BANK ACCESS ACROSS AMERICA*

Jung Sakong† Alexander K. Zentefis‡

September 18, 2022

Abstract

Low-income and Black households are less likely to own bank accounts and visit bank branches than high-income and White households. We assess whether unequal bank access can explain why. We develop a measure of bank access based on the locations of bank branches and consumer trips to those branches from mobile device data. Residents have better bank access if branches are located closer to them or if those branches have superior qualities that attract more visitors (like better loan rates or customer service). We find no evidence that low-income communities experience a lack of access, implying that their lower branch use is plausibly due to lower demand for banks. But in Black communities, a lack of access explains the entire drop off in branch use. For residents of these areas, weaker access is not from having lower quality branches, but from branches being located farther away from them. The results highlight areas of the country that would benefit the most from public policies that expand access to banking.

JEL classification: G21, J15, L87, R22
Keywords: banking, inequality, location economics, spatial analysis

*We give special thanks to Bo Honoré and Luojia Hu for suggesting parts of the econometric method we use in this article 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 Treb Allen, Costas Arkolakis, Asaf Bernstein, Lorenzo Caliendo, Judy Chevalier, Tony Cookson, Jonathan Dingel, Jane Dokko, Andrea Eisfeldt, Cecilia Fieler, Raffi Garcia (discussant), Paul Goldsmith-Pinkham, Gary Gorton, Jessie Handbury, Xuan Hung Do (discussant), Kristoph Kleiner (discussant), Sam Kortum, Simone Lenzu, Noura Kone (discussant), Runjing Lu (discussant), Yuhei Miyauchi, Luciana Orozco (discussant), Piyush Panigrahi, Karen Pence (discussant), Matthieu Picault (discussant), Roberto Robatto, Claudia Robles-Garcia, Rosa Sanchis-Guarner (discussant), Shri Santosh, Katja Siem, Fiona Scott Morton, Brad Shapiro (discussant), Kelly Shue, Mike Sinkinson, 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

The benefits of bank account ownership and use are well documented. They include improved access to credit (Fitzpatrick 2015), higher subjective well being (Prina 2015), greater liquid savings (Agarwal, Alok, Ghosh, Ghosh, Piskorski, and Seru 2017), and reduced reliance on costly payday lending (Melzer 2018). The mere exposure to local bank branches causally increases bank participation, leading to greater wealth accumulation (Célerier and Matray 2019) and enhanced financial literacy (Brown, Cookson, and Heimer 2019).

And yet, substantial disparities in bank account ownership and use persist in the United States. According to the 2019 FDIC Survey of Household Use of Banking and Financial Services, the share of Black households that are unbanked is 2.5 times the national average, and the share of low-income households earning less than $15,000 that are unbanked is over 4 times the national average. Disparities exist even in the use of bank branches, as both Black and low-income households are significantly less likely to visit a branch within the previous year compared to White and high-income households. This lower use of bank branches does not appear to be offset with greater use of online or mobile banking, as Black and low-income households report greater reliance on bank tellers/ATMs as their most common banking method.

Among policymakers and in the media, a common argument to explain these disparities is a lack of bank access (Friedline and Despard 2016; Dahl and Franke 2017; Davidson 2018). This reasoning has motivated policy proposals that seek to increase access, like investment in community development banks (Ellwood and Patel 2018) or the expansion of U.S. postal banking (Baradaran 2013; Warren 2014; Johnson 2017; Gillibrand 2021; Sanders 2021). But demand-related factors, such as low cash savings or distrust in banks, are equally reasonable explanations. Existing research has struggled to convincingly control for these demand-related factors, so as to isolate the extent to which weak access explains continuing disparities in U.S. bank participation.

In this paper, we implement a research design that separates out the impact of access on household use of bank branches. The design relies on a rich panel of consumer travel patterns between home Census block groups and bank branches throughout the United States, originating from mobile device data. We find no evidence that low-income communities experience a lack of bank access, which implies that lower demand for banks explains their lower use of bank branches. But in Black communities, a lack of access explains their entire drop off in branch use. These results reveal which parts of the country would benefit the most from policies that increase access to banks.

In the first part of the article, we estimate our measure of access based on the mobile device data.

We represent the consumer movements between home block groups and bank branches with a spatial gravity equation, which consists of block group × time fixed effects, bank branch × time fixed effects, and the distances between pairs of block groups and branches. This approach takes advantage of the network of trips that links block group residents to multiple branches. By using block group fixed effects, we compare how residents of the *same* block group visit different branches by varying amounts. The extent to which this *within* block group comparison fully absorbs the block group residents' demand for branch services, the estimated differences between block groups in overall branch use can plausibly be attributed to differential access.[1]

The gravity equation not only allows us to control for demand-related factors, but it also generates a local measure of bank access. One can break apart a block group's expected total number of branch goers into one part due to "demand" (the exponentiated block group fixed effect) multiplied by a second part due to "access." We use "demand" informally to capture all resident characteristics that contribute to visiting *any* bank branch (e.g., average wealth, income, financial sophistication, trust in banks, flexibility in time). The "access" part is an index of bank branches available to residents of the block group. Each branch in a block group's index is represented by the branch's attributes, as captured by the branch's fixed effect, and the transportation costs that block group residents must bear to travel to the branch. A branch's fixed effect proxies for its "quality," capturing all characteristics of the branch that make it a destination for residents of *any* block group in the period (e.g., the branch having better deposit or loan rates, higher staff attentiveness, or an efficient drive-through ATM).

By this index, residents have better bank access if branches are located closer or if the nearest branches have superior attributes that attract more visitors. Unlike pure supply-side measures of access, such as the number of branches per capita, which reveal only the opportunity set available to consumers, this measure of access embodies consumers' actual choices. The measure is conceptually related to indices in the economic geography and trade literature that describe an exporting country's access to the importing markets of other countries (e.g., Harris 1954; Head and Mayer 2004; Redding and Venables 2004; Hanson 2005; De Sousa, Mayer, and Zignago 2012; and Donaldson and Hornbeck 2016).

While the mobile device data on consumer flows are critical to estimate the local measure of bank access, the data are subject to differential privacy methods, which try to shield the personal identifiable information of individuals from becoming public. Noise is added to the number of visitors from a block group to a branch, and these visitor counts are either truncated or censored if the number is

[1]Using block group fixed effects to control for consumer demand is akin to using firm fixed effects to absorb credit demand shocks when estimating the consequences of credit supply shocks, in the style of Khwaja and Mian (2008).

too low. Because of these distortions, an OLS regression of the gravity equation would render biased estimates and contaminate the measure of access.

To account for the differential privacy, we instead use the Method of Simulated Moments (MSM) to estimate the gravity equation. Standard MSM is straightforward to implement in models with no or few fixed effects (McFadden 1989). But the gravity equation has hundreds of thousands of fixed effects across block groups and branches, which complicates the procedure. To handle such a large volume of fixed effects, we introduce an econometric method to estimate the gravity equation. The method can be implemented in other empirical settings involving high-dimensional fixed effects in MSM estimation.

We run the MSM 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 -1.25, 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 1.25% 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.

In the second part of the article, we use the gravity model estimates to uncover how bank access varies across the country. We document three geographic patterns: First, access varies substantially across regions. Residents of the eastern shores of New England and the Mid-Atlantic experience the highest access nationwide, whereas residents of the Deep South observe considerably weaker access. Second, the most pronounced differences in access are between urban and rural areas. Residents of big cities like Boston, Miami, Houston, and Minneapolis observe roughly three times higher access to banks than residents of rural towns. Third, even within a town or city, bank access varies greatly. A population-weighted block-group-level regression of bank access on Census tract fixed effects estimates that just 26.2% of cross block-group variance in access is within Census tracts. For example, in New York City, access is highest in Midtown Manhattan and lowest in the Bronx, and even within each borough there is large variation. Likewise, around Chicago, residents of the North side experience better access than residents of the South side; and in Greater Los Angeles, the neighborhoods of Beverly Hills and the Hollywood area observe meaningfully better access than parts south of the city, such as Compton and Long Beach.

After examining geographic variation in bank access, we evaluate how access covaries with the household incomes and racial makeups of local communities. We compute national statistics, but we also zero in on large Metropolitan core areas, because these parts of the country contain the majority of U.S. households, and they have Black population shares that resemble the national average.

Understanding bank access in these big cities is essential to explaining why Black households use branches less.

With controls for the racial and age population shares of block groups, we find that residents of low-income block groups both cross-country and in Metro cores experience better bank access than residents of high-income block groups. The better access can arise from either low-income residents living relatively closer to branches or visiting branches of higher quality. To determine which channel explains the result, we split access into two components such that their product equals the original measure. The first component equalizes the quality of all branches so that any variation in the component across block groups is entirely driven by residents' proximity to branches. The second component normalizes the total transportation costs of each block group's residents so that variation in the component is entirely driven by differences in the weighted average quality of branches that residents experience, where closer branches to the block group are given higher weights. Performing this separation, we find that the higher access in low-income communities is completely due to better branch proximity and not quality. We find no significant difference in the average quality of branches that residents of low-income and high-income areas experience.

Unlike residents of low-income block groups who observe relatively better bank access, we find that residents of block groups with high Black population shares experience *worse* access than residents living in block groups with high White population shares. This finding controls for a block group's median household income and the population shares of its residents' ages. Splitting access into the two components like before reveals that lower branch proximity is the reason behind the worse access. Residents of areas with high Black population shares experience *better* average branch quality compared to residents of areas with high White population shares. But because branches are located farther away from Black communities, those communities bear larger transportation costs to reach any branch. These costs more than offset the advantage of having the closest branches be of higher quality, leading to lower overall bank access.

In the third part of the article, we isolate the extent to which access or demand explains the observed racial and income disparities in branch use. We start by regressing the natural logarithm of the expected total number of branch goers per block group, as estimated from the gravity equation, onto demographic characteristics of the block group residents. Controlling for block group racial and age shares, we find that every doubling in a block group's median household income is associated with 15.5 percent more branch goers per month, both nationwide and in Metro cores. This income gradient in branch use is large, as the unconditional likelihood of a resident in the data visiting a bank branch during the year is about 72 percentage points. Examining differences by race, and controlling

for median household income and age shares, we find that a block group with a 100% Black population share observes roughly 5.6 percent fewer branch goers per month relative to a block group with a 100% White population share. In Metro cores, this Black-White gap in branch use is slightly smaller at 3.9 percent per month.

We next decompose the income gradient and the Black-White gap in branch use into constituent parts due to access and demand. Doing so simply requires regressing the natural logarithm of each block group's measure of access and the block group's estimated fixed effect onto the same set of block-group-level characteristics that we use to evaluate each block group's log expected total number of branch goers. By construction from the gravity equation, the estimated coefficients from the three regressions satisfy an identity along any demographic attribute. The identity separates the elasticity of branch use with respect to the demographic attribute into two pieces: the elasticity of access and the elasticity of demand.

Performing this decomposition, we find that the 15.5 percent income gradient in branch use nationwide consists of a -7.5 percent income gradient in access and a +23.0 percent income gradient in demand. Thus, while residents of low-income block groups have relatively better access to banks, they exhibit a lower propensity to visit any bank branch, which translates overall into lower branch use compared to residents of high-income block groups. The Black-White gap in branch use nationwide has a different explanation. Black communities exhibit no robust statistical difference in their demand for branch services nationwide compared to White communities. And yet, residents of Black communities visit branches less, which implies that the 5.6 percent Black-White gap in branch use cross-country is entirely due to weaker access.

In Metro cores, the decomposition by income reveals a similar pattern as the national estimates. Residents of low-income block groups experience better access but weaker demand, which combine overall into a lower use of branch services compared to residents of high-income block groups. Performing the decomposition by race, we find that residents of Black communities in big cities exhibit *higher* demand for branches compared to residents of White communities. But this higher demand is more than offset by weaker bank access, leading Black communities in big cities to use branches less. At the end of the article, we discuss several policy implications of our results.

Related Literature. This article relates to two major literatures. The first is the array of work that investigates financial access, financial use, and their joint relation to inequality. See Claessens (2006) and Claessens and Perotti (2007) for surveys. Much of this research has examined differences in access and use around the globe. Beck, Demirguc-Kunt, and Peria (2007) develop indicators of banking sector

outreach across 98 countries (e.g., the number of ATMs per capita, or the number of loans per capita). Beck, Demirgüç-Kunt, and Martinez Peria (2008) measure bank access barriers (e.g., minimum account and loan balances, or account fees) across 62 countries. See also Washington (2006), Blank (2008), Ho and Ishii (2011), and Goodstein and Rhine (2017), who investigate how improvements in various measures of bank access affect bank use in the United States. Yogo, Whitten, and Cox (2022) examine U.S. bank and retirement account participation by geography, income, and race using tax records. Nguyen (2019) studies how branch closings affect local small business access to credit. An advantage of this article's access measure is that it encapsulates the equilibrium choices of individual consumers, rather than reflecting only survey responses or being constructed from supply-side factors alone, such as the local branch density or the availability of low cost accounts. And while the causal benefits of increasing access are well established, less is known about the extent to which prevailing inequities in access sustain disparities in bank use. Our analysis advances this body of work with a research approach that isolates the effects of access on the use of bank branches.

Second, the article relates to the large literature in spatial economics on commuting flows and the arrangement of economic activity. See Redding (2013) and Redding and Rossi-Hansberg (2017) for surveys. Much of this work has focused on either firm or household location decisions (e.g., Lucas and Rossi-Hansberg 2002; Allen and Arkolakis 2014; Ahlfeldt, Redding, Sturm, and Wolf 2015) or agglomeration effects (e.g., Dekle and Eaton 1999; Rosenthal and Strange 2004). Recently, many articles in this literature have taken advantage of mobile device data to answer economic questions. An early example is Chen and Rohla (2018), who examine how political partisanship affects time spent together during Thanksgiving dinner. Athey, Blei, Donnelly, Ruiz, and Schmidt (2018) study consumer choice of restaurant dining. Chen, Haggag, Pope, and Rohla (2019) look at racial disparities in vote waiting times. Athey, Ferguson, Gentzkow, and Schmidt (2021) develop a measure of segregation based on where people visit over the course of a day. See also Büchel, Ehrlich, Puga, and Viladecans-Marsal (2020); Miyauchi, Nakajima, and Redding (2021); Kreindler and Miyauchi (2022); and Atkin, Chen, and Popov (2022). Many researchers have also used mobile device data to explore topics related to the Covid-19 pandemic (e.g., Coven, Gupta, and Yao Forthcoming; Almagro, Coven, Gupta, and Orane-Hutchinson 2021; Goolsbee and Syverson 2021; Couture, Dingel, Green, Handbury, and Williams 2022; and Chen, Chevalier, and Long 2021). We complement this literature by using micro-level observations of travel behaviors to quantify spatial patterns and transportation costs in banking activity.

The econometric method we propose in the article also relates to estimation procedures used in the spatial economics literature. Novel methods to handle econometric issues when estimating gravity models have been proposed before. In trade, Eaton and Tamura (1994); Helpman, Melitz, and

Rubinstein (2008); and Westerlund and Wilhelmsson (2011) suggest different methods to account for zero trade 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 procedure. Extensions of this estimator can handle a large number of fixed effects (Larch, Wanner, Yotov, and Zylkin 2019). Dingel and Tintelnot (2021) introduce a quantitative spatial model and estimation procedure for "granular" environments, where there is a finite number of individuals whose idiosyncratic choices impact equilibrium outcomes. But when the data in a gravity model's estimation is subject to differential privacy and many fixed effects require estimation, a different econometric approach is warranted. The article provides a way to implement the Method of Simulated Moments when many fixed effects require identification. Even in areas outside spatial economics, differential privacy algorithms are masking more economic data sets over time, including the 2020 Census tables and American Community Survey microdata (Ruggles, Fitch, Magnuson, and Schroeder 2019), transactions data (Karger and Rajan 2020), and health records (Allen, Bavitz, Crosas, Gaboardi, Hay, Honaker, King, Korolova, Mironov, Phelan, Vadhan, and Wu 2020). The article's econometric method can be helpful in estimating models in other economic environments that rely on privacy-preserving datasets. Finally, the article's conceptual measure of local access, empirically estimated using mobile device data, can be applied to other markets where physical retail locations matter, such as in grocery store and healthcare markets.

Outline. The article proceeds as follows. Section 2 provides background information on bank use in the United States. Section 3 presents a spatial model that serves as a conceptual framework for our measure of bank access. Section 4 describes the mobile device data we use to estimate the access measure. Section 5 details the econometric method for estimation. Section 6 analyzes variation in bank access by geography and demographic characteristics of local areas. Section 7 decomposes racial and income disparities in branch use by differences in access versus demand. Section 8 discusses policy implications of the results. Section 9 concludes.

2 Background on Bank Use in America

Every two years in June, the FDIC's Survey of Household Use of Banking and Financial Services is administered alongside the U.S. Census Bureau's Current Population Survey, which covers a representative sample of U.S. households. The FDIC survey queries both banked and unbanked households, and the most recent survey collected data from almost 33,000 respondents.

Our access measure is based on bank branches, but has the growth of online and mobile banking spoiled branches as important indicators of financial access? In Online Appendix B, we present evidence from the survey suggesting not. Roughly 81% of respondents answered having visited a branch in the past 12 months, and just over 29.7% reported having visited a branch 10 or more times. Traveling to a branch is the primary (i.e., most common) method of accessing bank accounts for about 25% of respondents. Mobile banking is more frequently cited as a primary method of use (31.4%) for banked households, but even in this group of respondents, 81.2% stated visiting a branch over the past year and about 1 in 5 in that group visited ten or more times.

From the survey, we also observe significant demographic variation in branch use. Controlling for age and race in a multivariate linear probability regression, we find that respondents in the highest income bracket ($\$75,000+$) are roughly 22% more likely to reply visiting a branch in the previous year than respondents in the lowest income bracket ($< \$15,000$). A substantial Black-White gap in reported branch use is also present. Controlling for reported income and age, we find that Black respondents are 10% less likely to answer having visited a branch in the past year than White respondents. Probit regressions provide similar estimates.

Focusing on the primary method of bank account access, we find that low-income and Black households do not appear to make up their lesser branch use with greater use of alternative banking methods like online or mobile banking. Both low income and Black households indicate relying on mobile/online banking less and bank branches more as their most common access method. Controlling for age and race, we find that respondents in the lowest income bracket are roughly 31% less likely than those in the highest income bracket to say that mobile or online bank is their primary method to access their bank accounts. Similarly, controlling for income and age, we find that Black respondents are about 6.6% less likely than White respondents to call mobile or online banking their primary access method.

