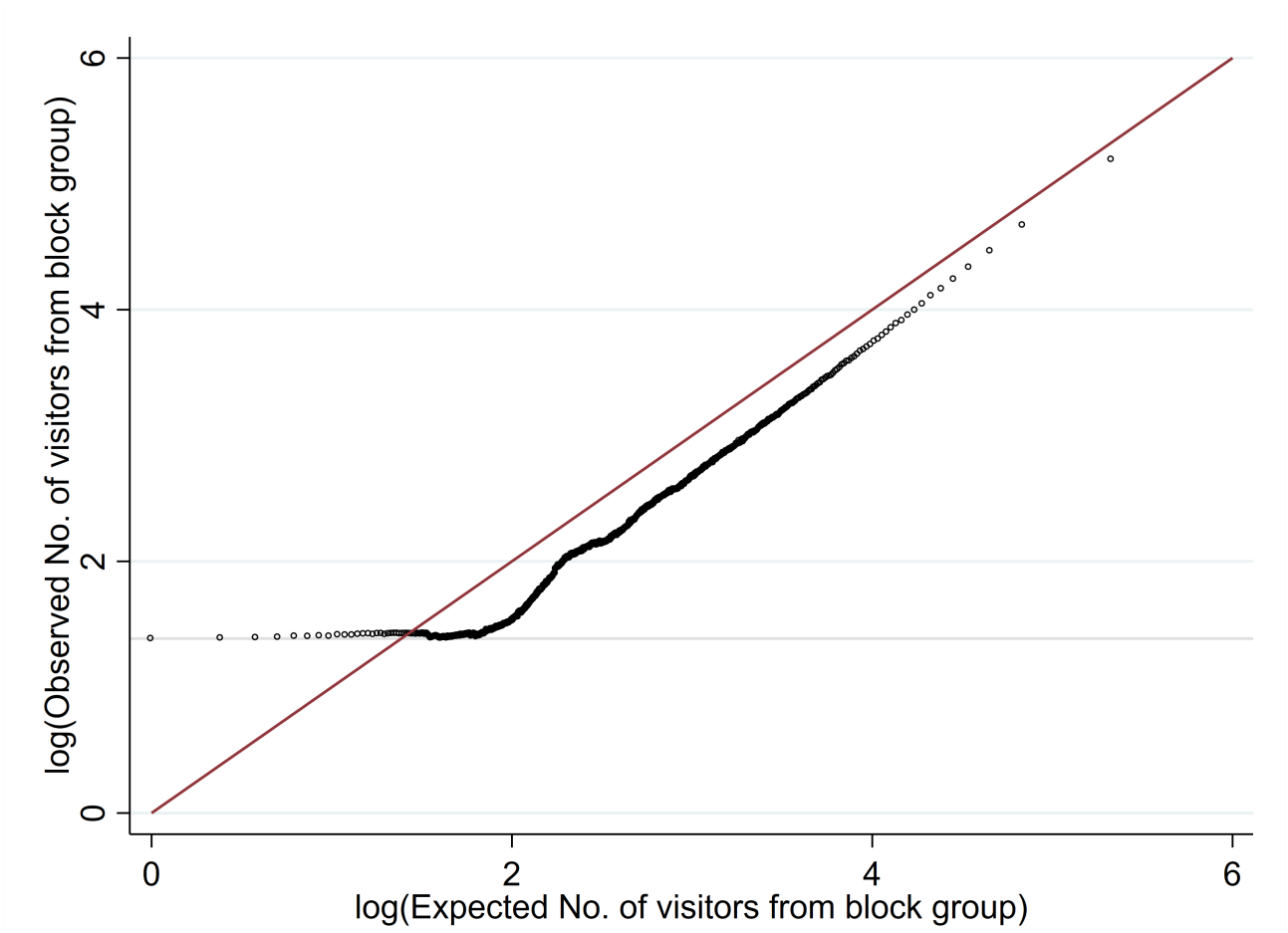
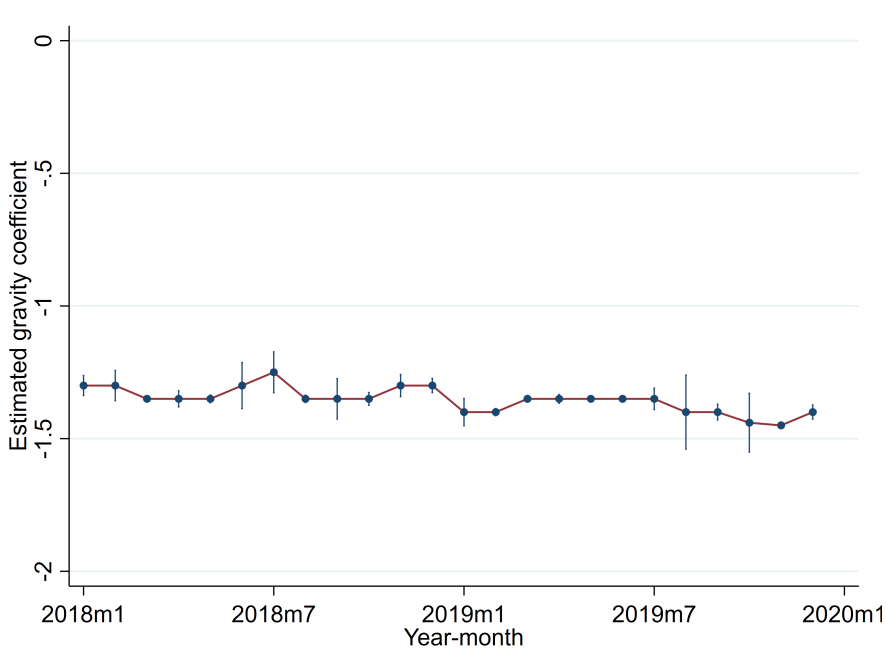


