Yale University

EliScholar – A Digital Platform for Scholarly Publishing at Yale

Cowles Foundation Discussion Papers

Cowles Foundation

7-10-2020

Measuring Movement and Social Contact with Smartphone Data: A Real-time Application to COVID-19

Victor Couture

Jonathan I. Dingel

Allison Green

Jessie Handbury

Kevin R. Williams

MEASURING MOVEMENT AND SOCIAL CONTACT WITH SMARTPHONE DATA: A REAL-TIME APPLICATION TO COVID-19

By

Victor Couture, Jonathan I. Dingel, Allison Green, Jessie Handbury, and Kevin R. Williams

July 2020

COWLES FOUNDATION DISCUSSION PAPER NO. 2241

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.yale.edu/

Measuring movement and social contact with smartphone data: a real-time application to COVID-19*

Victor Couture† Jonathan I. Dingel‡ Allison Green§

Jessie Handbury¶ Kevin R. Williams‖

3 July 2020

Abstract

Tracking human activity in real time and at fine spatial scale is particularly valuable during episodes such as the COVID-19 pandemic. In this paper, we discuss the suitability of smartphone data for quantifying movement and social contact. We show that these data cover broad sections of the US population and exhibit movement patterns similar to conventional survey data. We develop and make publicly available a location exposure index that summarizes county-to-county movements and a device exposure index that quantifies social contact within venues. We use these indices to document how pandemic-induced reductions in activity vary across people and places.

JEL codes: C8, R1, R4

*We are very grateful to Hayden Parsley and Serena Xu for outstanding research assistance under extraordinary circumstances. We thank Drew Breunig, Nicholas Sheilas, Stephanie Smiley, Elizabeth Cutrone, and the team at PlaceIQ for data access and helpful conversations. The views expressed herein are those of the authors and do not necessarily reflect the views of PlaceIQ, NBER, nor CEPR. This research was approved by the University of California, Berkeley Office for Protection of Human Subjects under CPHS Protocol No 2018-05-11122. This material is based upon work supported by the National Science Foundation under Grant No. 2030056, the Tobin Center for Economic Policy at Yale, the Fisher Center for Real Estate and Urban Economics at UC Berkeley, the Zell-Lurie Real Estate Center at Wharton, and the Initiative on Global Markets at Chicago Booth.

†University of British Columbia

‡University of Chicago Booth School of Business, NBER, and CEPR

§Princeton University

¶University of Pennsylvania and NBER

‖Yale School of Management and NBER

1 Introduction

Personal digital devices now generate streams of data that describe human behavior in great detail. The temporal frequency, geographic precision, and novel content of the "digital exhaust" generated by users of online platforms and digital devices offer social scientists opportunities to investigate new dimensions of economic activity. The COVID-19 pandemic has demonstrated the potential for real-time, high-frequency data to inform economic analysis and policymaking when traditional data sources deliver statistics less frequently and with some delay.

In this paper, we discuss the suitability of smartphone data for quantifying movement and social contact. We show that these data cover a significant fraction of the US population and are broadly representative of the general population in terms of residential characteristics and movement patterns. We use these data to produce a location exposure index ("LEX") that describes county-to-county movements and a device exposure index ("DEX") that quantifies the exposure of devices to each other within venues. These indices reveal substantial declines in inter-county travel and social contact in venues in March and April 2020. Compared to pre-pandemic levels, long-distance travel and the social contact of devices residing in more college-educated neighborhoods declined relatively more.

We publish these indices each weekday in a public repository available to all non-commercial users for research purposes.[1] Our aim is to reduce entry costs for those using smartphone movement data for pandemic-related research. By creating publicly available indices defined by documented sample-selection criteria, we hope to ease the comparison and interpretation of results across studies.[2] More broadly, this paper provides guidance on potential benefits and relevant caveats when using smartphone movement data for economic research.