In addition to the FDIC Survey, the Survey of Consumer Finances (SCF) also suggests that branches remain subjectively important to households in their use of banking services. According to the 2019 SCF, the locations of branches is cited most frequently as the most important reason for choosing an institution for their main checking account (43% of respondents), which is the most cited reason by far. Despite advances in mobile and online banking, the proportion of respondents citing branch locations as their most important reason has remained about the same since 1989 (between 43% and 49%).

Overall, bank branches remain integral to the financial lives of most U.S. households, and from survey evidence, we observe disparities in their use by both race and income. We turn next to providing a conceptual framework for household use of bank branches, which informs our measure of bank

access.

3 A Conceptual Framework for Bank Access

A continuum of residents choose destinations to visit from their home Census block groups per time period. Each resident r lives in one block group $i \in G$. Bank branches are located across the country, and each branch is indexed by $j \in B_t$, where the set of branches can vary over time from store openings and closings.

In every period, a resident chooses which single bank branch to visit so as to maximize utility. Residents may also choose not to visit a branch, either remaining home or visiting another point-of-interest. We index this outside option choice by $j = 0$. The indirect utility of resident r living in home block group i and visiting branch j at time t is

$$U_{rjt} = \frac{z_{rjt}\Lambda_{jt}}{\delta_{ij}}. \tag{1}$$

The term Λ_{jt} is an index of all attributes of branch j that make it a destination for residents of any block group at time t (e.g., the branch having attractive deposit or loan rates, higher customer service quality, or a wider variety of products). The term z_{rjt} is an idiosyncratic, unobserved error that captures individual differences in residents' personal preferences for banking at branch j (e.g., favoring Chase over Wells Fargo, relishing the branch's proximity to the children's daycare, or appreciating the building's historic architecture). Finally, the term δ_{ij} is an iceberg traveling cost that is defined as

$$\delta_{ij} = d_{ij}^{\kappa}, \tag{2}$$

where d_{ij} is the distance between home block group i and branch j, and $\kappa > 1$ controls the scale of the traveling costs.

To derive mathematically convenient functional forms for the branch choice behavior of the population, we follow McFadden (1974), Eaton and Kortum (2002), and Ahlfeldt et al. (2015) by assuming that the idiosyncratic component of utility z_{rjt} is drawn from an independent Fréchet distribution:

$$F\left(z_{rjt}\right) = e^{-H_{jt}z_{rjt}^{-\varepsilon}}, \quad H_{jt} > 0,\ \varepsilon > 1. \tag{3}$$

The branch-specific parameter $H_{jt} > 0$ influences the mean of the distribution. A larger H_{jt} implies that a high utility draw for branch j is more likely among residents of any block group. The term $\varepsilon > 1$ governs the heterogeneity of idiosyncratic utility. A smaller ε implies that residents are more

heterogeneous in their preferences for branches.[2]

Substituting the expression for U_{rjt} into the distribution of idiosyncratic tastes in Eq. (3), one can observe that residents of block group i at time t are presented with a distribution of utility across branches, $G_{ijt}(u) = \Pr\left[U_{rjt} \leq u\right] = F\left(u\delta_{ij}/\Lambda_{jt}\right)$, or

$$G_{ijt}(u) = e^{-\left[H_{jt}\left(\frac{\Lambda_{jt}}{\delta_{ij}}\right)^{\varepsilon}\right]u^{-\varepsilon}}. \tag{4}$$

We normalize the value from the outside point-of-interest $H_{0t}\Lambda_{0t}^{\varepsilon}\delta_{i0}^{-\varepsilon} = 1$. Each resident chooses a location to visit that yields the maximum utility. Hence, the distribution of utility across all possible locations that a resident would actually visit is

$$G_{it}(u) = \Pi_{j=0}^{B_t} G_{ijt}(u). \tag{5}$$

Inserting Eq. (4) into Eq. (5), one obtains the utility distribution:

$$G_{it}(u) = e^{-(1+\Phi_{it})u^{-\varepsilon}}, \tag{6}$$

where the parameter Φ_{it} of block group i's utility distribution is

$$\Phi_{it} = \sum_{j \in B_t} H_{jt}\Lambda_{jt}^{\varepsilon} d_{ij}^{-\kappa\varepsilon}. \tag{7}$$

Eq. (7) is the theoretical foundation of our measure of bank access. It summarizes information about the set of branches available to residents of a block group per time period. Given its form, Φ_{it} can be interpreted as an attribute-adjusted branch index that is unique to each block group. Each branch in a block group's index is represented by the average idiosyncratic utility $\left(H_{jt}\right)$ that residents of the block group assign to the branch's attributes $\left(\Lambda_{jt}\right)$ and the branch's distance $\left(d_{ij}\right)$ away. The impact of distance is influenced by the scale of traveling costs (κ) and the substitutability between branches (ε). While all branches in the country are available to all residents to visit, the costs of traveling to those branches vary by block group. Local areas have higher access if branches are relatively closer, especially branches with better attributes.

The object Φ_{it} is conceptually related to what some in the economic geography and trade literature have described as an exporting country's "access" to the importing markets of other countries (e.g., Harris 1954; Head and Mayer 2004; Redding and Venables 2004; Hanson 2005; and De Sousa et al. 2012). Donaldson and Hornbeck (2016) use a similar measure when estimating how U.S. railroad

[2]The parameter ε plays a role like the elasticity of substitution between bank branches in a model where residents have CES preferences over bank services from all branches. A smaller ε is akin to branches being less substitutable, which implies that residents find it worthwhile to travel to a branch despite resistance imposed by the geographic barrier δ_{ij}.

expansion in the 19th century affected a county's "market access" to trade with other areas. In our environment, we treat Φ_{it} as a local measure of residents' access to banks.[3]

In addition to providing a measure of access, the utility distribution also generates a gravity equation in visits between home block groups and bank branches. The share π_{ijt} of residents living in block group i who visit branch j at time t is

$$\pi_{ijt} = \frac{H_{jt}\Lambda_{jt}^{\varepsilon} d_{ij}^{-\kappa\varepsilon}}{1 + \Phi_{it}}. \tag{8}$$

The visitor share depends on the characteristics of the branch $\left(\Lambda_{jt}\right)$, the average utility draw of the branch $\left(H_{jt}\right)$, and the "bilateral resistance" derived from the intervening transportation costs $\left(d_{ij}^{-\kappa\varepsilon}\right)$. Other things equal, a resident is more likely to visit a branch if it has superior attributes, delivers higher average idiosyncratic utility, or is less costly to reach. In the denominator, Φ_{it} plays the role of "multilateral resistance," which affects residents' visitation to *all possible* branches. The probability that residents of a block group in, say, Palo Alto, visit a nearby Chase branch depends not only on the benefits of the branch and the costs of getting there, but also on the benefits and costs of visiting all other available branches.[4]

Eq. (8) offers a conceptual interpretation of the empirical gravity equation we employ to model consumer travel flows from block groups to bank branches. Estimates from the gravity equation are then aggregated to the block-group level to obtain local measures of bank access. We turn next to describing the microdata we use in the estimation.

4 Data on Branch Visits

We measure consumer travel patterns using mobile device data from SafeGraph between January 2018 and December 2019. The data are monthly and include both branch locations and information about branch visitors. The SafeGraph data are benefited by elaborate algorithms the company has developed to determine whether a mobile device visits a particular destination and to pinpoint a mobile device's home origin. A visitor 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. Unfortunately,

[3] While the spatial model implies that residents visit bank branches directly from home, in practice, they might drop by a branch during lunchtime at work or in the middle of other errands. Unfortunately, the mobile device data we will use in the estimation does not reveal the exact travel paths of visitors. The access measure is therefore best thought of as capturing bank access around residents' home block groups.

[4] We obtain the gravity relation in Eq. (8) by evaluating:

$$\pi_{ijt} = \Pr\left[u_{ijt} \geq \max\left\{u_{ijt}\right\}; \forall j\right] = \int_0^{\infty} \Pi_s \left[G_{is}(u)\right] dG_{ijt}(u)\, du.$$

we do not know the demographic attributes of the mobile device owners, nor their home addresses, nor their duration spent at a branch, nor what they do at the branch. Nevertheless, SafeGraph does provide the home Census block groups of branch visitors and the associated number of visitors from each block group. Online Appendix C provides background information on the SafeGraph data and a detailed explanation of the way we construct our primary sample. Here, we give a summary.[5]

4.1 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.[6]

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

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

[6]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.)

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.[7]

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. Online Fig. B.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.

4.2 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. II 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. We account for SafeGraph's differential privacy methods by estimating the gravity equation using the Method of Simulated Moments. Appendix A details the full procedure, and Section 5 provides a summary.

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

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. To address potential sampling bias from missing around 40% of U.S. branches, in Section 4.3 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.[8] 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. As we discuss in more detail in Online Appendix D, we do not know the precise demographic attributes of an individual bank branch visitor, and instead, we assign attributes to visitors according to the demographic characteristics of their home Census block groups. The 2019 FDIC Survey reports smartphone ownership rates by household characteristics. Overall, 85.4% 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.

To assess how well the mobile device visitation patterns align with the representative survey evidence on branch visits, in Online Appendix B.3, we compare the share of households in the 2019 FDIC survey who report having visited a bank branch in the previous 12 months to the share of mobile devices in SafeGraph that visit branches in the same period. The comparison is by reported household income of the respondent and median household income of the mobile device's home block group. There is a strong resemblance between the two sources, as both reported branch visitor shares from the FDIC survey and observed branch visitor shares from the mobility data are increasing and concave

[8] Online Fig. B.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.

in household income.

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%).[9]

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 than there are people. Squire (2019) also discusses this feature of the SafeGraph panel, and he interprets 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. For robustness, we weight block-group level regressions in Sections 6 to 7 by 2019 5-year American Community Survey (ACS) block group population counts. Population weighting down-weights block groups with disproportionately high mobile devices relative to their Census populations, and up-weights block groups with disproportionately low mobile devices relative to their populations, so as to make the visitor sample more representative.

4.3 Descriptive Statistics

Online Table B.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.

Online Table B.2 compares demographic characteristics of residents living in the geographic areas covered by our core sample of 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

[9] Couture et al. (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.

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 and significant, but overall, the economic magnitudes of the differences are small relative to the mean values across areas.

5 Bank Access Estimation

We compute local bank access by estimating a gravity model of consumer flows using the mobile device data. But SafeGraph's differential privacy methods would bias any OLS estimation. In this section, we describe our alternative econometric method and discuss the gravity estimates that enter the measure of access.

5.1 Econometric Method

We express the gravity equation in number of visitors rather than shares of residents to account for the differential privacy methods that SafeGraph applies to the visitor counts. Let V^*_{ijt} be the true number of visitors from block group i to branch j in period t that SafeGraph observes, and let R_{it} denote the number of residents of home block group i at time t. The true number of visitors can be expressed as:

$$V^*_{ijt} = \pi_{ijt} R_{it}, \tag{9}$$

where π_{ijt} is the share of block group i's residents visiting branch j at time t. 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:

$$V^+_{ijt} = V^*_{ijt} + L_{ijt}. \tag{10}$$

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 $\times$

branch visitor count is present in the sample. The selection equation is

$$z_{ijt} = \begin{cases} 1 & \text{if } \lfloor V^{+}_{ijt} \rfloor \geq 2, \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

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

$$V_{ijt} = \max \left\{ 4, \lfloor V^{+}_{ijt} \rfloor \right\}, \tag{12}$$

To estimate the gravity equation, we use the Method of Simulated Moments (MSM), relying on Eqs. (9) to (12). The MSM chooses model parameters to make the simulated model moments match the data moments as closely as possible. We run the estimation separately per year-month of the sample to account for branch openings and closings and to evaluate the stability of the estimates over time. A full description of our procedure is provided in Appendix A, but we provide a brief summary here.[10]

Specify the data generating process. In the simulations, we specify V^{*}_{ijt} as Poisson distributed with mean $\pi_{ijt} R_{it}$. We presume a Poission model for the count data rather than an alternative distribution such as Negative binomial because interpretations of the fixed effects and gravity coefficient are straightforward and because a Poisson model has been ubiquitous to estimation of gravity models since Silva and Tenreyro (2006).

Using the gravity equation in Eq. (8) from the conceptual framework to inform π_{ijt}, we express the true visitor count as following

$$V^{*}_{ijt} \sim \text{Pois}\left(\exp\left(\gamma_{it} + \lambda_{jt} - \beta \log \text{Distance}_{ij} \right) \right). \tag{13}$$

The term γ_{it} is a block group $\times$ year-month fixed effect that captures all characteristics of block group i that contribute to residents visiting any branch in year-month t. Informally, it represents factors that influence a block group's "demand" for branch services at any location (e.g., average wealth, income, financial sophistication, trust in banks, flexibility in time). The term λ_{jt} is a branch $\times$ year-month fixed effect that captures all characteristics of branch j that make it a destination for residents of any block group in year-month t. 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).[11]

[10] For textbook treatments of MSM, see Adda and Cooper (2003), Davidson and MacKinnon (2004), and Evans (2018).

[11] Using origin and destination fixed effects to estimate gravity equations has become standard practice in the trade literature since Harrigan (1996). See also Fally (2015) for a connection between fixed-effects and structural estimation of gravity models. Because we have a panel, the cross-sectional fixed effects are time-varying.

The parameter β is the elasticity of visitor flows with respect to distance. From the perspective of the conceptual framework, β can be interpreted as the product of residents' traveling cost scale (κ) and their elasticity of substitution between branches (ε). Distance is measured using the haversine formula, which accounts for the curvature of the Earth, and it is computed between branches and block groups' centers of population (see Footnote 18).

Eq. (13) imposes a log-linear relation between visitor counts and distance. Fig. I suggests that this structure represents the data fairly well. Panel A presents a binned scatter plot of the log number of visitors from block groups to visited branches by the log distance between the origin and destination, with the fixed effects removed. A clear negative relation between distance and visitation is visible, and that relation is nearly linear. The relation does flatten out when the log number of visitors approaches 1.4, but that change corresponds to SafeGraph's censored value of 4 visitors. When only block group × branch pairs exhibiting greater than 4 visitor counts are plotted in Panel B, the linear relation retains throughout.[12]

While the observed branch visits reasonably follow a gravity relation in distance, we do not claim that the relation is causal. It is true that the structural model of Eq. (13) implicitly assumes that the distances between branches and block groups is exogenous with respect to the parameters $\left\{\gamma_{it}, \lambda_{jt}, \beta\right\}$. This exogeneity assumption implies that knowledge of the distribution of distances is not required for inference of these parameters, and hence, we can validly condition the distribution of visitor counts on distance. But the distribution of distances is really the distribution of where people choose to live within block groups and where banks choose to build their branches. We treat these choices as exogenous not because we believe they satisfy this classification in reality, but because the additional joint modeling of both residential choices and branch location choices is outside the scope of our investigation. We tackle the much narrower goal of explaining the discrete choices of branch visitors, taking as given their home locations and the branch locations in each year-month. Finally, Eq. (13) might omit other block group × branch variables that influence branch visits (e.g., block-group targeted advertising or special block-group rates based on soft information). The extent to which these omitted variables have no predictive value for V^*_{ijt} after conditioning on distance, the parameter estimates will be consistent.

Sample the set of branches. Technically speaking, every possible block group i and branch j pair should enter Eq. (13). But our sample of over fifty-thousand branches, over two-hundred-thousand

[12]One might wonder whether driving time is a better measure of travel than distance. Online Table B.8 regresses driving times from block groups to branches over a random sample of around 1 million block group × branch pairs in our data. Over the full random sample and parts of the sample associated with different demographic attributes, such as including only block groups with Black population shares exceeding 80%, the R^2 values are very high, ranging from 0.972 to 0.993. Haversine distance is computationally easier to calculate, and these results suggest that it correlates highly with driving time.

block groups, altogether spanning twenty-four months, makes it computationally difficult to have all possible block group × branch pairs enter the gravity model estimation. Instead, we sample branches per block group using a technique that is similar in spirit to the procedure described in McFadden (1977) and Davis, Dingel, Monras, and Morales (2019). For each block group, we include the set of branches actually visited by the block group's residents in the year-month plus a random subset of all other *alternative* branches that could have been visited in the period.[13]

As we describe in more detail in Appendix A.2, we select the alternative branches using stratified sampling. We sample in the following manner: In each year-month, alternative branches for each block group are ranked by their distances away from the block group. We divide the number of alternative branches by 2,000 and round that value up. We then divide the distribution of ranked branches into these many quantiles and select one branch from each quantile with equal probability. For example, in a year-month in which a block group exhibits 38,020 alternative branches, the number of quantiles (and the number of sampled alternative branches) would be 20. The process implies that, on average, the random alternative set of branches per block group represents slightly higher than a 0.05% sample size of all possible alternative branches that residents could have visited. Because we use stratified sampling, we must apply frequency 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. Following standard practice, we use frequency weights that are the reciprocal of the likelihood of being sampled (subject to rounding). A larger sample size of 0.1% did not alter the estimation results.

Rather than using a stratified sampling of branches, we could have just as reasonably sampled branches uniformly in the style of McFadden (1977) and Davis et al. (2019). But stratifying by distance ensures that both near and faraway branches have equal chance of being selected, which affords the gravity model a more complete view of visitation rates by distance. Yet an alternative approach would be to restrict the set of branches per block group by a distance-based cutoff (e.g., including in residents' choice sets only the branches within a 10-mile radius of their block group). But doing so would contaminate the comparison of block group fixed effects and bank access between areas. For example, if a block group had twice as many branches within a 10-mile radius than another block group, the former's access measure would be twice as high as the latter, all else equal. But actual branch visitation might be similar between the two block groups. (Just because branch supply doubles does not mean visits double.) If so, the estimated fixed effect of the block group with more branches

[13] We say "similar in spirit" because, while McFadden (1977) and Davis et al. (2019) implement either uniform or stratified sampling of the set of available alternatives, consumers' choice processes in those articles obey the multinomial logit model, unlike here.

would mechanically have to lower to equalize visitation. It is important that the sets of branches available to each block group are sampled in the same manner. Finally, including all branches in the choice sets of residents allows us to uniquely pin down the branch fixed effects using the SafeGraph data, and we discuss how momentarily.