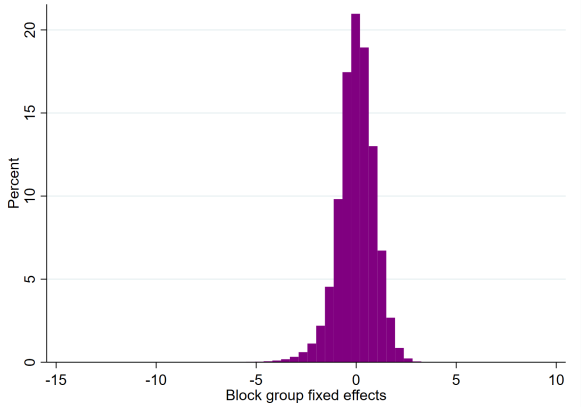
FIGURE 3
OBSERVED VS. EXPECTED BRANCH VISITORS PER CENSUS BLOCK GROUP

The figure presents a binned scatter plot of the log observed number of branch visitors from each Census block group (i.e., $\log V_{it} \equiv \log \sum_j V_{ijt}$, where V_{ijt} is given in Eq. (5)) versus the log expected number of branch visitors from each block group based on the month-by-month Method of Simulated Moments (MSM) estimates (i.e., $\log \hat{V}^a_{it} \equiv \hat{\gamma}_{it} + \log \hat{\Phi}^a_{it}$, where the access measure $\hat{\Phi}^a_{it} \equiv \sum_{j \in b_{it}} \omega^i_t \exp\left(\hat{\lambda}_{jt}\right) d_{ij}^{-\hat{\beta}_t}$ reflects the branch probability weights used in the stratified sampling and defined in Eq. (17)). The observed and expected number of visitors range over the full sample period from January 2018 to December 2019. Each dot represents a Census block group in a year-month. The red solid line is a 45° line and the light grey solid line cuts the y-axis at 1.4, which corresponds to SafeGraph's censoring at 4 visitor counts. The steps of the MSM procedure that generate the expected number of branch goers are in Section 4. To construct the binned scatter plot, we divide the x-axis values into 1,000 equal-sized bins. We then calculate the mean of the y-axis values and the mean of the x-axis values within each bin.