[1]The indices and related documentation can be downloaded from `https://github.com/COVIDExposureIndices`.

[2]Examples of research using our indices thus far include Gupta, Nguyen, Rojas, Raman, Lee, Bento, Simon, and Wing (2020), Monte (2020), Yilmazkuday (2020b), and Yilmazkuday (2020a).

Researchers in economics and other fields are turning to smartphone movement data to investigate a great variety of social-science questions. Chen and Pope (2020) use similar smartphone data covering almost 2 million users in 2016 to document cross-sectional variation in geographic movement across cities and income groups. We focus on the distinctive advantages of these data's frequency and immediacy. A growing body of both theoretical and empirical research investigates human movement, social contact, and economic activity in the context of the COVID-19 pandemic.[3] Our indices provide empirical measures of these phenomena, complementing private-sector real-time measures of social distancing and movement.[4] We describe properties of smartphone data, compare the residential distribution and movement patterns of devices to those in traditional data sources, produce publicly available indices that can be used to easily compare results across studies, and investigate potential measurement issues that arise in the context of the ongoing pandemic.

2 Data

Our smartphone movement data come from PlaceIQ, a location data and analytics firm. In this section, we describe how PlaceIQ processes devices' movements to define visits to venues, and how we select the devices, venues, and visits included when we compute our exposure indices. We then compare these devices and their movements to residential populations and movements reported in traditional data sources.

[3] Among many others, see Greenstone and Nigam (2020) on the value of social distancing, Maloney and Taskin (2020) on private social distancing, Brzezinski, Deiana, Kecht, and Van Dijcke (2020) on the effect of government-ordered lockdowns, Engle, Stromme, and Zhou (2020) on correlates of observed social distancing, Farboodi, Jarosch, and Shimer (2020) on optimal policy, Monte (2020) on mobility zones, and Xiao (2020) on the value of contact-tracing apps.

[4] For example, Unacast reports distance traveled; Google's community mobility reports capture visits to different venue types; and SafeGraph reports time spent at and away from home. Relative to these measures, our indices are designed to summarize travel and overlapping visits relevant for COVID-19 circumstances in an IRB-approved public release.

2.1 Device Visit Data

PlaceIQ aggregates GPS location data from different smartphone applications using each device's unique advertising identifier. The raw GPS data come as pings that register whenever the application requests location data from the device.[5] These pings are joined with a map of two-dimensional polygons, corresponding to buildings or outdoor features such as public parks, which we denote "venues." A timestamped set of pings within or in the close vicinity of a polygon constitutes a "visit."[6] Since a device's location is measured with varying precision, PlaceIQ assigns each visit an attribution score based on ping characteristics and geographic features. We retain all visits with an attribution score greater than a minimum threshold. See Appendix A.1 for details.

2.2 Sample Selection

2.2.1 Devices covered

For the typical smartphone in the PlaceIQ data, we observe about six months of movements, but there is considerable heterogeneity across devices. Each Android and iOS smartphone has an identifier that uniquely identifies the device at any given time, and the device's unique advertising identifier can be refreshed by the user and may be refreshed by some system updates. Thus, the average lifespan of an advertising identifier is less than that of a physical phone. Even devices observed over a long time period may not ping regularly. Ping frequency reflects a device's applications, settings, and movements.

To focus on devices whose (non-)movements can be reliably characterized, we restrict the set of devices included in the computation of our indices to those that pinged on at least 11 days over any 14-day period from November 1, 2019 through

[5] The set of applications is not revealed to us. Some applications collect location data only when in active use, while others collect location data at regular intervals.

[6] If a device pings multiple times during a visit, then we have information about visit duration.

the reporting date. The earliest date for which we report our indices is January 20, 2020, so this criterion selects a set of devices based on a window of at least 80 days of prior potential activity. Later reporting dates have longer windows. Given the reduced movement associated with the COVID-19 pandemic, a criterion using a fixed window of prior potential activity would exclude devices that temporarily reduced their movements. As of June 4, 2020, 53 million devices met this device selection criterion. On any given day, about 20 million of these devices ping at least once.