Simulate the visitor counts. We simulate visitor counts according to Eq. (13) and apply the differential privacy methods represented in Eqs. (10) to (12). The simulation process differs between the (i, j) pairs with positive observed visitor counts and those with 0 observed visitor counts from the stratified sampling. For the pairs with positive visitor counts, we draw Poisson random variables with distinct means given in Eq. (13). We then (i) add a Laplace draw to all non-zero true visitor counts to form a "noisy" block group $\times$ branch visitor count, (ii) round each noisy visitor count down to the nearest integer, (iii) make 0 all noisy visitor counts below 2, and (iv) replace all noisy visitor counts that equal 2 or 3 with 4.

For the (i, j) pairs arising from the stratified sampling, we do not draw random variables in the simulation. If these visitor counts were randomly drawn, they would have disproportionate impact on any computed moments because of the high frequency weights that would multiply them. Noise from the simulation would be amplified and make the estimation unstable. In other words, there is a peso problem. Rather than randomly drawing these pairs of visitor counts, we construct their implied empirical probability distribution given the parameter estimates. If an infinite number of observations were in fact simulated, their distribution would coincide with this constructed empirical distribution.[14]

That empirical distribution is a truncated and censored version of a Poisson probability mass function per Eq. (13) plus a Laplace distribution with mean 0 and scale $10/9$. Because the Laplace noise is added after the Poisson draw, this empirical distribution can be thought of as a truncated and censored Laplace distribution whose parameter location (mean) is the realization of the Poission draw. We construct 7 components of the empirical distribution: (i)-(ii) the probability that the visitor count equals 0, as well as exceeds 0; (iii)-(iv) the probability that the visitor count equals 4, as well as exceeds 4; (v) the expected visitor count; and (vi)-(vii) the expected natural logarithm of visitor counts, conditional on visitor counts exceeding 0, as well as exceeding 4. For more details, see Appendix A.4.

[14]Notice that we cannot apply this approach to the set of block group $\times$ branch pairs with positive visitor counts 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 the pairs with positive observed visitor counts, we simulate draws. However, the pairs in the set from the stratified sample are meant to represent the remaining population of (i, j) pairs, which are very high in number. One stratified sampled observation is meant to represent 2,000 observations from the same distribution. We need only construct the empirical distribution that these sampled pairs represent.

Iterate the fixed effects until convergence. The MSM uses the visitor data and the model parameters 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. (13) requires thousands of fixed effects be estimated. There are simply too many parameters to identify from the MSM minimization problem alone. Instead, we adopt an iterative routine to identify the fixed effects $\{\gamma_{it}, \lambda_{jt}\}$ and let the minimization problem identify β. First, given an estimate of β and the block group fixed effects, $\{\gamma_{it}\}$, we take advantage of another data field in SafeGraph: the total count of visitors to a branch, which is unaffected by SafeGraph's differential privacy methods. Because we presume that visitors can arrive from any block group, we can uniquely identify each branch's fixed effect from an "adding up" condition. Specifically, per branch, we sum the gravity equation across block groups, set that sum equal to the branch's total visitor count, and then invert the equation to extract the branch's fixed effect. Then, given an estimate of β and the branch fixed effects from the inversions, we repeatedly update the block group fixed effects, $\{\gamma_{it}\}$, until the differences in the average simulated visitor counts and average observed visitor counts of each block group i across all branches per year-month t become sufficiently small. When the number of fixed effects in each dimension is large, as in our setting, the routine produces consistent estimates. See Appendix A.5 for more details.

Construct the MSM estimator. After both dimensions of fixed effects are identified per estimate of β, the MSM minimization problem then selects the optimal β estimate that minimizes the weighted sum of squared errors (expressed in percentage points) between the simulated model moments and data moments. We use 6 unconditional moments that describe important parts of the distribution of visitor counts: (i)-(ii) the percent of visitor counts equaling 0 and equaling 4; (iii)-(iv) the average log distance, when visitor counts equal 0 and equal 4; and (v)-(vi) the OLS coefficient from regressing log visitor counts onto their associated log distances, when visitor counts equal 0 and equal 4. The model moments include both the simulated draws from the (i, j) pairs with positive observed visitor counts and components of the empirical distribution that represent the stratified sampled pairs. See Appendix A.6 for more details.

5.2 Gravity Estimates

Fig. II contrasts the distributions of observed "raw" visitor counts to simulated "true" visitor counts according to the month-by-month MSM estimation of the Poisson model in Eq. (13). The simulated visitor counts include all positive draws from all simulations across every year-month in the sample period. Among observed visitor counts across periods, 74.5% equal 4 due to SafeGraph's truncation

and censoring, whereas among simulated "true" visitor counts, 72.5% equal 4 or less. The MSM does a reasonable job spreading out the mass of visitors into the left tail of the distribution that is lost in the observed data. The two distributions also line up fairly well at values exceeding 4. For instance, around 4.3% of observed visitor counts equal 15 or more, whereas 5.1% of simulated visitor counts do. Also displayed in the figure is the distribution of simulated "manipulated" visitor counts, which are the "true" visitor counts after being manipulated by the differential privacy methods in Eqs. (10) to (12). The manipulated visitor counts line up fairly well with the observed visitor counts.[15]

Fig. III compares the observed number of visitors from each Census block group to their expected counterparts from the simulation. It presents a 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 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 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 data well in regions least affected by the differential privacy distortions, which one would hope for.

Fig. IV, Panel A presents the MSM estimates from the estimation, along with 95% confidence intervals. The monthly point estimates of the gravity coefficient range from about -1.25 to -1.45, and they are quite stable month-to-month. Thus, across the country, if a representative branch is located 1% farther away from a representative block group, the number of residents from that block group who travel to that branch will drop by around 1.25-1.45% 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. IV, Panels B and C present histograms of the estimated Census block group and and bank

[15] In the estimation, the sum of the observed visitor counts (in black) and simulated "manipulated" visitor counts (in red) match closely, but that may not seem to be the case from the figure. There, the simulated "manipulated" visitor counts tend to be smaller than the observed visitor counts. The reason behind this discrepancy is that the figure does not display the very high number of 0 visitor counts. The simulated "manipulated" visitor counts equaling 0 exceeds enough the observed counts equaling 0 to equalize the two sums.

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 distance. 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 biases any OLS estimation of the gravity equation and prompts an alternative econometric method like the MSM. But computing the OLS estimates is still useful to informally assess the magnitude of the bias. To this end, Online Table B.7 presents estimates from OLS regressions of a fixed-effects gravity equation using the observed visitor counts. We include a specification in which racial shares are also interacted with distance as independent variables. We run two sets of regressions: those including all block group $\times$ branch pairs and those that limit pairs with more than 4 branch visitors (which avoid SafeGraph's censoring).

When all block group $\times$ branch pairs are included, the OLS estimate $-\hat{\beta}_{\text{OLS}} = -0.053$, which is roughly one-third of the MSM estimate. When racial shares are interacted with distance, $-\hat{\beta}_{\text{OLS}}$ remains largely unchanged at -0.056. The coefficients on the racial interaction terms are small, which suggests that the relation between distance and visitor flows is universal across racial groups. But the coefficient on the interaction term with the Black share is precisely estimated and positive, which is consistent with residents from block groups with high Black population shares having a lower elasticity of substitution across branches.

When the sample is limited to block group $\times$ branch pairs with greater than 4 branch visitors, the estimate $-\hat{\beta}_{\text{OLS}}$ nationwide increases to -0.283, which still falls short of the MSM estimate. When racial shares are added as interaction terms, the OLS estimate of $-\hat{\beta}_{\text{OLS}} = -0.258$. The sign on the interaction term with the Black population share is still positive, but no longer precisely estimated. Overall, Online Table B.7 reveals the downward bias that SafeGraph's differential privacy methods introduce to an OLS estimation of the gravity equation, and it stresses the need for the alternative MSM procedure.

6 Bank Access in the United States

We next examine variation in bank access throughout the United States. Substituting the gravity equation estimates into the theoretical measure of access from Eq. (7), we obtain an empirical measure of bank access for residents of block group i in year-month t:

$$\hat{\Phi}_{it} \equiv \sum_{j \in B_t} \exp\left(\hat{\lambda}_{jt}\right) d_{ij}^{-\hat{\beta}}, \tag{14}$$

where B_t is the entire set of branches in our core sample in the year-month. Because all block group and branch pairs are included in the estimation from the stratified sampling, every branch has an estimated fixed effect, and thus, each home block group's measure of access includes all branches nationwide in the period.

The relation between bank access and branch visitor counts from the Poisson model of Eq. (13) bestows economic meaning to the magnitude of $\hat{\Phi}_{it}$. Because the access measure aggregates across all branches, it can be interpreted as a block group's expected total number of branch goers per month, when setting the block group's fixed effect to zero. The block group fixed effects control for block group populations, so residents of high-population areas do not mechanically have better access. We characterize bank access by evaluating its spatial heterogeneity over different parts of the country and by measuring its association with demographic characteristics of block group residents.

6.1 The Geography of Bank Access

Fig. V illustrates a dot density map of bank access by Census block groups cross-country. Each dot is positioned at a block group's center of population. We compute the access estimates month-by-month per block group, and the figure presents weighted monthly averages, where each month's weight is its share of the block group's total observed branch visitors over the sample period. We construct the map by grouping block groups into deciles and shading the dots so that higher-ordered colors in the rainbow gradient (indigo and violet) imply higher bank access and lower-ordered colors (red and orange) imply lower access. Block groups where no resident was recorded in SafeGraph as having visited a branch in the sample period are shaded white.

Across the country, three patterns in bank access are apparent. First, substantial variation in access is observed across regions. The eastern shores of New England and the Mid Atlantic, along with the upper Midwest experience the highest access nationwide. The Appalachian states, such as parts of West Virginia, Tennessee, and Kentucky, observe relatively better access than the Deep South, such as the southern parts of Mississippi and Alabama. A population weighted block-group-level regression

of bank access on Census region fixed effects suggests that 22.5% of the variation in access is within regions.

Second, the most pronounced differences in access are between urban and rural areas. Big cities like Boston, Richmond, Miami, Philadelphia, Houston, and Minneapolis observe substantially higher bank access than even nearby rural areas. Suburban areas that outlie large metropolitan centers also observe greater access than rural parts of the country. Online Fig. B.2 compares bank access by Rural-Urban Commuting Areas (RUCAs), as classified by the U.S. Department of Agriculture's Economic Research Service. RUCA codes separate census tracts by their urban/rural status and their commuting relationships with other areas using Census measures of population density, levels of urbanization, and daily home-to-work commuting. We categorize each block group into its RUCA code, and we present in the figure the population-weighted averages of bank access per RUCA. Access declines monotonically as one transitions from Metropolitan core to Metropolitan suburb, micropolitan suburb, and rural areas. Bank access in the typical rural area is about two-thirds less than the access in a typical Metropolitan core.

Third, even within a local area, bank access varies significantly. A population weighted block-group-level regression of bank access on census tract fixed effects estimates that just 26.2% of cross block-group variance in access is within census tracts. Fig. VI zeroes in on the four largest cities in the U.S. by population: New York City, Los Angeles, Chicago, and Houston. Around New York City, access is highest for residents of Midtown Manhattan and lowest for those living in the Bronx, and within each borough there is large variation. But relatively speaking, access is high overall for residents of New York City, such that the numbers in the city's lowest access neighborhoods are comparable to those in the highest access neighborhoods of Los Angeles. In Greater Los Angeles, residents of Beverly Hills and the Hollywood area observe substantially better access than residents living south of the city, such as in Compton and Long Beach. Residents of neighborhoods in the Palos Verdes Peninsula, San Fernando, and near the Anaheim Hills observe the lowest access. Residents of the Palos Verdes Peninsula experience weaker access despite the neighborhood being relatively affluent. Around Chicago, residents of the North side experience better access than those living in the South side, an area with a high Black population share. Residents of the Northwest suburbs near Lake Forest observe the highest access outside the Loop, which also corresponds to neighborhoods with high White population shares. Finally, in Houston, access is highest for residents living downtown, and access declines more and more, almost uniformly, as one moves farther away from the city into such areas as Sugar Land, Brookside Village, and Humble. The radial decline in access moving away from the downtown area mirrors the general pattern observed nationally in the neighborhoods surrounding

big cities.[16]

The heterogeneity in bank access displayed in these four big cities hint that access is correlated with the income and racial composition of local areas. We turn next to formally evaluating how access varies with those attributes.

6.2 Bank Access by Demographic Attributes

Because a central focus of our study is answering how bank access and use differs between Black and White households, we look at national comparisons, but also zero in on parts of the country that present Black population shares close to the national average. Online Table B.3 provides household counts and Black shares throughout the U.S. and within each RUCA. The national Black share is 12%. The commuting areas with Black population shares closest to this national number are Metropolitan area core (Metro core), having a 15% Black share, and Micropolitan area core (micro core), having a 9% Black share. Although Metro and micro cores share similar racial shares, Metro cores vastly outnumber micro cores in household counts (99.5 million vs. 8.5 million), and Metro cores capture roughly 72% of the 138.9 million total households in the U.S. For this reason, we supplement our national estimates of bank access with local estimates in Metro cores.

Table I presents weighted OLS regressions of $\log \hat{\Phi}_{it}$ on demographic attributes of block group residents. Observations in the regressions are at the level of a home Census block group per year-month over the core sample period, they are weighted by 2019 5-year ACS block-group population counts, and standard errors are clustered at the block-group level. Independent variables are population-based shares from the 2019 5-year ACS. The five racial/ethnic groups used in the regressions are non-Hispanic Asian, non-Hispanic Black, non-Hispanic White, non-Hispanic Other Races, and Hispanic. In all specifications, we control for the log number of mobile devices residing in the block group, which is unaffected by SafeGraph's differential privacy methods. The coefficients from the table can be interpreted as the percent change in a block group's expected number of branch goers per month, assuming the block group fixed effect is held constant at zero.

Our access measure embodies information about the per-mile cost of travel through the elasticity parameter β, which we take to be constant nationwide for simplicity. But commuting costs indeed vary substantially throughout the diverse U.S. landscape. A mile in downtown Chicago is much

[16] A popular measure of bank access is an area's density of branches. In Online Table B.9, we regress our access measure on branch density at both the county and census tract levels. At these geographies, between 44-62% of the variation in access is explained by density when state fixed effects are included, suggesting that our measure of bank access captures more information than just supply-side factors. In addition, the coefficient in the regression changes sign from positive to negative as the landscape enlarges from census tracts to counties, which reveals the importance of measuring access over as narrow a terrain as possible.

costlier to traverse in a car, train, or bus than a mile in the surrounding Cook County suburbs. Because transportation costs might differ even within counties, county fixed effects are insufficient as controls. To control for variation in traveling times and make the notion of "distance" as comparable as possible across different types of areas (urban, rural, and suburban), we add RUCA fixed effects to our specifications in the table.

Column (1) conditions the bank access regressions on median household income and race. Residents of block groups with higher median household income experience lower bank access. A doubling in a block group's median household income is associated with its residents experiencing about 11.0% weaker access. Residents of block groups with high Black population shares also experience weaker access. Extrapolating from the coefficients on the racial shares, we estimate that a hypothetical block group with a 100% Black population share would observe roughly 8.2% fewer branch goers per month than a comparable block group with a 100% White population share, holding constant the block group fixed effect at zero.

An important factor that might drive branch visits are differences in financial savvy or technical sophistication from differences in age or education (Caskey and Peterson 1994; Caskey 1994; Hogarth and O'Donnell 1997; Hogarth, Anguelov, and Lee 2005; Blank and Barr 2009; Rhine and Greene 2013). Including income in column (1) already proxies for the permanent component of human capital, but in column (2), we add age shares. Controlling for age in column (2) still preserves the negative relation between income and access, though the magnitude is cut to -7.6%. Also controlling for age, we observe block groups with large Black population shares experiencing weaker access. Extrapolation of the coefficient implies that a hypothetical block group with a 100% Black population share observes about 5.3% poorer access than a comparable block group with a 100% White population share.

Column (3) restricts the sample to block groups in Metropolitan core areas. There, the negative coefficients on income and the Black share are sharper. In big cities, a doubling of a block group's median household income is associated with its residents observing about 12.6% weaker access, and the Black-White gap in access is 10.7%. Controlling for age shares in column (4), we find that the coefficients on income and the Black population share remain negative, though smaller in magnitude (-8.7% and -6.4%, respectively).

6.3 Explaining Access: Branch Proximity and Branch Quality

The measure of bank access combines information about the "quality" of branches available to residents and the costs of traveling to those branches. We next evaluate how each component

contributes to differential access across block groups. Bank access in Eq. (14) can be rewritten as

$$\hat{\Phi}_{it} \equiv \underbrace{\left(\sum_{j \in B_t} d_{ij}^{-\hat{\beta}}\right)}_{\text{Branch Proximity}} \times \underbrace{\left(\sum_{j \in B_t} \pi_{ij} \exp\left(\hat{\lambda}_{jt}\right)\right)}_{\text{Branch Quality}}, \tag{15}$$

where the weights $\pi_{ij} \equiv \frac{d_{ij}^{-\hat{\beta}}}{\sum_{j \in B_t} d_{ij}^{-\hat{\beta}}}$. Eq. (15) separates access into two components. The first is branch proximity, measured as the sum of the inverse of the transportation costs that residents would bear to reach bank branches across the country. It can be interpreted as residents' hypothetical bank access if branch quality were equalized across all branches (and normalized to one). The second component is branch quality, measured as the weighted average quality of branches that residents experience, where closer branches are assigned higher weight. It can be interpreted as residents' hypothetical access if their total transportation costs were normalized to one. (Notice that this normalization does not *equalize* transportation costs across block groups; if costs were equalized, average branch quality would be the same across block groups.)

Columns (5)-(8) of Table I present the coefficients of weighted OLS regressions of these two access components on block-group median household income, racial shares, and age shares. Columns (5) and (7) provide nationwide estimates, whereas columns (6) and (8) focus on Metro cores. Nationwide, a doubling in a block group's median household income is associated with a 7.6% decline in its residents' proximity to all bank branches. In Metro cores, the drop in proximity is higher at 8.6%. Looking at differences in branch quality in columns (7)-(8), we find no statistical difference in the average quality of branches residents experience in low-income versus high-income communities, both nationwide and in big cities. Put together, the decomposition reveals that the higher bank access for residents of low-income block groups is entirely driven by their greater proximity to bank branches and not from experiencing a higher quality of branches.