(A) Time Series of $-\hat{\beta}_{t,\text{MSM}}$



(B) Census Block Group Fixed Effects



(C) Bank Branch Fixed Effects

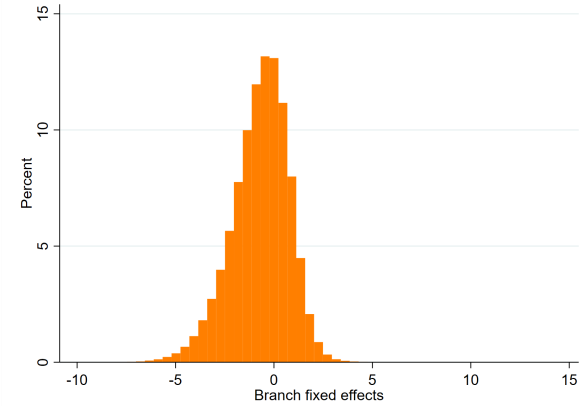


FIGURE 4
METHOD OF SIMULATED MOMENTS PARAMETER ESTIMATES

The figure presents the parameter estimates from the month-by-month Method of Simulated Moments (MSM) estimation of the visitor count gravity relation in Eq. (2). Panel A illustrates the monthly time series of the $-\hat{\beta}_{t,\text{MSM}}$ gravity coefficient estimates, along with 95% confidence intervals. Panel B presents a histogram of the estimated Census block group fixed effects, $\left\{\hat{\gamma}_{it}^{\infty}\right\}$, and Panel C presents a histogram of the estimated bank branch fixed effects, $\left\{\hat{\lambda}_{jt}^{\infty}\right\}$. In each histogram, the fixed effects are grouped into 50 equally-sized bins, and the estimated fixed effects for all months in the sample period are presented. A summary of the MSM estimation is provided in Section 4.

TABLE 1
DESCRIPTIVE STATISTICS - CORE SAFEGRAPH SAMPLE

	Mean	Std. Dev.	P10	P25	P50	P75	P90	N
No. of Visits	67	180	6	14	35	78	147	919,076
No. of Visitors	40	94	5	10	23	48	90	919,076
Med. Dist. from Home (mi)	5	16	2	3	4	6	9	822,569
Med. Dwell Time (min)	49	102	6	7	9	30	152	919,076
Device Type - iOS	52%							19,238,792
Device Type - Android	46%							17,207,356

The table reports descriptive statistics of key variables related to bank branch visitation. All values are based on our core sample of geolocation data, which consists of businesses in SafeGraph with NAICS codes equal to 522110 (Commercial Banking), 522120 (Savings Institutions), or 551111 (Offices of Bank Holding Companies) for which we have visitor data and whose brands are also listed in the FDIC's 2019 Summary of Deposits. Data are monthly, at the branch level, and range from January 2018 - December 2019. *No. of Visits* is the total number of visits to a typical bank branch in a month. *No. of Visitors* is the total number of visitors (i.e., mobile devices) to a typical branch in a month. *Med. Dist. from Home (mi)* is the median distance in miles that visitors travel to a branch from their home (among visitors whose home is identified). *Med. Dwell Time (min)* is the median amount of time in minutes that visitors stay at a branch. *Device Type* is the fraction of total branch visitors using Google Android vs. Apple iOS mobile devices. The number of observations *N* used in the first four rows is the total number of branch-year-months. The number of observations used in the last two rows is the total number of mobile devices with device-type information over the core sample period.

TABLE 2
DEMOGRAPHIC ATTRIBUTES OF RESIDENTS IN AREAS REPRESENTED IN CORE SAFEGRAPH SAMPLE VS. FDIC SOD

	Core Sample $\hat{\mu}_1$ $\hat{\sigma}_1$	SOD $\hat{\mu}_2$ $\hat{\sigma}_2$	Diff $\hat{\mu}_1 - \hat{\mu}_2$ (se)		Core Sample $\hat{\mu}_1$ $\hat{\sigma}_1$	SOD $\hat{\mu}_2$ $\hat{\sigma}_2$	Diff $\hat{\mu}_1 - \hat{\mu}_2$ (se)
N branch	51,369	86,374	-35,005	< HS	0.104	0.104	-0.000
					0.075	0.075	(0.000)
White	0.799	0.805	-0.006	HS degree	0.260	0.263	-0.003
	0.184	0.183	(0.001)		0.103	0.106	(0.001)
Black	0.103	0.095	0.008	Some college	0.199	0.198	0.001
	0.153	0.146	(0.001)		0.052	0.054	(0.000)
Asian	0.046	0.047	-0.001	College degree	0.297	0.295	0.002
	0.076	0.082	(0.000)		0.090	0.091	(0.001)
Hispanic	0.109	0.106	0.003	> College	0.140	0.140	0.000
	0.155	0.156	(0.001)		0.098	0.100	(0.001)
Homeowner	0.645	0.643	0.002	Unemp rate	0.050	0.050	0.000
	0.168	0.173	(0.001)		0.026	0.028	(0.000)
Age 15-34	0.188	0.190	-0.001	HH income ($)	70,338	69,802	536
	0.089	0.093	(0.001)		29,657	30,052	(166)
Age 35-54	0.350	0.347	0.004	In poverty	0.127	0.129	-0.002
	0.069	0.069	(0.000)		0.082	0.084	(0.000)
Age 55-64	0.195	0.195	-0.000	Urban	0.815	0.787	0.028
	0.038	0.039	(0.000)		0.295	0.319	(0.002)
Age 65+	0.267	0.269	-0.002	Home value ($)	286,881	290,086	-3,205
	0.086	0.085	(0.000)		243,368	254,504	(1,383)