For a subset of devices, we can assign a residential location with reasonable confidence, based on the duration of their residential visits since November 1, 2019. Appendix A.2 describes our home assignment algorithm. In short, we assign home locations based on where devices repeatedly spend time at night. We use Census-reported demographic characteristics for block groups, which contain about 600 to 3,000 people, as proxies for device demographics. Since many people temporarily moved to other residential locations during the pandemic, we assign a device to a block group of residence based on the block group of its first home location after November 1, 2019. As of June 4, 2020, 30 million devices have an assigned block group of residence.

In the context of the COVID-19 pandemic, a potential concern is that devices may not generate pings when "sheltering in place", due to their lack of movement. Indeed, there was a general decline in the number of devices generating pings in March 2020, presumably due to pandemic-induced declines in movement. When defining our exposure indices in the next section, we discuss how they are impacted by devices sheltering in place and suggest potential adjustments.

Even absent a pandemic, the number of devices appearing in the data varies meaningfully over time. That variation may reflect changes in smartphone ownership patterns, smartphone device settings, app usage, PlaceIQ app coverage, seasonal variation in behavioral patterns, or an Android or iOS operating system update. These are unlikely explanations for the general decline starting in March 2020, as that decline coincides with the COVID-19 outbreak in the United States and there has not been a major OS update or major shift in PlaceIQ app coverage since the beginning of 2020. When publishing our indices, we also publish the number of devices underlying these values so that researchers can assess when changes in the exposure indices may not reflect true changes in behavior.[7]

2.2.2 Venues covered

Venues include commercial establishments, public parks, residential locations, and polygons lacking an identified business category. When assigning devices' homes, only residential locations are relevant. When tracking devices' movements across geographic units in the LEX, visits to all such venues are informative.

When measuring potential social contact by the DEX defined in Section 3, we restrict attention to venue categories in which most venues are sufficiently small that visiting devices would be exposed to each other. In particular, we omit the categories "Residential", "Nature and Outdoor", "Theme Parks", "Airports", "Universities", as well as venues without a category identified by PlaceIQ. Finally, note that PlaceIQ excludes certain venue categories for privacy reasons, such as hospitals, schools, and places of worship.

The commercial categories included in our DEX calculations account for three-quarters of a million venues. Since a venue corresponds to a building, certain types of buildings can belong to multiple categories. For instance, a building with a coffee shop inside a book store would map to two categories (restaurant and retail). In most categories, the coverage of chains is high, but we observe a smaller share of independent businesses.[8] For instance, the largest category is restaurants,

[7]For example, the number of devices drops about 10 percent during April 14-18, 2020. In the absence of an obvious nationwide shock, this presumably reflects a change in smartphone data provision rather than a common change in behavior. Such variation will be absorbed by day fixed effects in difference-in-differences research designs.

[8]See Appendix C of Couture, Gaubert, Handbury, and Hurst (2020) for details.

which has about 200,000 distinct venues containing 370,000 restaurants.[9] Table A.2 reports the number of venues within each venue category in the DEX. There is little variation in the number of venues within January to June 2020.

2.2.3 Locations covered

We report our indices for all US states and most US counties. Many US counties have few residents and therefore few devices in the PlaceIQ data. The indices we report are restricted to counties with reasonably large device samples. To implement this restriction, we assign each device to a unique daily "residential county", where that device had the highest (cumulative) duration of time at residential locations on that date. We report our indices only for the 2,018 counties that were the residential county of at least 1,000 devices on every day from January 6 to 12, 2020.