Focusing on the racial differences in the two access components, we find that residents of block groups with high Black population shares experience significantly lower branch proximity. Nationwide, residents of a block group with a 100% Black share observe 14.3% lower branch proximity compared to residents of a block group with a 100% White share. In Metro cores, the drop in proximity to branches for Black communities is 16.8%. However, both nationwide and in Metro cores, residents of block groups with high Black population shares experience better average branch quality (9.0% higher quality cross-country and 10.4% higher quality in big cities). This result implies that reduced proximity to branches, rather than lower quality branches, explains the lower bank access in Black communities. Put differently, the branches nearest Black communities tend to be of higher average quality, but those

branches are located relatively farther away from those communities and hence, are more costly to reach.[17]

Overall, results from Sections 6.2 to 6.3 show that (1) across the country and in Metro cores, residents of low-income block groups observe higher bank access, and we find no meaningful difference in their average quality of branches compared to high-income communities; and (2) residents of areas with high Black population shares experience weaker bank access, and this weaker access is driven by bank branches being located relatively farther away from Black communities, rather than these communities experiencing lower average quality branches.

7 Bank Branch Use: Access versus Demand

Our gravity model takes advantage of the network of consumer flows linking block-group residents to multiple branches. By using block group fixed effects, it isolates how the unequal distribution of bank branches explains disparities in household use of those branches, by controlling for block-group differences in demand for branch services. Since the comparison is across branches for the *same* block group, block-group specific demand is absorbed by the block group fixed effect.

As an added benefit, the gravity model in Eq. (13) generates a simple decomposition of branch visitor counts into two parts. From that equation, the natural logarithm of the expected total number of residents of block group i who visit any branch in year-month t is

$$\log \hat{V}^*_{it} = \log \hat{\Phi}_{it} + \hat{\gamma}_{it}. \tag{16}$$

Eq. (16) separates a block group's total count of branch goers into (1) the count if branch demand were equalized across block groups, plus (2) the count if access instead were equalized. This decomposition thus separates the part of branch use that is due to access versus demand.

We can also evaluate the extent to which *variation* in branch use across population groups can be attributed to differences in each component. An OLS projection of $\log \hat{V}^*_{it}$ onto a vector X_i of resident characteristics at the block group level gives

$$\log \hat{V}^*_{it} = X_i \theta_V + \varepsilon_{V,it}. \tag{17}$$

Similar projections of both the access measure and the estimated block-group × year-month fixed

[17] If the coefficients on the Black population share in columns (7) and (8) registered the opposite sign, bank access would have been doubly bad for Black communities. Not only would branches have been located farther away from them, but their nearest branches would be of lower quality.

effects onto X_i produce

$$\log \hat{\Phi}_{it} = X_i \theta_{\Phi} + \varepsilon_{\Phi,it}, \tag{18}$$

$$\hat{\gamma}_{it} = X_i \theta_{\gamma} + \varepsilon_{\gamma,it}. \tag{19}$$

Along any demographic attribute x, the estimated coefficients from Eqs. (17) to (18) satisfy the identity:

$$\hat{\theta}_{V,x} \equiv \hat{\theta}_{\Phi,x} + \hat{\theta}_{\gamma,x}, \tag{20}$$

which separates the elasticity of branch use with respect to the demographic attribute into two constituent parts: the elasticity of bank access, $\hat{\theta}_{\Phi,x}$, and the elasticity of demand for branch services, $\hat{\theta}_{\gamma,x}$.

We begin the empirical analysis of this section by running regressions of the sort in Eq. (17) and Eq. (19). Regressions of the sort in Eq. (18) were presented earlier in Section 6.2. We then use the estimates to decompose disparities in branch use into parts explained by differential bank access and demand for branch services.

7.1 Bank Branch Use

Table II presents weighted OLS regressions of bank branch visitors by demographic attributes, both nationwide and in Metro cores. Like in the access regressions of Table I, observations are at the level of a home Census block group per year-month over the sample period, they are weighted by 2019 5-year ACS block-group population counts, and standard errors are clustered at the block-group level.

In columns (1)-(4), the dependent variable is $\log \hat{V}^*_{it} \equiv \log \sum_j \hat{V}^*_{ijt}$, where $\hat{V}^*_{ijt}$ is the predicted mean of V^*_{ijt} in Eq. (13) and β varies by month. Column (1) reports coefficients on median household income and racial shares, with year-month, county, and RUCA fixed effects. High income block groups have more expected bank branch goers per month (about 18.6% more for every doubling in the block group's median household income). Block groups with high Black population shares have fewer expected branch goers compared to block groups with high White population shares. A hypothetical block group with a 100% Black share would have roughly 4.1% fewer branch goers per month relative to a comparable block group with a 100% White share. Controlling for block group age shares in column (2), we find that the coefficient on income drops slightly (0.155), but the coefficient on the Black share is more sharply negative (-0.056). Extrapolations of the coefficient suggest that a block group with a 100% Black population share would expect 5.6% fewer branch visitors per month relative to a comparable block group with a 100% White share.

Moving away from nationwide estimates, we next focus on Metro core areas in columns (3) and (4). Having controls for age shares in column (4), we find that the positive income gradient in branch use in big cities remains the same as the cross-country estimate (0.155). The magnitude of the coefficient on the Black population share is -0.039, which is slightly narrower than the national estimate. Thus, in big cities, we detect a Black-White gap in branch use of roughly 3.9% per month.

7.2 Bank Branch Demand

In columns (5)-(8) of Table II, we report coefficients from weighted OLS regressions of the estimated block group fixed effects, $\hat{\gamma}_{it}$, from Eq. (13) on demographic attributes of block group residents, both cross-country and in Metro cores. In all columns, lower-income block groups observe lower fixed effects, which implies that residents of these areas have a lower propensity to visit any bank branch. A 1% decline in a block group's median household income is associated with a roughly 0.23-0.32% drop in its residents' "demand" for branch services. If the block group's access were held fixed and normalized to one, this range would correspond with the decline in the block group's number of branch goers per month.

We cannot say with certainty why residents of low-income areas exhibit lower demand for bank branches. But evidence from the 2019 FDIC Survey provides some potential explanations. Among unbanked households, the top 5 reasons cited for not having a bank account are (i) not having enough money to meet minimum balance requirements (48.9% of respondents), (ii) not trusting banks (36.3%), (iii) avoiding a bank gives more privacy (36.0%), (iv) bank account fees are too high (34.2%), and (v) fees are too unpredictable (31.3%). The extent to which these reasons correlate with a respondents' income, they can be explanations for the lower demand among residents of low-income block groups. But our sample also includes banked, and quite likely, underbanked residents, the latter of whom have bank accounts but still rely on alternative financial services like payday loans. Low-income banked and underbanked residents might exhibit lower demand for banks for the same reasons as unbanked residents. But they might also have less demand for the kinds of premium services that require visiting a branch, such as storing valuables in a safety deposit box, using notary services, or consulting with a banker about more complex financial issues like wealth management or small business banking. Finally, nonbank financial institutions like check cashers might target their services and advertising to low income customers, shifting their demand away from banks.

Turning to racial differences in bank branch demand, we find that block groups with high Black population shares observe larger fixed effects than block groups with high White shares, with controls

for income alone in column (5). Extrapolations of the coefficient suggest that a block group with a 100% Black share would observe about 4.1% more branch visitors per month compared to a comparable block group with a 100% White share, if bank access were held constant and set to one. But when age controls are added, the sign of the coefficient on the Black share turns negative and is no longer precisely estimated, suggesting that demand for branch services nationwide between Black and White communities is not robustly different. In Metro cores, block groups with high Black population shares do observe precisely estimated higher fixed effects without age controls (8.9% higher demand in column 7), and the coefficient stays precisely estimated once age controls are added (2.5%).

Overall, results from Sections 7.1 to 7.2 show that (1) across the country and in Metro cores, residents of low-income block groups use branches less than residents of high-income block groups, but residents of these low-income block groups exhibit substantially lower demand for branches via lower block-group fixed effects, and (2) residents of areas with high Black population shares also use branches less cross-country and in big cities than comparable block groups with high White population shares. But residents of Black communities show no robust difference in their demand for branch services nationwide, and they exhibit slightly higher demand for branch services in big cities.

7.3 Decomposing Branch Use

Looking at different retail markets, researchers have studied how weaker access to vital products, such as grocery stores and hospitals, depresses use of those products and disrupts welfare (Yantzi, Rosenberg, Burke, and Harrison 2001; Inagami, Cohen, Finch, and Asch 2006; Nicholl, West, Goodacre, and Turner 2007). Furthermore, access to these products varies by socioeconomic status (Currie and Reagan 2003; Hamrick, Hopkins et al. 2012)). See also Allcott, Diamond, Dubé, Handbury, Rahkovsky, and Schnell (2019) for an analysis that separates out demand and supply explanations for the inequalities in nutrition observed by income. When it comes to banking markets, we can decompose the variation in branch use into parts due to differences in access versus demand. We do so by inserting the estimates from the bank access regressions in Table I and the estimates from the branch use and demand regressions in Table II into the identity of Eq. (20).

Nationwide, the income gradient in branch use from column (2) of Table II is 15.5%. This income gradient consists of a demand gradient of +23.0% from column (6) of the table and an access gradient of -7.5% in column (2) of Table I. The decomposition thus shows that the greater demand for branch services among residents of high-income block groups dominates their weaker access and leads to their higher overall branch use. When it comes to racial disparities, column (2) of Table II reveals a

Black-White gap in branch use of 5.6%. This gap consists of 0.3% lower demand and 5.3% lower access. But the difference in demand is imprecisely estimated, implying that the Black-White gap in branch use is entirely due to a Black-White gap in bank access rather than in demand.

In Metro cores, the income gradient in branch use (column 4 in Table II) is +15.5%, which can be separated into an income gradient in access of -8.7% (column 4 in Table I) and an income gradient in demand of +24.2% (column 8 in Table II). As was the case nationally, residents of high-income block groups experience weaker access but higher demand, with the latter eclipsing the former enough to induce higher overall branch use. As for racial differences in big cities, residents of areas with high Black population shares experience weaker access (an elasticity of -6.4%) compared to areas with high White population shares, and residents of Black communities exhibit slightly higher demand (an elasticity of +2.5%). Thus, in big cities, the lack of access in Black communities is powerful enough to upset the higher demand for bank branches and induce a 3.9% Black-White gap in branch use.

To summarize, we find that low-income communities observe better bank access, both nationally and in big cities, but their demand for branch services is lower compared to high-income communities, which induces a lower use of branches. For these low-income communities, a lack of bank access cannot explain their lower branch use. Black communities, on the other hand, experience weaker access compared to White communities both cross-country and in big cities. Black communities exhibit no robust difference nationwide in demand for branches compared to White communities and mildly stronger demand for branch services in big cities. But their relative lack of access is substantial enough to overcome their higher demand, leading to a Black-White gap in branch use in big cities as well.

8 Discussion of Policy Implications

Researchers and policymakers have proposed several programs to increase bank access, aiming to remedy disparities in bank participation (Dahl and Franke 2017; Davidson 2018). Example proposals are investing in community development banks, which are certified commercial banks that principally serve minority and low-income communities (Ellwood and Patel 2018); or expanding U.S. postal banking, which would add checking, savings, and possibly credit services to some or all U.S. Post Office branches (Baradaran 2013). Our results suggest that these policies would have their largest impact on bank access in Black communities, particularly those in big cities.

In Online Appendix E, we asses how expanding U.S. postal banking might affect both access to and use of bank branches. In September 2021, the U.S. Postal Service launched a pilot program in four post office locations that allowed customers to cash payroll and business checks in the form of

gift cards (Heckman 2022). Adding more financial services and expanding the program to all Post Office locations would require congressional legislation. To evaluate the potential expansion of the policy, we re-estimate each block group's bank access after adding all Post Office locations registered in SafeGraph to the set of available branches. We use the time-series estimates of β from before; we leave unchanged each private bank's estimated fixed effects, λ_{jt}; and we retain the same estimates of each block group fixed effects, γ_{it}. This kind of analysis, which ignores the general equilibrium effects under the policy change, is a "partial policy" evaluation of postal banking, akin to what the trade literature calls a "partial trade impact" of a policy change in trade costs, such as tariffs (Head and Mayer 2014).

Whereas private banks in the analysis retain their original estimated fixed effects, we consider three different scenarios of fixed effects for the USPS locations. In each scenario, all postal locations share the same fixed effect for simplicity. In the first scenario, Postal branches are "low quality," in that they all share the estimated fixed effect of the 10^{th} percentile of private banks per year-month. In the second scenario, Postal branches are "medium quality," having the estimated fixed effect of the 50^{th} percentile of private banks per year-month. In the third scenario, Postal branches are "high quality," having the estimated fixed effect of the 90^{th} percentile of private banks per year-month.

In the low- and medium-quality scenario, we find that an expansion of the postal banking system would *widen* the Black-White gap in access, both nationwide and in big cities. Extending banking services to Post Office locations would mechanically increase bank access for everyone, but it would enhance access relatively more for residents of predominately White areas than for residents of predominately Black areas. The racial gap in access widens because Post Offices also tend to be located comparatively closer to White communities than Black communities, just like private bank branches. Only in the case of a high quality postal banking system would the Black-White gap in access shrink. Nationwide, it would shrink from 5.3% to 4.8% per month, and in Metro cores, it would shrink from 6.4% to 5.3% per month. Because block group fixed effects retain their same values from before, any change in the racial gap in access would equal the relative amounts that branch use would change between groups. By raising access for everyone, postal banking would increase branch use in all groups, but the increase would be relatively smaller in Black communities than in White communities. The exception is a high quality system, which would shrink the Black-White gap in branch use by 0.5 percentage points cross-country and 1.1 percentage points in Metro cores.

Our partial policy evaluation ignores any changes to block group residents' demand for banking services that postal banking might bring about, which is a shortcoming of the analysis. Distaste for private banks might partly explain the low estimated block group fixed effects of some communities. In the 2019 Survey of Consumer Finances (SCF), the choice "do not like dealing with banks" is the

second-most cited reason for families not having a checking account, and the fraction of unbanked respondents selecting this option as their reason has increased steadily over time (from 15% in 1989 to 22.9% in 2019). Proponents of postal banking argue that the unbanked will perceive Post Office bank branches as more trustworthy than private banks, making the unbanked more likely to utilize post office locations (Office of the USPS Inspector General 2014; Baradaran 2015). If this is true, postal banking could improve bank participation by raising demand for bank branch services. Our main analysis suggests that low-income communities would benefit the most from this potential outcome of the policy, as residents of these areas observed the lowest estimated block group fixed effects.

An expansion of postal banking could also trigger a competitive reaction from private banks, leading to an endogenous change in their branch fixed effects, something we also do not account for. Private banks might develop different fee and service structures to attract unbanked and underbanked residents. If this were to happen, a postal banking policy could indirectly increase access and consumer welfare by inducing private banks to improve their branch quality in response to competitive pressures. But why do private banks not already cater to the unbanked and underbanked? We found that Black communities exhibit slightly higher demand for branch services in big cities, but they experience significantly weaker access there. Why do private banks not enter these areas? It is beyond the article's scope to answer this question, but it is important to stress that we measure branch demand from branch visitation, which might not perfectly correlate with bank profits. Private banks might not enter because doing so is not profitable. If that were the case, another policy that could remedy weak bank access would be allocating tax credits to banks that establish new branches in Black communities, or subsidizing community development banks that operate there.

The policy responses discussed so far relate to changes in the locations or quality of bank branches. But another policy that could improve bank access is enhancing broadband Internet connectivity to reduce the costs of mobile or online banking. The two groups we identify making the least use of bank branches (low-income and Black households), simultaneously rely on branches *more* as their primary method of banking, rather than mobile or online. Weak access to broadband could be a reason why. Black adults are about 9 percentage points less likely to have home broadband than White adults (71% vs. 80%), and low income families earning less than $30,000 are 13 percentage points less likely to have broadband at home than families earning more than $75,000 (86% vs. 99%) (Pew Research Center 2021). Programs to expand broadband connectivity in low-income and Black communities could encourage residents to make greater use of mobile or online banking as a substitute for visiting a physical branch.

Expanding broadband could also raise consumer access to Fintech firms, non-depository institutions offering financial services entirely online. A purported goal of Fintech is increasing the financial

inclusion of groups whom traditional banks have historically underserved. But evidence so far on Fintech's success at achieving that goal is mixed. Erel and Liebersohn (2022) find that Fintech lenders expanded credit from the Paycheck Protection Program to small business owners located in areas with more low-income and minority residents. But Fuster, Plosser, Schnabl, and Vickery (2019) find that Fintech borrowers of purchase mortgages tend to have higher incomes and are less likely to be minorities. More time is needed to observe the full potential of Fintech on closing disparities in financial access. But in every wave of the Survey of Consumer Finances since 1989, the location of a bank's branches is cited as the most important reason by far for choosing an institution for a main checking account (43.1% of respondents). This persistent stated preference suggests that programs targeting the geographic distribution of retail locations where consumers receive their financial services will remain important for public policy.

9 Conclusion

We use anonymous location data from mobile devices to develop a local measure of bank access. The measure is derived from a spatial gravity model and is an expression of a local area's distance from surrounding bank branches and branch quality. Both the model and the rich data of consumer travel patterns allows us to implement a research design that isolates the impact of access on branch use, controlling for potential differences in demand for branch services.

To overcome distortions in the mobile device data that safeguard user privacy, we estimate the access measure using the Method of Simulated Moments (MSM), and we introduce an econometric method that can handle hundreds of thousands of fixed effects in MSM routines. The method can be of use in other applications relying on big data that is potentially contaminated by differential privacy methods.

Controlling for block group racial and age shares, we find that access is better for residents of low-income communities, which implies that a lack of bank access cannot explain their lower use of bank branches. In contrast, residents of areas with high Black population shares experience weaker access, largely because branches are located farther away from them. The relative lack of bank access for Black communities in big cities is so significant that it eclipses their slightly higher demand for branch services. The consequence is their overall lower use of bank branches, despite Black residents reporting greater reliance on branches as their primary means of banking.

References

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

Agarwal, S., S. Alok, P. Ghosh, S. Ghosh, T. Piskorski, and A. Seru (2017): "Banking the unbanked: What do 255 million new bank accounts reveal about financial access?" Working paper. Columbia University, New York, NY.

Agarwal, S., J. B. Jensen, and F. Monte (2018): "The geography of consumption," 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, 2127–2189.

Allcott, H., R. Diamond, J.-P. Dubé, J. Handbury, I. Rahkovsky, and M. Schnell (2019): "Food deserts and the causes of nutritional inequality," *Quarterly Journal of Economics*, 134, 1793–1844.

Allen, J., C. Bavitz, M. Crosas, M. Gaboardi, M. Hay, J. Honaker, G. King, A. Korolova, I. Mironov, M. Phelan, S. Vadhan, and A. Wu (2020): "The OpenDP White Paper," Working Paper.