The table compares demographic characteristics of residents of geographic areas represented in our core sample of bank branches with characteristics of residents of geographic areas represented in the FDIC's 2019 Summary of Deposits (SOD). Our core sample consists only of businesses in SafeGraph with NAICS codes equal to 522110 (Commercial Banking), 522120 (Savings Institutions), or 551111 (Offices of Bank Holding Companies) for which we have visitor data and whose brands are also listed in the SOD. Demographic characteristics in the table are taken from the 2019 5-year American Community Survey and are averaged at the level of the Census Bureau's zip code tabulation areas (ZCTA). *White* includes both Hispanic and non-Hispanic White racial shares. The only non-demographic variable is *N branch*, which is the number of branches in the core sample and the SOD. In columns (1), (2), (4), and (5), the first row of each demographic attribute is its sample mean across ZCTAs, whereas the second row is the sample standard deviation. In columns (3) and (6), the first row is the difference in sample means between the core sample and the SOD, whereas the second row is the heteroskedasticity-robust standard error of the estimated difference between the two sample means.

TABLE 3
COMPARING GRAVITY EQUATION ESTIMATION METHODS

		MSM		PPML		OLS		OLS where ≥ 4	
Year	Month	β	s.e.	β	s.e.	β	s.e.	β	s.e.
2018	1	-1.26	(0.035)	-0.066	(0.003)	-0.038	(0.001)	-0.331	(0.030)
	2	-1.31	(0.227)	-0.072	(0.004)	-0.042	(0.001)	-0.319	(0.023)
	3	-1.32	(0.019)	-0.076	(0.003)	-0.046	(0.001)	-0.295	(0.018)
	4	-1.33	(0.033)	-0.073	(0.002)	-0.045	(0.001)	-0.287	(0.016)
	5	-1.32	(0.011)	-0.075	(0.003)	-0.045	(0.001)	-0.297	(0.017)
	6	-1.30	(0.007)	-0.072	(0.002)	-0.045	(0.001)	-0.288	(0.017)
	7	-1.27	(0.043)	-0.069	(0.002)	-0.043	(0.001)	-0.278	(0.018)
	8	-1.29	(0.053)	-0.079	(0.003)	-0.047	(0.001)	-0.317	(0.018)
	9	-1.34	(0.304)	-0.082	(0.002)	-0.049	(0.001)	-0.340	(0.022)
	10	-1.37	(0.090)	-0.086	(0.003)	-0.051	(0.001)	-0.303	(0.016)
	11	-1.31	(0.032)	-0.086	(0.003)	-0.051	(0.001)	-0.293	(0.014)
	12	-1.31	(0.035)	-0.091	(0.003)	-0.053	(0.001)	-0.269	(0.014)
2019	1	-1.40	(0.018)	-0.089	(0.003)	-0.053	(0.001)	-0.300	(0.014)
	2	-1.43	(0.030)	-0.089	(0.002)	-0.053	(0.001)	-0.286	(0.015)
	3	-1.37	(0.035)	-0.096	(0.003)	-0.056	(0.001)	-0.279	(0.014)
	4	-1.39	(0.016)	-0.098	(0.003)	-0.056	(0.001)	-0.268	(0.012)
	5	-1.40	(0.023)	-0.106	(0.003)	-0.061	(0.001)	-0.258	(0.010)
	6	-1.38	(0.177)	-0.096	(0.002)	-0.057	(0.001)	-0.274	(0.010)
	7	-1.35	(0.106)	-0.095	(0.003)	-0.056	(0.001)	-0.261	(0.011)
	8	-1.40	(0.039)	-0.103	(0.003)	-0.061	(0.001)	-0.270	(0.010)
	9	-1.41	(0.034)	-0.108	(0.003)	-0.060	(0.001)	-0.290	(0.011)
	10	-1.45	(0.031)	-0.102	(0.003)	-0.059	(0.001)	-0.291	(0.012)
	11	-1.43	(0.015)	-0.099	(0.003)	-0.058	(0.001)	-0.290	(0.013)
	12	-1.41	(0.033)	-0.105	(0.003)	-0.062	(0.001)	-0.285	(0.010)
Panel				-0.091	(0.002)	-0.053	(0.001)	-0.283	(0.008)