2.3 Representativeness

Smartphone data cover a significant fraction of the US population. However, differences in smartphone ownership and app use, sample selection rules specific to research applications, and the use of small geographic units may produce unrepresentative samples.[10] For example, older adults are less likely to own smartphones, making smartphone-derived samples unbalanced across age groups.[11]

In this section, we compare the residential distribution and movement patterns of devices in our sample to those in traditional data sources. This analysis requires restricting our sample to devices assigned a residential block group, which

[9]US County Business Patterns reports there were about 570,000 establishments in NAICS 7225 in 2017.

[10]For instance, SafeGraph, another location data provider, found that about 10 percent of block groups contain 30 to 40 percent of the devices in their data, leading to "disproportionately and sometimes impossibly high" numbers of devices relative to the Census-reported residential population (Squire, 2019).

[11]The Pew Research Center estimates that 81 percent of US adults own a smartphone. That rate varies from 96 percent for ages 18-29 to only 53 percent for those over 65 years. See `https://www.pewresearch.org/internet/fact-sheet/mobile/`.

constitute about 80 percent of the devices in our sample.[12]

Panel A of Figure 1 shows that geographic units with larger residential population have more devices in our sample residing in them. Regressing the log number of devices on the US Census Bureau's 2019 estimate of log residential population yields an R^2 of 0.96 for states and 0.95 for counties. On average, the number of devices in our sample is about one-tenth of the total population.

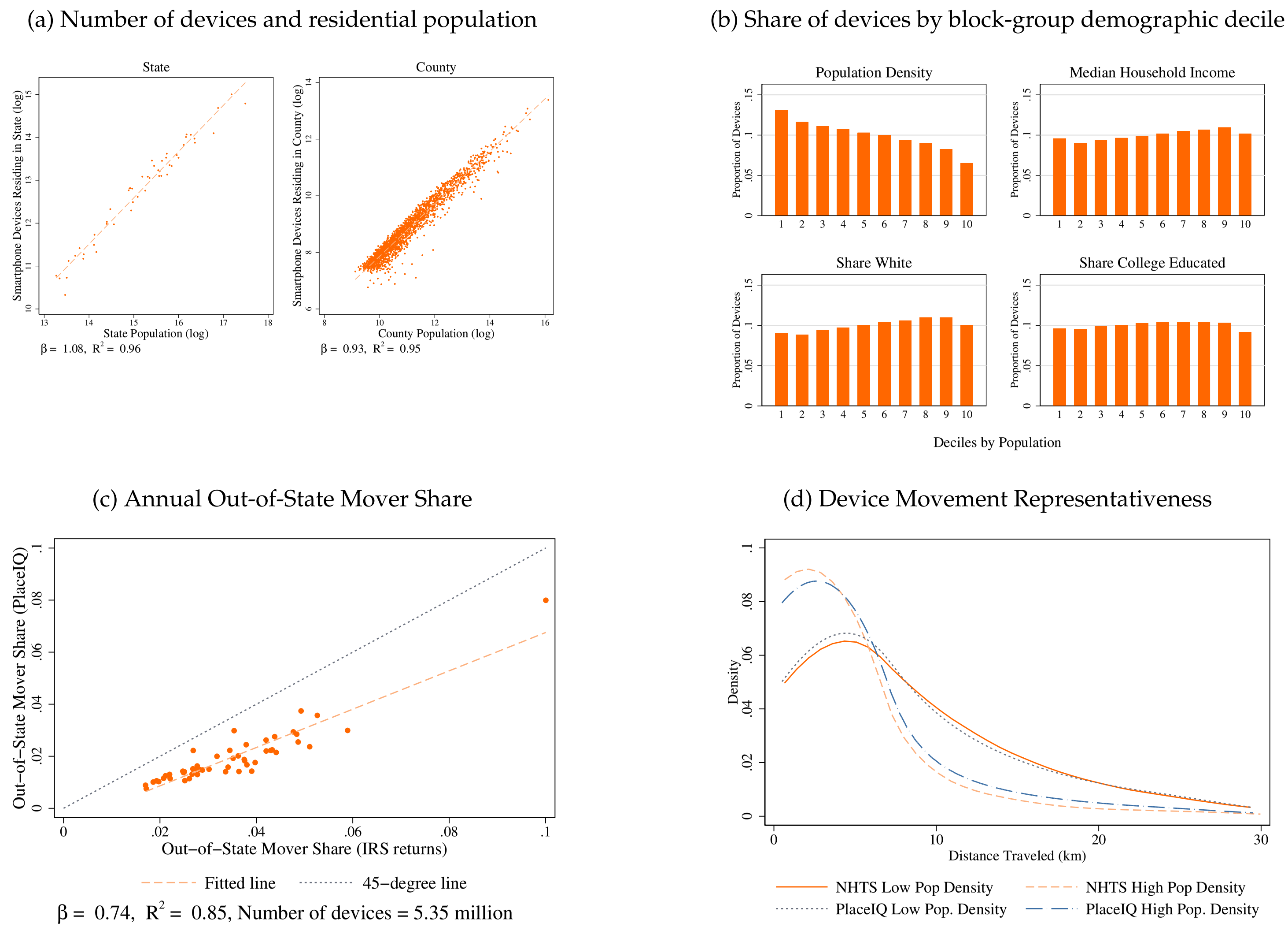
Panel B of Figure 1 investigates the distribution of devices across residential block groups within each county. The panel shows the share of devices living in block groups in ten population deciles ranked by income, share white, education, and population density. For instance, the top-right chart shows that about 10 percent of devices live in each decile of a county's block group median household income distribution. Similarly, about 10 percent of devices live in each decile when we rank block groups within their county by the share of their residents who are white or college graduates. When looking at deciles ranked by population density, denser block groups are somewhat underrepresented: only about 7 percent of devices live in block groups in the highest population-density decile.