Allen, T. and C. Arkolakis (2014): "Trade and the Topography of the Spatial Economy," *Quarterly Journal of Economics*, 129, 1085–1140.

Almagro, M., J. Coven, A. Gupta, and A. Orane-Hutchinson (2021): "Racial disparities in frontline workers and housing crowding during COVID-19: Evidence from geolocation data," Working paper. New York University, New York, NY.

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, 277–297.

Athey, S., D. Blei, R. Donnelly, F. Ruiz, and T. Schmidt (2018): "Estimating heterogeneous consumer preferences for restaurants and travel time using mobile location data," in *AEA Papers and Proceedings*, vol. 108, 64–67.

Athey, S., B. Ferguson, M. Gentzkow, and T. Schmidt (2021): "Estimating experienced racial segregation in US cities using large-scale GPS data," *Proceedings of the National Academy of Sciences*, 118, 1–9.

Atkin, D., M. K. Chen, and A. Popov (2022): "The returns to face-to-face interactions: Knowledge spillovers in silicon valley," Working paper no. 30147. National Bureau of Economic Research, Cambridge, MA.

Baradaran, M. (2013): "It's time for postal banking," *Harvard Law Review Forum*, 127, 165–175.

——— (2015): *How the other half banks: Exclusion, exploitation, and the threat to democracy*, Harvard University Press.

Beck, T., A. Demirgüç-Kunt, and M. S. Martinez Peria (2008): "Banking services for everyone? Barriers to bank access and use around the world," *World Bank Economic Review*, 22, 397–430.

Beck, T., A. Demirguc-Kunt, and M. S. M. Peria (2007): "Reaching out: Access to and use of banking services across countries," *Journal of Financial Economics*, 85, 234–266.

Blank, R. M. (2008): "Access to financial services, savings, and assets among the poor," *National Poverty Center Policy Brief (13)*.

Blank, R. M. and M. S. Barr (2009): *Insufficient funds: Savings, assets, credit, and banking among low-income households*, Russell Sage Foundation.

Brown, J. R., J. A. Cookson, and R. Z. Heimer (2019): "Growing up without finance," *Journal of Financial Economics*, 134, 591–616.

Büchel, K., M. V. Ehrlich, D. Puga, and E. Viladecans-Marsal (2020): "Calling from the outside: The role of networks in residential mobility," *Journal of Urban Economics*, 119, 103277.

Caskey, J. P. (1994): *Fringe banking: Check-cashing outlets, pawnshops, and the poor*, Russell Sage Foundation.

Caskey, J. P. and A. Peterson (1994): "Who has a bank account and who doesn't: 1977 and 1989," *Eastern Economic Journal*, 20, 61–73.

Célerier, C. and A. Matray (2019): "Bank-branch supply, financial inclusion, and wealth accumulation," *Review of Financial Studies*, 32, 4767–4809.

Chen, M. K., J. A. Chevalier, and E. F. Long (2021): "Nursing home staff networks and COVID-19," *Proceedings of the National Academy of Sciences*, 118.

Chen, M. K., K. Haggag, D. G. Pope, and R. Rohla (2019): "Racial disparities in voting wait times: Evidence from smartphone data," *Review of Economics and Statistics*, 1–27.

Chen, M. K. and R. Rohla (2018): "The effect of partisanship and political advertising on close family ties," *Science*, 360, 1020–1024.

Claessens, S. (2006): "Access to financial services: A review of the issues and public policy objectives," *World Bank Research Observer*, 21, 207–240.

Claessens, S. and E. Perotti (2007): "Finance and inequality: Channels and evidence," *Journal of Comparative Economics*, 35, 748–773.

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.

Coven, J., A. Gupta, and I. Yao (Forthcoming): "Urban flight seeded the Covid-19 pandemic across the United States," *Journal of Urban Economics: Insights, Forthcoming*.

Currie, J. and P. B. Reagan (2003): "Distance to hospital and children's use of preventive care: Is being closer better, and for whom?" *Economic Inquiry*, 41, 378–391.

Dahl, D. and M. Franke (2017): "'Banking deserts' become a concern as branches dry up," *Population*, 12, 5–9.

Davidson, C. (2018): "Lack of access to financial services impedes economic mobility," Online.

Davidson, R. and J. G. MacKinnon (2004): *Econometric Theory and Methods*, vol. 5, Oxford University Press New York.

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

De Sousa, J., T. Mayer, and S. Zignago (2012): "Market access in global and regional trade," *Regional Science and Urban Economics*, 42, 1037–1052.

Dekle, R. and J. Eaton (1999): "Agglomeration and land rents: Evidence from the prefectures," *Journal of Urban Economics*, 46, 200–214.

Dingel, J. I. and F. Tintelnot (2021): "Spatial economics for granular settings," Working paper. The University of Chicago, Chicago, IL.

Donaldson, D. and R. Hornbeck (2016): "Railroads and American economic growth: A "market access" approach," *Quarterly Journal of Economics*, 131, 799–858.

Eaton, J. and S. Kortum (2002): "Technology, geography, and trade," *Econometrica*, 70, 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, 478–510.

Ellwood, D. T. and N. G. Patel (2018): "Restoring the american dream," Online.

Erel, I. and J. Liebersohn (2022): "Can FinTech reduce disparities in access to finance? Evidence from the Paycheck Protection Program," *Journal of Financial Economics*, 146, 90–118.

Evans, J. W. (2018): "Simulated Method of Moments (SMM) Estimation," *QuantEcon Notes*.

Fally, T. (2015): "Structural gravity and fixed effects," *Journal of International Economics*, 97, 76–85.

Fitzpatrick, K. (2015): "The effect of bank account ownership on credit and consumption: Evidence from the UK," *Southern Economic Journal*, 82, 55–80.

Friedline, T. and M. Despard (2016): "Life in a banking desert," Online.

Fuster, A., M. Plosser, P. Schnabl, and J. Vickery (2019): "The role of technology in mortgage lending," *Review of Financial Studies*, 32, 1854–1899.

Gillibrand, K. (2021): "Senators Gillibrand And Sanders, Representatives Ocasio-Cortez, Pascrell, and Kaptur call on Congress to implement Postal Banking pilot programs," *Kirsten Gillibrand | U.S. Senator for New York*.

Goodstein, R. M. and S. L. Rhine (2017): "The effects of bank and nonbank provider locations on household use of financial transaction services," *Journal of Banking & Finance*, 78, 91–107.

Goolsbee, A. and C. Syverson (2021): "Fear, lockdown, and diversion: Comparing drivers of pandemic economic decline 2020," *Journal of Public Economics*, 193, 104311.

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

Hamrick, K. S., D. Hopkins, et al. (2012): "The time cost of access to food–Distance to the grocery store as measured in minutes," *International Journal of Time Use Research*, 9, 28–58.

Hanson, G. H. (2005): "Market potential, increasing returns and geographic concentration," *Journal of International Economics*, 67, 1–24.

Harrigan, J. (1996): "Openness to trade in manufactures in the OECD," *Journal of International Economics*, 40, 23–39.

Harris, C. D. (1954): "The market as a factor in the localization of industry in the United States," *Annals of the Association of American Geographers*, 44, 315–348.

Head, K. and T. Mayer (2004): "Market potential and the location of Japanese investment in the European Union," *Review of Economics and Statistics*, 86, 959–972.

——— (2014): "Gravity equations: Workhorse, toolkit, and cookbook," in *Handbook of International Economics*, vol. 4, 131–195.

Heckman, J. (2022): "USPS continues postal banking pilot, despite House Republicans' objections," *Federal News Network*.

Helpman, E., M. Melitz, and Y. Rubinstein (2008): "Estimating trade flows: Trading partners and trading volumes," *Quarterly Journal of Economics*, 123, 441–487.

Ho, K. and J. Ishii (2011): "Location and competition in retail banking," *International Journal of Industrial Organization*, 29, 537–546.

Hogarth, J. M., C. E. Anguelov, and J. Lee (2005): "Who has a bank account? Exploring changes over time, 1989–2001," *Journal of Family and Economic Issues*, 26, 7–30.

Hogarth, J. M. and K. H. O'Donnell (1997): "Being accountable: A descriptive study of unbanked households in the US," in *Proceedings of the Association for Financial Counseling and Planning Education*, 58–67.

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, 381–405.

Inagami, S., D. A. Cohen, B. K. Finch, and S. M. Asch (2006): "You are where you shop: Grocery store locations, weight, and neighborhoods," *American Journal of Preventive Medicine*, 31, 10–17.

Johnson, R. K. (2017): "How the United States Postal Service (USPS) could encourage more local economic development," *Chicago Kent Law Rev.*, 92, 593–615.

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.

Khwaja, A. I. and A. Mian (2008): "Tracing the impact of bank liquidity shocks: Evidence from an emerging market," *American Economic Review*, 98, 1413–42.

Kreindler, G. E. and Y. Miyauchi (2022): "Measuring commuting and economic activity inside cities with cell phone records," *Review of Economics and Statistics*, 1–48.

Larch, M., J. Wanner, Y. V. Yotov, and T. Zylkin (2019): "Currency unions and trade: A PPML re-assessment with high-dimensional fixed effects," *Oxford Bulletin of Economics and Statistics*, 81, 487–510.

Lucas, R. E. and E. Rossi-Hansberg (2002): "On the internal structure of cities," *Econometrica*, 70, 1445–1476.

McFadden, D. (1974): "The measurement of urban travel demand," *Journal of Public Economics*, 3, 303–328.

——— (1977): "Modelling the choice of residential location," *Cowles Foundation Discussion Paper, no. 477*.

——— (1989): "A method of simulated moments for estimation of discrete response models without numerical integration," *Econometrica*, 995–1026.

Melzer, B. T. (2018): "Spillovers from costly credit," *Review of Financial Studies*, 31, 3568–3594.

Miyauchi, Y., K. Nakajima, and S. J. Redding (2021): "The economics of spatial mobility: theory and evidence using smartphone data," Working paper no. 28497. National Bureau of Economic Research, Cambridge, MA.

Nguyen, H.-L. Q. (2019): "Are credit markets still local? Evidence from bank branch closings," *American Economic Journal: Applied Economics*, 11, 1–32.

Nicholl, J., J. West, S. Goodacre, and J. Turner (2007): "The relationship between distance to hospital and patient mortality in emergencies: An observational study," *Emergency Medicine Journal*, 24, 665–668.

Office of the USPS Inspector General (2014): "Providing non-bank financial services for the underserved," *USPS Office of the Inspector General, Working Paper, Washington, DC*.

Pew Research Center (2021): "Internet/Broadband Fact Sheet," Online.

Prina, S. (2015): "Banking the poor via savings accounts: Evidence from a field experiment," *Journal of Development Economics*, 115, 16–31.

Redding, S. and A. J. Venables (2004): "Economic geography and international inequality," *Journal of International Economics*, 62, 53–82.

Redding, S. J. (2013): "Economic geography: A review of the theoretical and empirical literature," *Palgrave Handbook of International Trade*, 497–531.

Redding, S. J. and E. Rossi-Hansberg (2017): "Quantitative spatial economics," *Annual Review of Economics*, 9, 21–58.

Rhine, S. L. and W. H. Greene (2013): "Factors that contribute to becoming unbanked," *Journal of Consumer Affairs*, 47, 27–45.

Rosenthal, S. S. and W. C. Strange (2004): "Evidence on the nature and sources of agglomeration economies," in *Handbook of Regional and Urban Economics*, vol. 4, 2119–2171.

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*, vol. 109, 403–08.

Sanders, B. (2021): "Fair banking for all," *Bernie Sanders Official Website*.

Silva, J. S. and S. Tenreyro (2006): "The log of gravity," *Review of Economics and statistics*, 88, 641–658.

Squire, R. F. (2019): "Quantifying sampling bias in SafeGraph Patterns," *SafeGraph Blog*.

Thaenraj, P. (2021): "Identifying CBG Sinks," *SafeGraph Blog*.

Warren, E. (2014): "The big benefits of postal service banking," *U.S. News*.

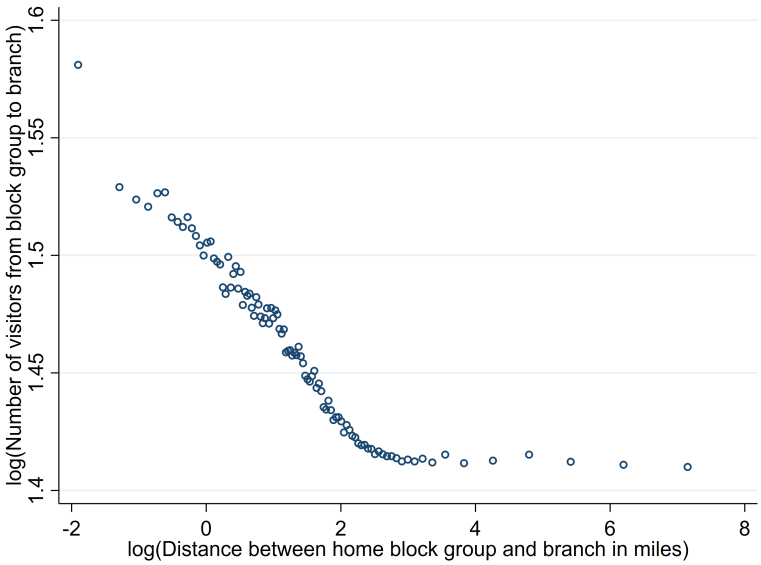
Washington, E. (2006): "The impact of banking and fringe banking regulation on the number of unbanked Americans," *Journal of Human Resources*, 41, 106–137.

Westerlund, J. and F. Wilhelmsson (2011): "Estimating the gravity model without gravity using panel data," *Applied Economics*, 43, 641–649.

Yantzi, N., M. W. Rosenberg, S. O. Burke, and M. B. Harrison (2001): "The impacts of distance to hospital on families with a child with a chronic condition," *Social Science & Medicine*, 52, 1777–1791.

Yogo, M., A. Whitten, and N. Cox (2022): "Financial Inclusion Across the United States," Working paper. Princeton University, Princeton, NJ.

(A) All Block Group × Branch Pairs



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

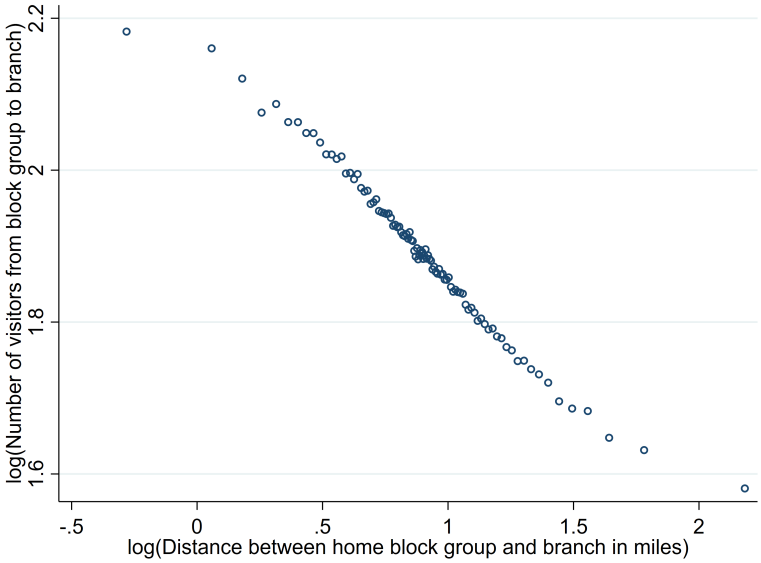


FIGURE I

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 the branch. Centers of population are from the 2010 Census, and we use the haversine formula to compute distance (see Footnote 18). Panel A presents the observed (raw) 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 × 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.

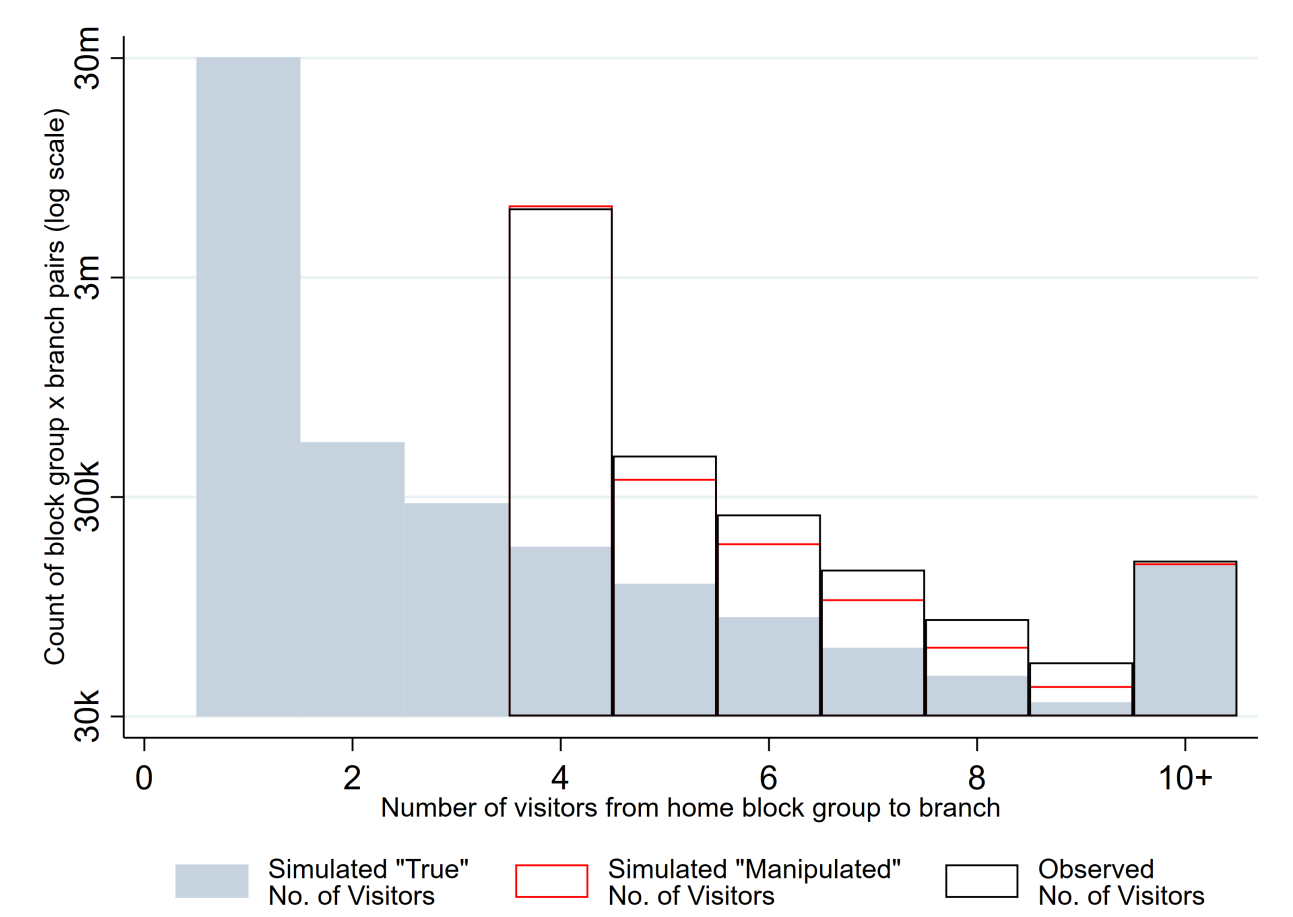


FIGURE II
DISTRIBUTIONS OF VISITOR COUNTS

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. (12), are the raw 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. (13), 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. (10) to (12). The simulated values are computed from the month-by-month Method of Simulated Moments estimation described in Appendix A. 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 × branch pairs with visitor counts exceeding 10 is assigned to 10+ visitors in the panel.