The table reports estimates and standard errors of the gravity coefficient β_t from the fixed-effects gravity model in Eq. (1):

$$\log\left(\text{No. of visitors}_{ijt}\right) = \gamma_{it} + \lambda_{jt} - \beta_t \log\left(\text{Distance}_{ij}\right) + \varepsilon_{ijt}.$$

Columns (3) and (4) present estimates from the Method of Simulated Moments estimation described in Section 4. Columns (5) and (6) present estimates from an unweighted Poisson pseudo-maximum-likelihood (PPML) estimation, as in Silva and Tenreyro (2006), run using `ppmlhdfe` in Stata. Columns (7)-(10) present estimates from an unweighted OLS regression. The PPML and OLS estimations use the raw number of visitors from home Census block groups to bank branches based on our core sample of geolocation data, which consists of businesses in SafeGraph with NAICS codes equal to 522110 (Commercial Banking), 522120 (Savings Institutions), or 551111 (Offices of Bank Holding Companies) for which we have visitor data and whose brands are also listed in the FDIC's 2019 Summary of Deposits. Columns (9) and (10) restrict the sample to visitor counts of at least 4, which circumvent SafeGraph's truncation and censoring. The MSM, PPML, and OLS gravity coefficient estimates are calculated month-by-month over the sample period (January 2018 - December 2019). PPML and OLS estimates are also calculated over the full sample panel. Standard errors of the MSM estimates are described in Section 4. Standard errors of the PPML and OLS estimates are two-way clustered by both Census block groups and bank branches.

Appendix

A Core Sample Construction

Here, we supply background information on the SafeGraph geolocation data and a detailed explanation of how we construct our core sample.

A.1 SafeGraph Geolocation Data

We use two of SafeGraph's primary datasets: Core Places and Patterns. Both datasets have information on millions of points-of-interest (POIs) in the United States, which SafeGraph defines as "specific location[s] where consumers can spend money and/or time."[10] Locations such as restaurants, grocery stores, parks, museums and hospitals are included, but not residential homes or apartment buildings.

The Core Places dataset provides the establishment name (e.g., Salinas Valley Ford Lincoln), brand (e.g., Ford), six-digit NAICS code, latitude and longitude coordinates, address, phone number, hours open, when the establishment opened, and when SafeGraph began tracking information about the establishment. SafeGraph describes creating this dataset using thousands of diverse sources. We use the January 2021 version of the Core Places dataset, which was the most up-to-date and accurate as of the time of our analysis.

The Patterns dataset contains information on visitors to different locations. A visitor is identified via his or her mobile device, and one device is treated as one visitor. SafeGraph collects this information from third-party mobile application developers. Through these mobile applications, SafeGraph gathers a device's advertisement identifier, the latitude and longitude coordinates of the device at a designated time, and the horizontal accuracy of the geographic coordinates.[11] In this dataset, SafeGraph aggregates the visitor data and provides several bits of information, including the number of visits and unique visitors to a POI during a specified date range, the median distance from home that visitors traveled to reach the POI, the median dwell time spent at the POI, and the number of visitors using Apple's iOS or Google's Android operating system. The Patterns dataset is backfilled to reflect the Core Places from the January 2021 version.