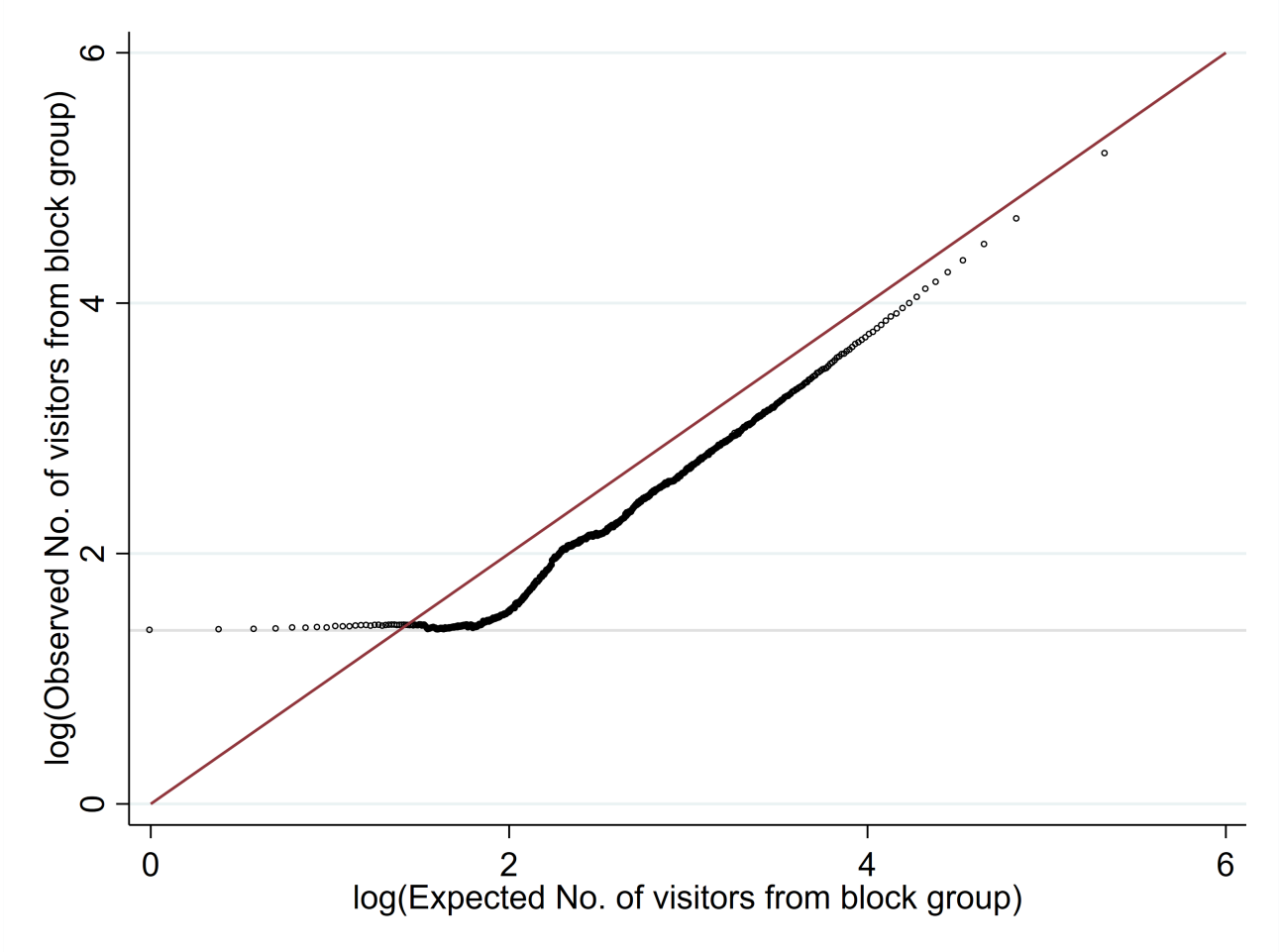


FIGURE III
OBSERVED VS. EXPECTED BRANCH VISITORS PER CENSUS BLOCK GROUP

The figure presents a scatter plot of the log observed number of branch visitors from each Census block group (i.e., $V_{it} \equiv \sum_j V_{ijt}$, where V_{ijt} is given in Eq. (12)) versus the log expected number of branch visitors from each block group based on the month-by-month MSM estimates (i.e., $\hat{V}^a_{it} \equiv \exp(\hat{\gamma}_{it})\hat{\Phi}^a_{it}$, where the access measure $\hat{\Phi}^a_{it} \equiv \sum_{j \in B_{it}} \omega^i_{ijt} \exp\left(\hat{\lambda}_{jt}\right) d_{ij}^{-\hat{\beta}}$ reflects the frequency weights used in the stratified sampling). 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 Appendix A.

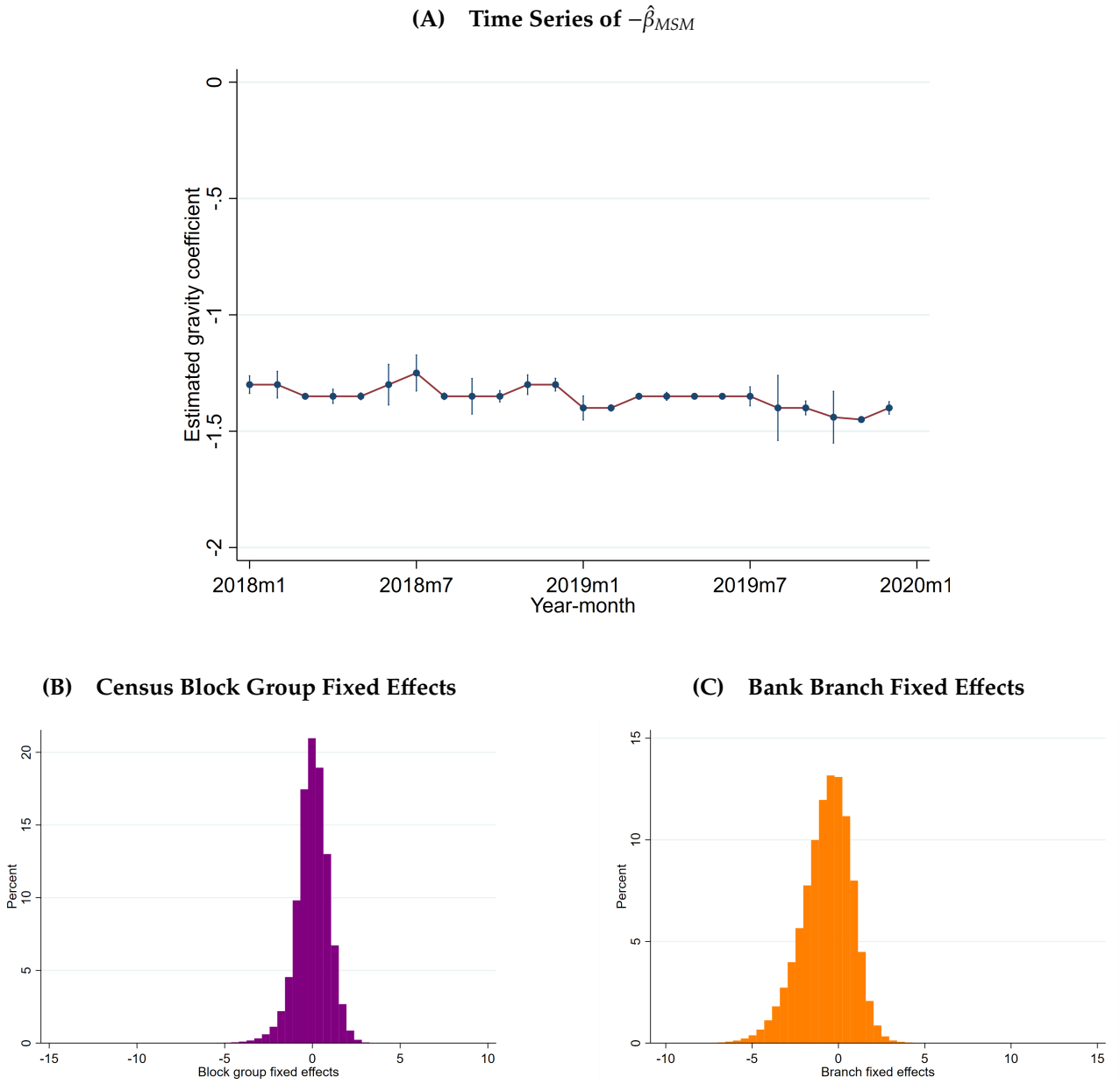


FIGURE IV
METHOD OF SIMULATED MOMENTS PARAMETER ESTIMATES

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. (13). Panel A illustrates the monthly time series of the $-\hat{\beta}_{\text{MSM}}$ 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. The step-by-step details of the MSM estimation are provided in Appendix A.

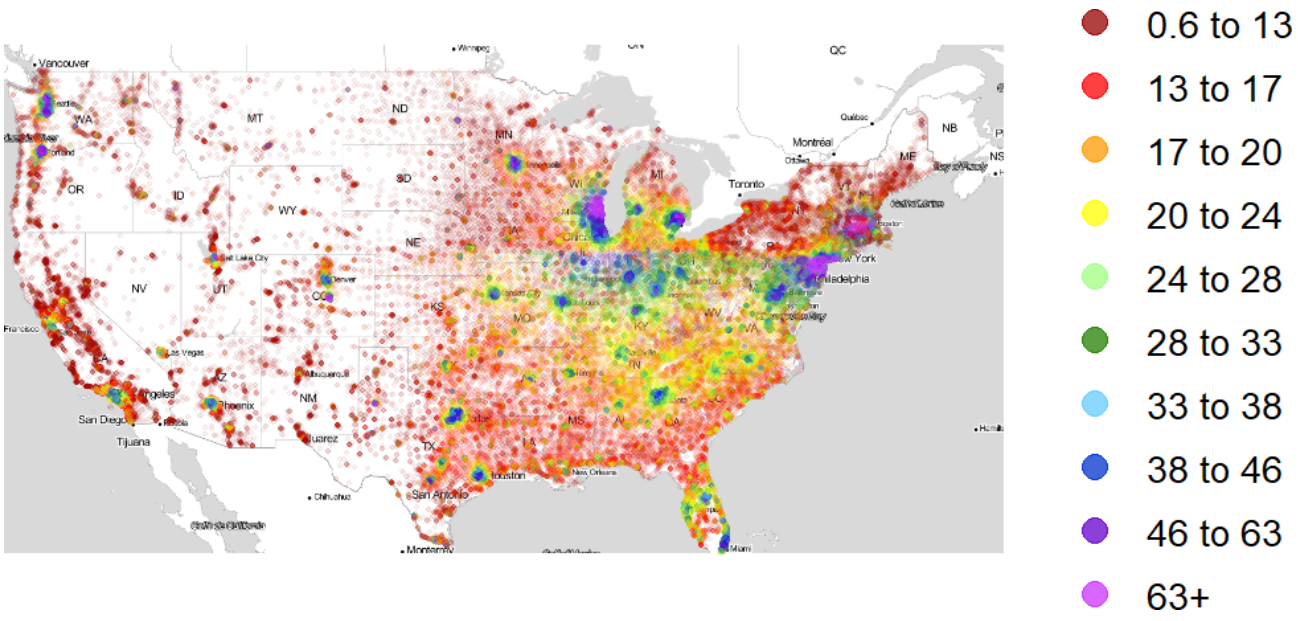


FIGURE V
BANK ACCESS NATIONWIDE

The figure illustrates a dot density map of bank access by Census block groups nationwide. The figure is based on our core SafeGraph sample of bank branches; i.e., 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. Each dot is positioned at a block group's center of population. Bank access estimates are calculated from Eq. (14) and are based on the Method of Simulated Moments estimation described in Appendix A. Access estimates are calculated month-by-month per block group, and the figure presents weighted monthly averages, where each month's weight is its share of the block group's total observed branch visitors over the core sample period (January 2018 - December 2019). The map is constructed by grouping block groups into deciles and shading the dots so that higher-ordered colors in the rainbow gradient (indigo and violet) imply higher bank access values and lower-ordered colors (red and orange) imply lower access values. Block groups where no resident was recorded in SafeGraph as having visited a branch in the sample period are shaded white.

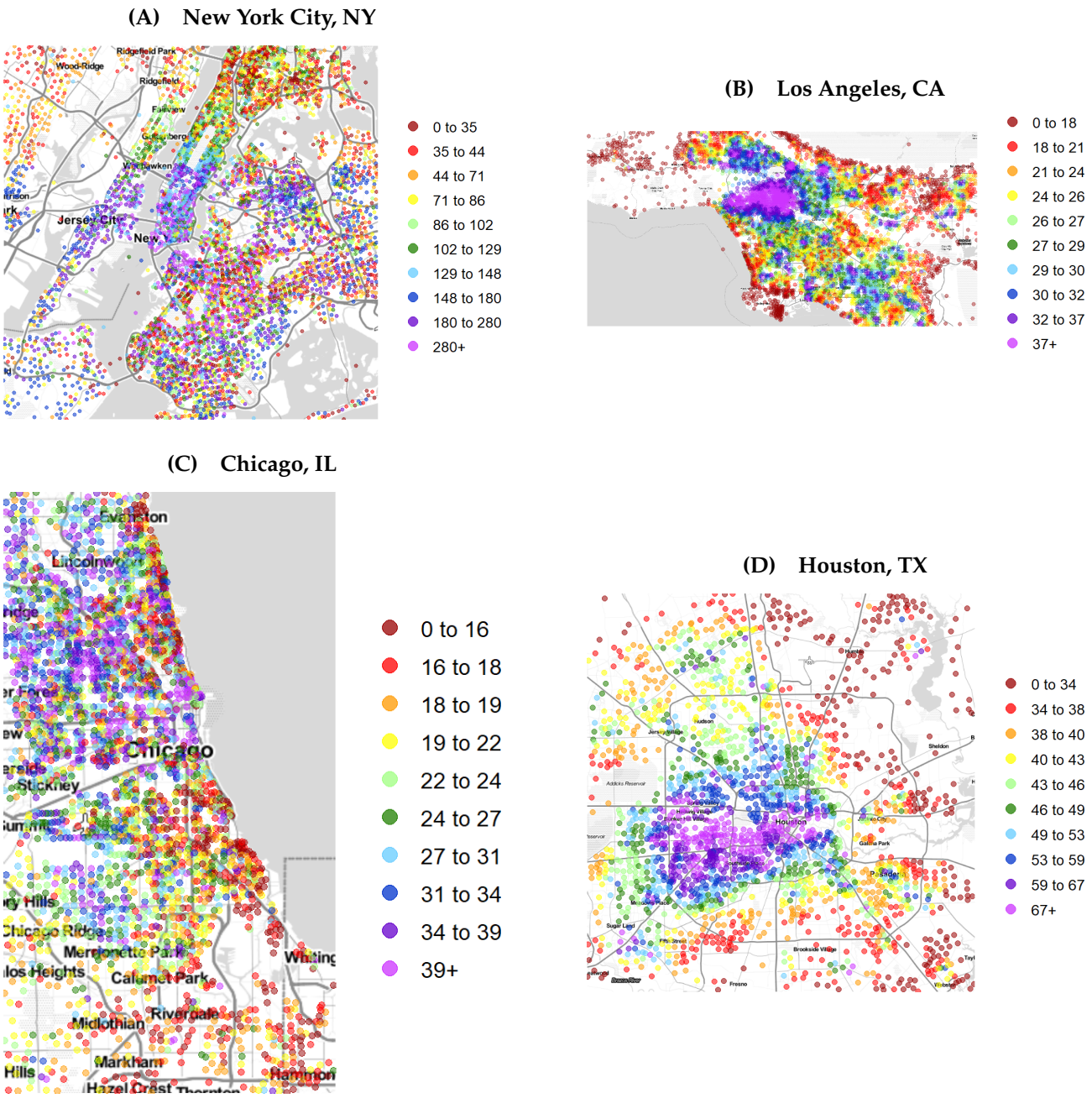


FIGURE VI
BANK ACCESS IN THE FOUR LARGEST U.S. CITIES

The figure illustrates dot density maps of bank access by Census block groups in the four largest U.S. cities by population as of the 2020 Census. The figure is based on our core SafeGraph sample of bank branches; i.e., 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. Each dot in a panel is positioned at a block group's center of population. Bank access estimates are calculated from Eq. (14) and are based on the Method of Simulated Moments estimation described in Appendix A. Access estimates are calculated month-by-month per block group, and the panels present weighted monthly averages, where each month's weight is its share of the block group's total observed branch visitors over the core sample period (January 2018 - December 2019). Each panel's map is constructed by grouping block groups within the panel into deciles and shading the dots so that higher-ordered colors in the rainbow gradient (indigo and violet) imply higher bank access values and lower-ordered colors (red and orange) imply lower access values. Block groups where no resident was recorded in SafeGraph as having visited a branch in the sample period are shaded white.