Most importantly for us, the Patterns dataset contains the home Census block groups of visitors, and the number of visitors from each of those home block groups. To protect user privacy, SafeGraph employs differential privacy methods to the visitor home block group data. First, it adds Laplace noise to each block group's visitor count (when it observes at least one visitor from the block group). Second, after the noise is added, Safegraph rounds the visitor counts down to their nearest integers. Third, SafeGraph then truncates the rounded visitor counts by only reporting data from block groups with at least two visitors. Fourth, home block groups with only two, three, or four visitors are reported as having four visitors.

SafeGraph determines a visitor's home Census block group using an algorithm. A brief description of that algorithm is as follows. The algorithm starts by clustering GPS signals from a device during the nighttime hours between 6pm - 7am local time. The Census block group with the most clusters is recorded as the device's potential home location for the day. SafeGraph reviews the previous six weeks of the device's daily home locations and identifies the most frequent one as the device's home Census block group. This home location applies for the device over the next thirty days, at which point the home location is updated. New devices that appear in the panel require at least five days of data before they are eligible to have their home locations identified. Finally, SafeGraph computes a confidence score for each device's calculated home block group. Only high-confidence home locations are included; otherwise, the device's home location is classified as unknown.[12]

[10] See the SafeGraph Places Manual and Data Guide for more details.
[11] See the SafeGraph Privacy Policy for more details.
[12] Full details of the algorithm are found here: Home Identification Algorithm.

A.2 FDIC Summary of Deposits

To construct our core sample, we rely on branch information from the Federal Deposit Insurance Corporation (FDIC). Branch data are from the FDIC's 2019 Summary of Deposits (SOD).[13] We rely on the SOD to confirm that branch locations we use from SafeGraph belong to actual depository institutions, instead of other financial institutions that SafeGraph might mistakenly label as a "bank," but do not take deposits, such as an investment advisory firm.

A.3 Construction Process

Our core sample can be thought of as consisting of two components: (i) a set of locations and (ii) consumer movement to those locations. We call these two components "places" and "visitors." In our case, the places and visitors are specific to bank branches. SafeGraph is our only source of visitor data, and so, we rely on it exclusively. The visitors data field we use that contains the home Census block groups of the visitors to a branch is `VISITOR_HOME_CBGS`. As we describe in the text, this data field is subject to SafeGraph's differential privacy.

Places data, on the other hand, are available in both SafeGraph and the SOD. Before we detail how we make use of both sources, we first need to introduce *placekey*, which is a crucial way we identify a place.

A.3.1 Placekey

Placekey is a free, standardized identifier of physical locations. It supplants a location's address and latitude-longitude geocode with a unique identifier. Using this identifier overcomes the challenge of linking locations by addresses that are spelled differently (e.g., 1215 Third Street, Suite 10 vs. 1215 3rd St., #10) or by latitude-longitude geocodes that differ slightly but refer to the same place.

A business's placekey consists of two parts (called "What" and "Where"), and it is written as What@Where. The What component encodes an address and a point-of-interest. The point-of-interest piece adjusts if a new business opens at the same address of a previous business that closed. For example, if a bank branch closed, but its building converted into a bakery, the two businesses would share the same address, but different points-of-interest; and therefore, they would be assigned different placekeys.

The Where component consists of a unique character sequence. It encodes a hexagonal region on the surface of the Earth based on the latitude and longitude of the business. The hexagon contains the centroid of the business, and the Where component is the full encoding of the hexagon. To consider an example Placekey, take the Chase branch at 1190 S. Elmhurst Rd. in Mount Prospect, IL 60056. This branch's placekey is 223-222@5sb-8gg-jn5. Additional technical information about Placekey can be found in their white paper located here: Placekey White Paper.

A.3.2 Choosing the Set of Places

Both the SOD and SafeGraph have bank branch locations. SafeGraph locations are already identified by their placekeys. We generate placekeys for the SOD locations using Placekey's free API. To construct an accurate and comprehensive set of places, we take advantage of place information in SafeGraph and the SOD. The *quality* of SafeGraph places is higher than those in the SOD. Often, an address in the SOD has an invalid placekey, and a Google Maps search confirms that no physical place exists at that address. (The place's absence is not due to a branch closing.) A higher quality set of places from SafeGraph should come at little surprise, as the success of the company's business relies in part on providing highly accurate place information.

[13] FDIC SOD data are located here: SOD.