TABLE I
BANK ACCESS BY DEMOGRAPHIC ATTRIBUTES

Dep. var.:	log(Bank access of block groups)				log(Branch proximity)		log(Branch quality)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
log(Income)	-0.110	-0.076	-0.126	-0.087	-0.076	-0.086	0.001	-0.001
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.001)	(0.002)
Black	-0.082	-0.053	-0.107	-0.064	-0.143	-0.168	0.090	0.104
	(0.005)	(0.005)	(0.006)	(0.006)	(0.006)	(0.006)	(0.003)	(0.003)
Asian	0.470	0.438	0.429	0.398	0.422	0.375	0.016	0.023
	(0.014)	(0.013)	(0.014)	(0.013)	(0.012)	(0.013)	(0.007)	(0.007)
Other	0.023	0.020	0.078	0.073	0.006	0.064	0.015	0.009
	(0.023)	(0.023)	(0.033)	(0.033)	(0.022)	(0.031)	(0.011)	(0.015)
Hispanic	0.046	0.081	0.021	0.071	-0.003	-0.030	0.085	0.101
	(0.007)	(0.007)	(0.007)	(0.008)	(0.007)	(0.008)	(0.004)	(0.004)
Age <15		-0.721		-0.814	-0.757	-0.855	0.036	0.041
		(0.017)		(0.020)	(0.017)	(0.019)	(0.010)	(0.011)
Age 35-54		-0.238		-0.204	-0.253	-0.209	0.015	0.005
		(0.017)		(0.020)	(0.017)	(0.020)	(0.009)	(0.011)
Age 55-64		-0.551		-0.573	-0.675	-0.691	0.124	0.117
		(0.019)		(0.022)	(0.019)	(0.023)	(0.010)	(0.012)
Age 65+		-0.245		-0.262	-0.256	-0.277	0.010	0.014
		(0.013)		(0.015)	(0.013)	(0.015)	(0.007)	(0.008)
log(No. of devices)	-0.050	-0.053	-0.057	-0.059	-0.063	-0.072	0.010	0.012
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
Observations	2,549,020	2,549,020	1,847,252	1,847,252	2,549,020	1,847,252	2,549,020	1,847,252
Adjusted R^2	0.704	0.708	0.634	0.640	0.820	0.774	0.684	0.646
Sample	Core	Core	MC	MC	Core	MC	Core	MC
Year-month FE	O	O	O	O	O	O	O	O
County FE	O	O	O	O	O	O	O	O
RUCA FE	O	O			O		O	

Each column reports coefficients from a multivariate, weighted OLS regression with standard errors clustered at the Census-block-group level reported in parentheses. One observation is a block group per month per year in the sample period from January 2018 - December 2019. Block groups where no resident was recorded in SafeGraph as having visited a bank branch in the year-month are dropped. Observations are weighted by block-group population counts from the 2019 5-year American Community Survey (ACS). All columns use our core sample of branch locations, 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 (SOD). Demographic independent variable observations are population-based decimal shares from the 2019 5-year ACS. Income is median household income. In columns (1)-(4), the dependent variable is the natural logarithm of the estimated bank access measure, $\log \hat{\Phi}_{it}$, from Eq. (14). In columns (5)-(6), the dependent variable is the natural logarithm of the "branch proximity" component in the decomposition of $\hat{\Phi}_{it}$ in Eq. (15), whereas in columns (7)-(8), the dependent variable is the natural logarithm of the "average branch quality" component in that decomposition. All dependent variables are computed from the month-by-month Method of Simulated Moments estimation described in Appendix A. Columns (1), (2), (5), and (7) include all block groups for which we have branch visitor data, whereas columns (3), (4), (6), and (8) restrict the sample to block groups with Rural-Urban Commuting Areas (RUCA) codes equaling 1 (Metropolitian area core). The omitted demographic groups are non-Hispanic Whites and age range 15-34.

TABLE II
BANK BRANCH USE AND BLOCK GROUP FIXED EFFECTS BY DEMOGRAPHIC ATTRIBUTES

Dep. var.:	log(Expected no. of visitors)				Block group fixed effects			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
log(Income)	0.186	0.155	0.190	0.155	0.296	0.230	0.315	0.242
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)
Black	-0.041	-0.056	-0.017	-0.039	0.041	-0.003	0.089	0.025
	(0.009)	(0.009)	(0.010)	(0.010)	(0.010)	(0.010)	(0.011)	(0.011)
Asian	0.238	0.241	0.227	0.224	-0.232	-0.197	-0.202	-0.174
	(0.020)	(0.020)	(0.020)	(0.019)	(0.023)	(0.022)	(0.023)	(0.022)
Other	0.034	0.041	0.009	0.016	0.010	0.020	-0.068	-0.058
	(0.028)	(0.029)	(0.038)	(0.039)	(0.033)	(0.033)	(0.046)	(0.046)
Hispanic	0.008	-0.016	0.021	-0.012	-0.038	-0.097	-0.000	-0.083
	(0.009)	(0.010)	(0.010)	(0.011)	(0.010)	(0.011)	(0.011)	(0.012)
Age <15		0.341		0.343		1.061		1.157
		(0.022)		(0.025)		(0.026)		(0.030)
Age 35-54		0.486		0.529		0.724		0.732
		(0.023)		(0.026)		(0.025)		(0.029)
Age 55-64		0.101		0.053		0.651		0.626
		(0.025)		(0.030)		(0.029)		(0.034)
Age 65+		0.248		0.253		0.494		0.515
		(0.018)		(0.020)		(0.021)		(0.023)
log(No. of devices)	0.606	0.606	0.601	0.601	0.656	0.659	0.658	0.660
	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Observations	2,549,020	2,549,020	1,847,252	1,847,252	2,549,020	2,549,020	1,847,252	1,847,252
Adjusted R^2	0.380	0.381	0.339	0.340	0.310	0.314	0.309	0.313
Sample	Core	Core	MC	MC	Core	Core	MC	MC
Year-month FE	O	O	O	O	O	O	O	O
County FE	O	O	O	O	O	O	O	O
RUCA FE	O	O			O	O		

Each column reports coefficients from a multivariate, weighted OLS regression with standard errors clustered at the Census-block-group level reported in parentheses. One observation is a block group per month per year in the sample period from January 2018 - December 2019. Block groups where no resident was recorded in SafeGraph as having visited a bank branch in the year-month are dropped. Observations are weighted by block-group population counts from the 2019 5-year American Community Survey (ACS). Dependent variable observations are based on our core sample of branch locations, 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. Demographic independent variable observations are population-based decimal shares from the 2019 5-year ACS. Income is median household income. The dependent variable in columns (1)-(4) is the natural logarithm of the expected number of branch goers from each block group based on the month-by-month Method of Simulated Moments estimates; i.e., $\log \hat{V}^*_{it} \equiv \log \sum_j \hat{V}^*_{ijt}$, where $\hat{V}^*_{ijt}$ is the predicted mean of V^*_{ijt} in Eq. (13) and β is time-varying. The dependent variable in columns (5)-(8) is the estimated block group fixed effects, $\hat{\gamma}_{it}$, from the gravity relation in Eq. (13). Columns (1), (2), (5), and (6) include all block groups for which we have branch visitor data, whereas columns (3), (4), (7), and (8) restrict the sample to block groups with Rural-Urban Commuting Areas (RUCA) codes equaling 1 (Metropolitian area core). The omitted demographic groups are non-Hispanic Whites and age range 15-34.

Appendix

A Method of Simulated Moments

This section describes the steps used in the Method of Simulated Moments (MSM) to estimate the parameters of the gravity relation in Eq. (13) that enter our measure of local bank access in Eq. (14). The MSM is run separately per year-month of our sample period from January 2018 - December 2019.

A.1 Specify the DGP for visitors

The data generating process (DGP) we simulate is the number of visitors between block groups and branches through time. To account for measurement error in the SafeGraph data and unobservable factors that influence visits, we assume that the true visitor count V^*_{ijt} is Poisson distributed with mean $\pi_{ijt}R_{it}$, where R_{it} is the number of residents of block group i in year-month t and π_{ijt} is the share of residents living in block group i that visit branch j in year-month t. Using the gravity equation from the conceptual framework in Eq. (8) to inform π_{ijt}, we express the true visitor count as obeying Eq. (13). 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.[18]

A.2 Sample branches included in estimation

Technically speaking, every branch in our sample is available to each resident and ought to enter a block group's measure of bank access. A resident of a block group in, say, New York City, could travel to a First Republic Bank branch in San Francisco three thousand miles away. But the chances are low, and a sample of over fifty thousand branches, over two hundred thousand block groups, altogether spanning twenty-four months, makes it computationally challenging to calculate an access measure that accounts for all possible block group × branch pairs.

Instead, we sample branches per block group using a technique that is similar in spirit to the procedure described in McFadden (1977) and Davis et al. (2019). Specifically, for each block group, we include the locations of branches that block group residents actually visited in the year-month plus a random subset of all *alternative* branches that they could have visited in the period. We select these alternative branches using stratified sampling in the following manner: Per block group in each year-month, alternative branches are ranked by their distance away. We divide the number of alternative branches by 2,000 and round that value up. We then divide the distribution of ranked branches into these many quantiles and select one branch from each quantile with equal probability. For example, in a year-month in which a block group exhibits 38,020 alternative branches, the number of quantiles would be 20, and hence, the block group's number of randomly sampled alternative branches would be 20.

Notice that, in the observed SafeGraph data, the number of residents in a block group who visit a branch in the block group's random alternative set of branches is 0. This means that either no resident truly visited that branch or the true visitor count was hidden by SafeGraph's differential privacy.

To establish notation for the samples of block groups, branches, and block group × branch pairs, we let B_{it} denote the sample of branches for block group i in year-month t, which is the union of the branches with positive observed visitor counts, denoted B^1_{it}, and those with 0 visitor counts from the random sample of alternative branches, denoted B^0_{it}. Similarly, let H_{jt} denote the sample of block groups for branch j in year-month t, which is the union of the block

[18] 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 traveling distance is the road driving distance between locations. The haversine formula is simple, fairly accurate, and convenient to compute. In Online Table B.8, 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 above the median share from the 2019 ACS. Across the samples, the regressions produce very high R^2, ranging from 0.972 to 0.993, implying that the haversine distance captures variation in driving times quite well.

groups with residents who are observed in the data to have visited the branch, denoted H^1_{jt}, and the random sample of alternative block groups with 0 visitor counts, denoted H^0_{jt}. Finally, let N_{it} denote the sample of pairs of block groups and branches in year-month t, which is the union of pairs with positive observed visitor counts, denoted N^1_t, and the pairs with 0 visitor counts from the random stratified sampling, denoted N^0_t. Notice that, by definition, $\sum_i |B_{it}| = |N_t|$ and $\sum_j |H_{jt}| = |N_t|$, where $|\cdot|$ denotes cardinality.

The stratified sampling suggests that we must apply frequency weights to any variable measured at the (block group i, branch j) level, such as visitor counts or pairwise distance, so as to rebalance the data and make it represent the target population as closely as possible. Following standard practice, the frequency weights are the reciprocal of the likelihood of being sampled (subject to rounding). Although the frequency weights are at the (i, j) level, we need to use them when computing both i-level and j-level fixed effects. Thus, we construct two sets of frequency weights per year-month, one from the perspective of block group i and the other from the perspective of branch j. In both cases, because (i, j) pairs with positive visitor counts are always included in the sample, we assign these pairs a frequency weight of 1. On the other hand, the (i, j) pairs with zero visitor counts (which make up the random alternative set of pairs per year-month) are assigned a weight equaling the number of pairs they represent. Let ω^i_{ijt} denote the frequency weight from block group i's perspective in year-month t, and let ω^j_{ijt} denote the frequency weight from branch j's perspective. The weights from the branch perspective, ω^j_{ijt}, are only used in the calculation of the branch fixed effects. The weights $\omega^i_{ijt} = 1$ and $\omega^j_{ijt} = 1$ for (i, j) pairs with positive observed visitor counts $\left(\text{i.e., } \forall\, (i, j) \in N^1_t\right)$. Weights for (i, j) pairs from the stratified sampling satisfy the equations:

$$\omega^i_{ijt}|B^0_{it}| + 1|B^1_{it}| = \text{Total no. of branches in year-month } t \qquad \forall\, (i, j) \in N^0_t, \tag{A.21}$$

$$\omega^j_{ijt}|H^0_{jt}| + 1|H^1_{it}| = \text{Total no. of block groups in year-month } t, \qquad \forall\, (i, j) \in N^0_t. \tag{A.22}$$

To better understand the calculation of the frequency weights, consider the set of weights from a block group's perspective in Eq. (A.21). These weights will be used in computing any statistic at the block group level, using either simulated or observed data. Continuing with the example from earlier, if the total number of branches in the year-month minus the number of branches with positive visitor counts from block group i (i.e., $|B^1_{it}|$) equaled 38,020, this implies that the frequency weight used in computing statistics for the block group (such as its fixed effects) would be 1901, since $|B^0_{it}| = 20$. On average, from a block group's perspective, the random alternative set of branches per block group represents slightly higher than a 0.05% sample size of all possible alternative branches from our core set of branches (i.e., $[(^{38,000}/_{2,000})\,/2,000 \times 100\%] = 0.05\%$).

A.3 Initialize the fixed effects routine

The MSM uses the visitor data v and the model parameters $\psi \equiv \left\{\beta, \gamma_{it}, \lambda_{jt}\right\}$ 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. (13) requires hundreds of thousands of fixed effects to be estimated. There are simply too many parameters to identify from the MSM minimization problem alone. Instead, we adopt an iterative routine to identify the fixed effects $\left\{\gamma_{it}, \lambda_{jt}\right\}$ and let the minimization problem identify β. Holding fixed an estimate of β 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, the MSM minimization problem then chooses the optimal β estimate that satisfies the moment conditions. In the month-by-month estimation, we estimate ψ per year-month of the sample, which means we run the MSM 24 times and produce 24 estimates of ψ. We initialize the fixed effects routine with guessed estimates $\hat{\gamma}^0_{it} = \hat{\lambda}^0_{jt} = 1$ for all i and j.

A.4 Simulate visitor counts

We run $S = 10$ simulations of the number of visitors per block group × branch pair. The S simulations are run per year-month of the sample. Recall that N^1_t denotes the sample of block group × branch pairs with positive visitor counts in year-month t, and N^0_t denotes the sample of block group × branch pairs with 0 visitor counts in year-month t, coinciding with the number of (i, j) pairs that arise from the stratified sampling. We distinguish these two groups when simulating visitor counts because they enter the simulation differently, precisely because the stratified sampled pairs are adjusted to represent the target population.

Consider first the set N_t^1 of pairs with positive observed visitor counts. 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 $\hat{\beta}$ 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. (13). Each Poisson draw is a "true" block group × branch visitor count. Then, 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. 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.

Consider next the set N_t^0 of block group- × branch pairs with 0 visitor counts that arise from the stratified sampling. If an extra $|N_t^0| \times S$ pairs of visitor counts were simulated in the same manner described in the previous paragraph, they would have disproportionate impact on any computed moments because of the high frequency weights that would multiply them. Noise from the simulation would be amplified and make the estimation unstable. Rather than simulating these observations, we construct their implied empirical probability distribution according to the parameter estimate ψ in each iteration. If an infinite number of visitor counts 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 block group × branch pairs with positive visitor counts 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 pairs in the set N_t^0 are sampled and meant to represent the remaining population of (i, j) pairs, 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.

That empirical distribution is a truncated and censored version of a Poisson probability mass function per Eq. (13) plus a Laplace distribution with mean 0 and scale $^{10}/_{9}$. Because the Laplace noise is added to the Poisson draw, this empirical distribution can be thought of as a truncated and censored Laplace distribution whose parameter location (mean) is the realization of the Poission draw. With this in mind, let $G(y, k)$ be the CDF of a Laplace distribution with location parameter k and scale $^{10}/_{9}$. And let $\hat{\mu}_{ijt}$ denote the estimated mean of the Poisson visitor count in Eq. (13). 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.

We construct 7 components of the empirical distribution that we use in the moments of the estimation, provided in Appendix A.6. 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:**

$$\Pr\left(\tilde{V}_{ijt} = 0 | \hat{\mu}_{ijt}\right) \approx p\left(0, \hat{\mu}_{ijt}\right) + \sum_{k=1}^{K} p\left(k, \hat{\mu}_{ijt}\right) \times G(2, k). \tag{A.23}$$

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. (A.23), 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. (A.23). In that term, the Laplace draw has mean k to adjust for different possible positive draws of the Poisson, and the CDF value of the Laplace 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:**

$$\Pr\left(\tilde{V}_{ijt} > 0 | \hat{\mu}_{ijt}\right) \approx \sum_{k=1}^{K} p\left(k, \hat{\mu}_{ijt}\right) \times \left(1 - G\left(2, k\right)\right). \tag{A.24}$$

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:**

$$\Pr\left(\tilde{V}_{ijt} = 4 | \hat{\mu}_{ijt}\right) \approx \sum_{k=1}^{K} p\left(k, \hat{\mu}_{ijt}\right) \times \left(G\left(5, k\right) - G\left(2, k\right)\right). \tag{A.25}$$

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 to 4.

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

$$\Pr\left(\tilde{V}_{ijt} > 4 | \hat{\mu}_{ijt}\right) \approx \sum_{k=1}^{K} p\left(k, \hat{\mu}_{ijt}\right) \times \left(1 - G\left(5, k\right)\right). \tag{A.26}$$

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:**

$$E\left(\tilde{V}_{ijt} | \hat{\mu}_{ijt}\right) \approx \sum_{k=1}^{K} p\left(k, \hat{\mu}_{ijt}\right) \left[4 \times \left\{G\left(5, k\right) - G\left(2, k\right)\right\} + \sum_{l=5}^{L} l \times \left\{G\left(l+1, k\right) - G\left(l, k\right)\right\}\right]. \tag{A.27}$$

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, giben by $G\left(l+1, k\right) - G\left(l, k\right)$.

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

$$E\left(\log \tilde{V}_{ijt} | \tilde{V}_{ijt} > 0, \hat{\mu}_{ijt}\right) \approx \frac{\sum_{k=1}^{K} p\left(k, \hat{\mu}_{ijt}\right) \left[\log 4 \times \left\{G\left(5, k\right) - G\left(2, k\right)\right\} + \sum_{l=5}^{L} \left\{\log l \times \left(G\left(l+1, k\right) - G\left(l, k\right)\right)\right\}\right]}{\Pr\left(\tilde{V}_{ijt} > 0 | \hat{\mu}_{ijt}\right)}. \tag{A.28}$$

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. (A.27). 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. (A.24), which is the way to compute the mean of a truncated random variable.