On the other hand, the *quantity* of places is higher in the SOD than in SafeGraph. In SafeGraph, bank branches are classified by their 6 digit NAICS codes (522110 for Commercial Banking, 522120 for Savings Institutions, and 551111 for Offices of Bank Holding Companies). The number of places in SafeGraph under these categories is less than the number of branches in the SOD. So that we can link places information to visitor information, all places we analyze must be included in SafeGraph. For example, a branch in the SOD that is not part of SafeGraph whatsoever has no visitor information to study. But we can use place information from the SOD to choose the set of places from SafeGraph that balances quality and quantity. Doing so constructs our core sample, which we define next.

Our **core sample** includes only SafeGraph places with brands that are included in the SOD and for which we have visitor geolocation data from SafeGraph. In the SOD, the field `CERT` identifies a unique banking institution. We rely on this field to select the list of unique banks, and we use the union of the SOD fields `namefull` and `namehcr` to identify a bank's brand. In SafeGraph, we use the field `LOCATION_NAME` to identify a bank brand name. For example, Wells Fargo & Company and SunTrust Banks, Inc. are two bank brands with locations in the SOD. All Wells Fargo and SunTrust Bank places in SafeGraph would be included, and their locations would be identified by SafeGraph's placekeys for them. All SOD locations (and their placekeys) are ignored.

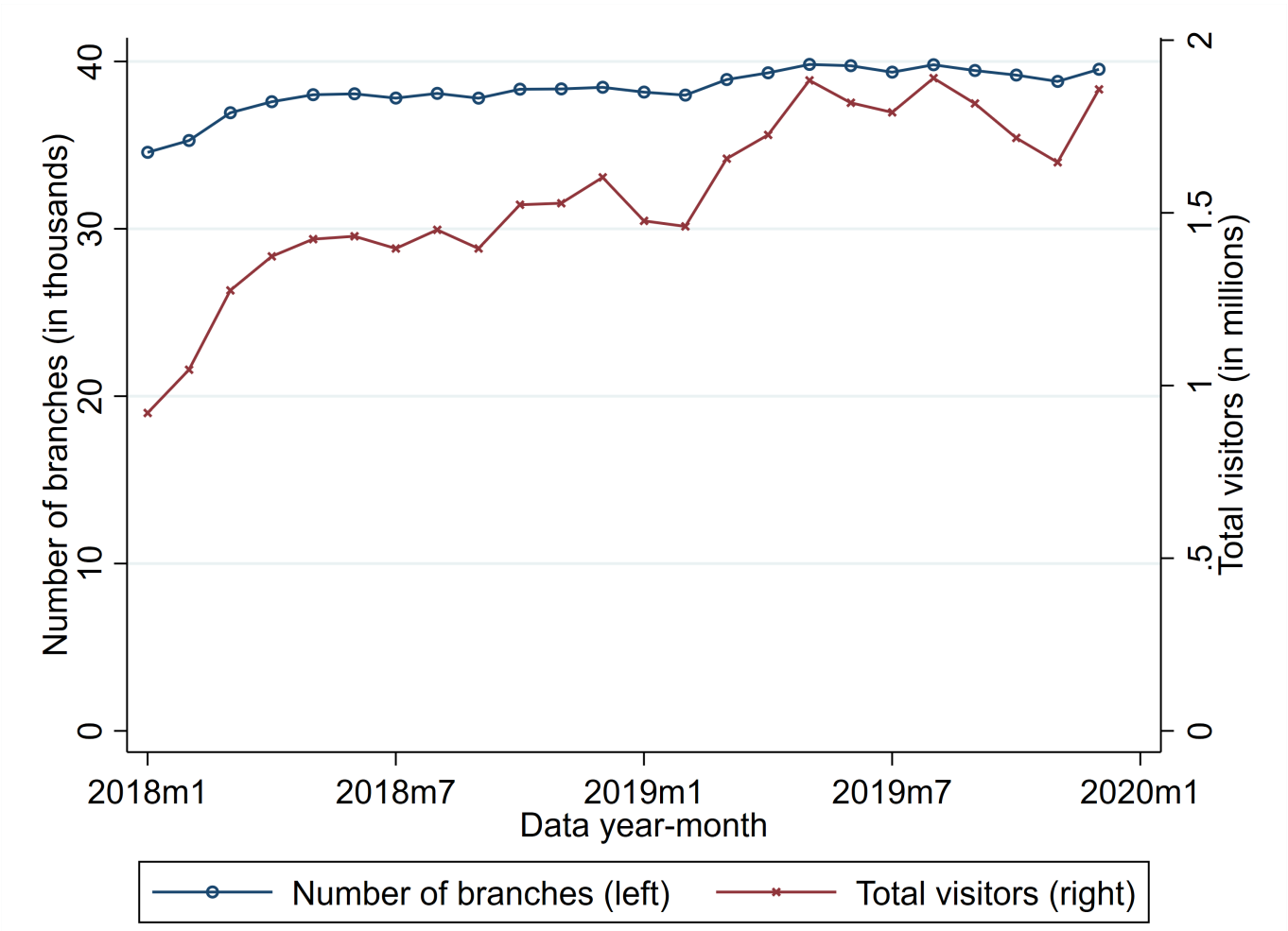


FIGURE A.1
NUMBER OF BANK BRANCHES AND BRANCH VISITORS - CORE SAMPLE

The figure presents the number of bank branches and number of branch visitors each year-month in our core sample. The core sample includes only businesses in SafeGraph with NAICS codes equal to 522110 (Commercial Banking), 522120 (Savings Institutions), or 551111 (Offices of Bank Holding Companies) for which we have visitor data and whose brands are also listed in the FDIC's 2019 Summary of Deposits.

Table A.1
Driving Time versus Haversine Distance

Dep. var.:	Driving time b/w block group and visited branch								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Haversine distance	0.641	0.634	0.631	0.632	0.649	0.652	0.634	0.647	0.632
b/w block group and visited branch	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Constant	57.610	60.684	58.863	64.234	51.796	51.427	58.749	50.702	67.206
	(0.177)	(0.150)	(0.494)	(0.156)	(0.319)	(0.289)	(0.195)	(0.303)	(0.171)
Observations	995,000	725,000	35,000	498,000	497,000	498,000	497,000	508,000	487,000
Adjusted R^2	0.982	0.991	0.993	0.992	0.972	0.973	0.990	0.975	0.990
Sample	Core	MC	Core	Core	Core	Core	Core	Core	Core
Black > 0.8			O						
Black ≥ Med. Black				O					
Black < Med. Black					O				
White ≥ Med. White						O			
White < Med. White							O		
log(Income) ≥ Med. log(Income)								O	
log(Income) < Med. log(Income)									O

Each column reports coefficients from a univariate, weighted OLS regression with heteroskedasticity-robust standard errors reported in parentheses. One observation is a block group × branch pair from our core sample of Census block groups and bank branches, where the branches consist of businesses in SafeGraph with NAICS codes equal to 522110 (Commercial Banking), 522120 (Savings Institutions), or 551111 (Offices of Bank Holding Companies) for which we have visitor data and whose brands are also listed in the FDIC's 2019 Summary of Deposits (SOD). Observations are weighted by block-group population counts from the 2019 5-year American Community Survey (ACS). Dependent variable observations are the driving times from the population-weighted centers of block groups to branches, where driving times are computed using the Origin-Destination Cost Matrix of ArcGIS Pro under the default settings. Centers of population are from the 2010 Census. Independent variable observations are the corresponding haversine distances between block groups and branches. 995,000 block group × branch pairs were drawn randomly. Column (1) includes the entire random sample of block group × branch pairs. Column (2) restricts the sample to block groups with Rural-Urban Commuting Areas (RUCA) codes equaling 1 (Metropolitian area core). Column (3) restricts the sample to block groups with Black population shares exceeding 80%. Column (4) restricts the sample to block groups with Black population shares at or exceeding the median Black population share across all block groups in the entire random sample. Column (5) restricts the sample to block groups with Black population shares below the median Black population share across all block groups in the entire random sample. Column (6) restricts the sample to block groups with White population shares at or exceeding the median White population share across all block groups in the entire random sample. Column (7) restricts the sample to block groups with White population shares below the median White population share across all block groups in the entire random sample. Column (8) restricts the sample to block groups with the natural logarithm of median household income at or exceeding the median of the natural logarithm of median household income across all block groups in the entire random sample. Column (9) restricts the sample to block groups with the natural logarithm of median household income below the median of the natural logarithm of median household income across all block groups in the entire random sample. Racial shares and median household income are from the 2019 5-year ACS.