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

$$\mathbb{E}\left(\log \tilde{V}_{ijt} | \tilde{V}_{ijt} > 4, \hat{\mu}_{ijt}\right) \approx \frac{\sum_{k=1}^{K} p\left(k, \hat{\mu}_{ijt}\right)\left[\sum_{l=5}^{L} \{\log l \times (G(l+1,k) - G(l,k))\}\right]}{\Pr\left(\tilde{V}_{ijt} > 4 | \hat{\mu}_{ijt}\right)}. \tag{A.29}$$

This conditional mean is even simpler to compute than the one in Eq. (A.28). The formula consists of just the second component in the numerator of Eq. (A.28), and the re-weighting in the denominator is the probability of the visitor count exceeding 4, given in Eq. (A.26).

A.5 Iterate the fixed effects until convergence

Under a fixed estimate $\hat{\beta}$, the next step is to iterate the estimated fixed effects until they converge. We iterate the estimated fixed effects sequentially, beginning with the estimated branch fixed effects $\left\{\hat{\lambda}_{jt}\right\}$, while holding constant the estimated block group fixed effects $\{\hat{\gamma}_{it}\}$ at $\hat{\gamma}_{it}^{0} = 1, \forall i$. The iteration routine is as follows. Suppose we are on the k-th iteration. To estimate the branch fixed effects, we take advantage of another data field in SafeGraph: the total count of visitors to a branch. The SafeGraph name for this field is `RAW_VISITOR_COUNT`. Unlike the number of visitors from a block group to the branch, the branch's total visitor count 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 visitor count to uniquely pin down the estimate of the branch's fixed effect. Let V_{jt}^{T} denote branch j's total visitor count in year-month t.

From Eq. (13), 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

$$\hat{V}_{ijt}^{k} = \exp\left(\hat{\lambda}_{jt}^{k}\right) \exp\left(\hat{\gamma}_{it}^{k}\right) d_{ij}^{-\hat{\beta}}. \tag{A.30}$$

Summing across block groups, and adjusting for the frequency weights from the stratified sampling, gives a branch's expected total visitor count:

$$\hat{V}_{jt}^{k} = \exp\left(\hat{\lambda}_{jt}^{k}\right) \sum_{i \in H_{jt}} \omega_{ijt}^{j} \exp\left(\hat{\gamma}_{it}^{k}\right) d_{ij}^{-\hat{\beta}}. \tag{A.31}$$

Notice that the frequency weights, ω_{ijt}^{j}, are from a branch's perspective. For (i, j) pairs with positive observed visitor counts, these weights equal 1. For (i, j) pairs from the stratified sample, they satisfy Eq. (A.22).

Given $\hat{\beta}$ and the k-th iteration of the estimated block group fixed effects, $\hat{\gamma}_{it}^{k}$, 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. (A.31), with the branch's observed total visitor count, V_{jt}^{T}. Mathematically speaking, the branch's fixed effect estimate satisfies:

$$\hat{\lambda}_{jt}^{k} = \log V_{jt}^{T} - \log \sum_{i \in H_{jt}} \omega_{ijt}^{j} \exp\left(\hat{\gamma}_{it}^{k}\right) d_{ij}^{-\hat{\beta}} \tag{A.32}$$

Per iteration, Eq. (A.32) pins down each branch's estimated fixed effect as a function of the estimated block group fixed effects (and the estimate of β). The estimated block group fixed effects will iterate until they converge, and by Eq. (A.32), once the estimated block group fixed effects converge, so too do the estimated branch fixed effects, given an estimate of β.

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}^{1} that have positive observed visitor counts, by the average simulated visitor counts across all branches in set B_{it} and all simulations S. With this in mind, let the average observed visitor count of block group i be

$$\overline{V}_{it} = \frac{1}{|B_{it}|} \sum_{j \in B_{it}} V_{ijt}. \tag{A.33}$$

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

$$\overline{\tilde{V}}_{it}^{k}(s) = \frac{\sum_{j \in B_{it}^{1}} \tilde{V}_{ijt}^{k}(s) + \sum_{j \in B_{it}^{0}} \omega_{ijt}^{i} E\left(\tilde{V}_{ijt}^{k} | \hat{\mu}_{ijt}\right)}{\sum_{j \in B_{it}} \omega_{ijt}^{i}}, \tag{A.34}$$

where $E\left(\tilde{V}_{ijt}^{k}|\hat{\mu}_{ijt}\right)$ is provided in Eq. (A.27). Because the calculation is at the block group-level, the frequency weights used are from the block group perspective, and they either equal 1 or satisfy Eq. (A.21). Averaging across simulations delivers the mean simulated visitor count of block group i as

$$\overline{\tilde{V}}_{it}^{k} = \frac{1}{S}\sum_{s}\overline{\tilde{V}}_{it}^{k}(s)\,. \tag{A.35}$$

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

$$\chi_{it}^{k} = \frac{\overline{V}_{it}}{\overline{\tilde{V}}_{it}^{k}} \tag{A.36}$$

We take ratios of averages rather than differences of averages because the fixed effects in the visitor count model in Eq. (13) are exponentiated. These block group-level ratios then multiplicatively update each block group's estimated fixed effect:

$$\hat{\gamma}_{it}^{k+1} = \hat{\gamma}_{it}^{k} \times \left(\chi_{it}^{k}\right)^{h}, \tag{A.37}$$

where h 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 $N_t \times S$ Uniform random variables into Poisson random variables using (i) the estimate $\hat{\beta}$; (ii) the updated block group fixed effect estimates, $\hat{\gamma}_{it}^{k+1}$; and (iii) the updated branch fixed effect estimates, $\hat{\lambda}_{jt}^{k+1}$, based on Eq. (A.32). We then apply differential privacy methods to the "updated" simulated data. The process iterates until the estimated block group fixed effects converge.[19]

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 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

$$\left[\frac{1}{N_t}\sum_{i} B_{it}\left(\overline{\tilde{V}}_{it}^{k+1} - \overline{V}_{it}\right)^2 - \frac{1}{N_t}\sum_{i} B_{it}\left(\overline{\tilde{V}}_{it}^{k} - \overline{V}_{it}\right)^2\right]^2 < \varepsilon \tag{A.38}$$

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

After the condition in Eq. (A.38) is met, we have converged fixed effects estimates, denoted $\hat{\gamma}_{it}^{\infty}$ and $\hat{\lambda}_{jt}^{\infty}$, for a given estimate of β. The final piece of the estimation is to select the optimal β estimate that minimizes the distance between simulated and data moments.

A.6 Select the moments

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

[19] 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).

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

$$m_1(v) \equiv \frac{\sum_{(i,j)\in N_t} \mathbb{I}\left(V_{ijt}=0\right)\omega^i_{ijt}}{\sum_{(i,j)\in N_t}\omega^i_{ijt}}, \tag{A.39}$$

$$m_1(\tilde{v}|\psi) \equiv \frac{\sum_{(i,j)\in N^1_t} \mathbb{I}\left(\tilde{V}_{ijt}=0\right) + \sum_{(i,j)\in N^0_t} \Pr\left(\tilde{V}_{ijt}=0|\hat{\mu}_{ijt}\right)\omega^i_{ijt}}{\sum_{(i,j)\in N_t}\omega^i_{ijt}}, \tag{A.40}$$

where $\mathbb{I}(\cdot)$ stands for the indicator function and $\Pr\left(\tilde{V}_{ijt}=0|\hat{\mu}_{ijt}\right)$ is from Eq. (A.23). The data moment $m_1(v)$ is straightforward. The simulated moment $m_1(\tilde{v}|\psi)$ adds the fraction of the simulated visitor counts equaling 0 to the probability of the stratified sample of visitor counts equaling 0, adjusted by the frequency weights.

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

$$m_2(v) \equiv \frac{\sum_{(i,j)\in N_t} \mathbb{I}\left(V_{ijt}=4\right)\omega^i_{ijt}}{\sum_{(i,j)\in N_t}\omega^i_{ijt}}, \tag{A.41}$$

$$m_2(\tilde{v}|\psi) \equiv \frac{\sum_{(i,j)\in N^1_t} \mathbb{I}\left(\tilde{V}_{ijt}=4\right) + \sum_{(i,j)\in N^0_t} \Pr\left(\tilde{V}_{ijt}=4|\hat{\mu}_{ijt}\right)\omega^i_{ijt}}{\sum_{(i,j)\in N_t}\omega^i_{ijt}}, \tag{A.42}$$

where $\Pr\left(\tilde{V}_{ijt}=4|\hat{\mu}_{ijt}\right)$ is from Eq. (A.25).

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

$$m_3(v) \equiv \frac{\sum_{(i,j)\in N_t} \mathbb{I}\left(V_{ijt}=0\right)\omega^i_{ijt}\log d_{ij}}{\sum_{(i,j)\in N_t} \mathbb{I}\left(V_{ijt}=0\right)\omega^i_{ijt}}, \tag{A.43}$$

$$m_3(\tilde{v}|\psi) \equiv \frac{\sum_{(i,j)\in N^1_t} \mathbb{I}\left(\tilde{V}_{ijt}=0\right)\log d_{ij} + \sum_{(i,j)\in N^0_t} \Pr\left(\tilde{V}_{ijt}=0|\hat{\mu}_{ijt}\right)\omega^i_{ijt}\log d_{ij}}{\sum_{(i,j)\in N_t} \mathbb{I}\left(\tilde{V}_{ijt}=0\right)\omega^i_{ijt}}. \tag{A.44}$$

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

$$m_4(v) \equiv \frac{\sum_{(i,j)\in N_t} \mathbb{I}\left(V_{ijt}=4\right)\omega^i_{ijt}\log d_{ij}}{\sum_{(i,j)\in N_t} \mathbb{I}\left(V_{ijt}=4\right)\omega^i_{ijt}}, \tag{A.45}$$

$$m_4(\tilde{v}|\psi) \equiv \frac{\sum_{(i,j)\in N^1_t} \mathbb{I}\left(\tilde{V}_{ijt}=4\right)\log d_{ij} + \sum_{(i,j)\in N^0_t} \Pr\left(\tilde{V}_{ijt}=4|\hat{\mu}_{ijt}\right)\omega^i_{ijt}\log d_{ij}}{\sum_{(i,j)\in N_t} \mathbb{I}\left(\tilde{V}_{ijt}=4\right)\omega^i_{ijt}}. \tag{A.46}$$

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

$$y_{ijt} = \left\langle \log V_{ijt} \right\rangle_{(i,j)\in N^1_t} \tag{A.47}$$

$$X_{ijt} = \left[\ \langle 1\rangle_{(i,j)\in N^1_t}\ ,\quad \left\langle \log d_{ij}\right\rangle_{(i,j)\in N^1_t}\ \right]. \tag{A.48}$$

Here, $\langle\cdot\rangle_{(i,j)\in N^1_t}$ denotes a vector with length equaling the number of elements in the set N^1_t. 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

$$m_5(v) \equiv \text{Second element of } \left(X'_{ijt}X_{ijt}\right)^{-1}\left(X'_{ijt}y_{ijt}\right) \tag{A.49}$$

Notice that, because the moment reflects only positive observed visitor counts, the frequency 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 frequency weights are not all 1. With this in mind, the observation weights of the WLS are defined as

$$\tilde{\eta}_{ijt} \equiv \begin{bmatrix} \langle 1 \rangle_{(i,j) \in N_t^1 : \tilde{V}_{ijt} > 0} \\ \left\langle \omega_{ijt}^i \Pr\left(\tilde{V}_{ijt} > 0 | \hat{\mu}_{ijt} \right) \right\rangle_{(i,j) \in N_t^0} \end{bmatrix}, \tag{A.50}$$

where $\Pr\left(\tilde{V}_{ijt} > 0 | \hat{\mu}_{ijt} \right)$ is from Eq. (A.24). The observation weights consist of (1) a vector of ones with length equaling the number of (i, j) pairs with positive observed visitor counts *that also* have positive simulated visitor counts, and (2) a vector of frequency-weighted probabilities that simulated visitor counts from the set of (i, j) pairs belonging to the stratified sample exceed 0.

The dependent variable in the WLS is defined as

$$\tilde{y}_{ijt} \equiv \sqrt{\tilde{\eta}_{ijt}} \odot \begin{bmatrix} \left\langle \log \tilde{V}_{ijt} \right\rangle_{(i,j) \in N_t^1 : \tilde{V}_{ijt} > 0} \\ \left\langle \mathbb{E}\left(\log \tilde{V}_{ijt} | \tilde{V}_{ijt} > 0, \hat{\mu}_{ijt} \right) \right\rangle_{(i,j) \in N_t^0} \end{bmatrix}, \tag{A.51}$$

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. (A.28). The dependent variable consists of (1) a weighted vector of log simulated visitor counts with length equaling the number of (i, j) pairs with positive observed visitor counts *that also* have positive simulated visitor counts, and (2) a weighted vector of mean log simulated visitor counts from the set of (i, j) pairs belonging to the stratified sample, conditional on the simulated visitor counts exceeding 0.

The independent variable in the WLS is defined as

$$\tilde{X}_{ijt} \equiv \begin{bmatrix} \sqrt{\tilde{\eta}_{ijt}}, & \sqrt{\tilde{\eta}_{ijt}} \odot \begin{pmatrix} \left\langle \log d_{ij} \right\rangle_{(i,j) \in N_t^1 : \tilde{V}_{ijt} > 0} \\ \left\langle \log d_{ij} \right\rangle_{(i,j) \in N_t^0} \end{pmatrix} \end{bmatrix}. \tag{A.52}$$

The independent variable consists of (1) a weighted vector of ones with length equaling the number of (i, j) pairs with positive observed visitor counts, (2) a weighted vector of associated log distances from the set of (i, j) pairs with positive observed visitor counts *that also* have positive simulated visitor counts, and (3) a weighted vector of associated log distances from the set of (i, j) pairs belonging to the stratified sample.

With these terms established, we set the simulated moment as

$$m_5(\tilde{v} | \psi) \equiv \text{Second element of } \left(\tilde{X}_{ijt}' \tilde{X}_{ijt} \right)^{-1} \left(\tilde{X}_{ijt}' \tilde{y}_{ijt} \right). \tag{A.53}$$

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

$$q_{ijt} = \left\langle \log V_{ijt} \right\rangle_{(i,j) \in N_t^1 : V_{ijt} > 4} \tag{A.54}$$

$$Z_{ijt} = \begin{bmatrix} \langle 1 \rangle_{(i,j) \in N_t^1 : V_{ijt} > 4}, & \left\langle \log d_{ij} \right\rangle_{(i,j) \in N_t^1 : V_{ijt} > 4} \end{bmatrix}. \tag{A.55}$$

The data moment is then

$$m_6(v) \equiv \text{Second element of } \left(Z_{ijt}' Z_{ijt} \right)^{-1} \left(Z_{ijt}' q_{ijt} \right). \tag{A.56}$$

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

$$\tilde{\xi}_{ijt} \equiv \begin{bmatrix} \langle 1 \rangle_{(i,j) \in N_t^1 : \tilde{V}_{ijt} > 4} \\ \left\langle \omega_{ijt}^i \Pr\left(\tilde{V}_{ijt} > 4 | \hat{\mu}_{ijt} \right) \right\rangle_{(i,j) \in N_t^0} \end{bmatrix}, \tag{A.57}$$

where $\Pr\left(\tilde{V}_{ijt} > 4|\hat{\mu}_{ijt}\right)$ is from Eq. (A.26). The dependent variable in the WLS is defined as

$$\tilde{q}_{ijt} \equiv \sqrt{\tilde{\xi}_{ijt}} \odot \begin{bmatrix} \left\langle \log \tilde{V}_{ijt} \right\rangle_{(i,j)\in N_t^1 : \tilde{V}_{ijt}>4} \\ \left\langle \mathbb{E}\left(\log \tilde{V}_{ijt} | \tilde{V}_{ijt} > 4, \hat{\mu}_{ijt}\right) \right\rangle_{(i,j)\in N_t^0} \end{bmatrix}, \tag{A.58}$$

where $\mathbb{E}\left(\log \tilde{V}_{ijt} | \tilde{V}_{ijt} > 4, \hat{\mu}_{ijt}\right)$ is from Eq. (A.29). Likewise, the independent variable in the WLS is defined as

$$\tilde{Z}_{ijt} \equiv \begin{bmatrix} \sqrt{\tilde{\xi}_{ijt}}, & \sqrt{\tilde{\xi}_{ijt}} \odot \begin{pmatrix} \left\langle \log d_{ij} \right\rangle_{(i,j)\in N_t^1 : \tilde{V}_{ijt}>4} \\ \left\langle \log d_{ij} \right\rangle_{(i,j)\in N_t^0} \end{pmatrix} \end{bmatrix}. \tag{A.59}$$

With these terms established, we set the simulated moment as

$$m_6\left(\tilde{v}|\psi\right) \equiv \text{Second element of } \left(\tilde{Z}'_{ijt}\tilde{Z}_{ijt}\right)^{-1}\left(\tilde{Z}'_{ijt}\tilde{q}_{ijt}\right). \tag{A.60}$$

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

$$\hat{m}\left(\tilde{v}|\psi\right) = \frac{1}{S}\sum_{S} m\left(\tilde{v}_s|\psi\right). \tag{A.61}$$

The final step of the MSM procedure is to find the estimate of β that minimizes the distance between the data moments and simulated model moments.

A.7 Construct the MSM estimator

The MSM estimator $\hat{\beta}_{\text{MSM}}$ 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 use the error function $e\left(\tilde{v}, v|\psi\right)$, which is the percent difference between the two vectors of moments:

$$e\left(\tilde{v}, v|\psi\right) \equiv \frac{\hat{m}\left(\tilde{v}|\psi\right) - m\left(v\right)}{m\left(v\right)}. \tag{A.62}$$

In this case, the MSM estimator is

$$\hat{\beta}_{MSM} = \underset{\beta}{\text{argmin}} \;\; e\left(\tilde{v}, v|\psi\right)' W e\left(\tilde{v}, v|\psi\right), \tag{A.63}$$

where W is an 8×8 weighting matrix that controls how each moment is weighted in the minimization problem. Notice that each candidate β in Eq. (A.63) 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 we found that approach to produce 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 $\hat{\beta}_{MSM}$ as

$$\widehat{\text{Var}}\left(\hat{\beta}_{MSM}\right) = \left(1 + \frac{1}{S}\right)\left[\frac{\partial \hat{m}\left(\tilde{v}|\psi\right)}{\partial \beta}' W^{-1} \frac{\partial \hat{m}\left(\tilde{v}|\psi\right)}{\partial \beta}\right]^{-1}, \tag{A.64}$$

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