In Appendix Figure B.1, we reproduce Panel B of Figure 1 using national population deciles instead of within-county population deciles. In that case, we find greater overrepresentation of block groups with low population densities and large shares of white residents.[13] Given that our sample is more representative within counties than across counties, we suggest that researchers focus on applications of our indices that rely on within-county variation or on intertemporal cross-county variation in relative changes. Applications relying on cross-county differences in levels may be prone to sample-selection bias.

[12] This restricted sample is the same that we will later use to compute our indices broken down by demographic group.

[13] When examining SafeGraph data, Squire (2019) reports the opposite pattern: SafeGraph data have fewer devices in block groups with more white residents. This suggests that representativeness may vary across smartphone data providers or sample-selection criteria.

Figure 1: Spatial and Demographic Balance of Device Populations



Panel A compares the number of devices residing in a geographic unit (vertical axis) to the Census's estimated 2019 residential population (horizontal axis) for all states, and for the 2,018 counties in the DEX and LEX. Panel B depicts the share of devices with residing in block groups in each within-county decile of population density, median household income, share of white residents, and share of residents over 25 years with a bachelor's degree or higher. These block group characteristics are from the 2014-2018 American Community Survey. Panel C compares, for each state, the share of devices that resided in another state over the last year (vertical axis) to the share of tax filers who reported an address in a different state on the previous year's tax return in the 2017-2018 IRS Migration Data (horizontal axis). Panel D depicts a kernel density plot of trip length in kilometers, for trips from home to a commercial venue in the PlaceIQ data and in the 2017 NHTS, for residents of block groups in the top and bottom quartile of the population-density distribution.

Panel C of Figure 1 examines residential migration. For each state, Panel C compares the share of devices that moved from another state during the prior year to the share of new residents in the 2017-2018 Internal Revenue Service (IRS) Migration Data. To facilitate this comparison, we restrict attention to the 5.4 million devices in the PlaceIQ data with non-missing home assignments in both the first and last week of 2019. Using this sample, we compute the share of devices in each state in the last week of 2019 that were residing in a different state in the first week of 2019. At the state level, this share of devices and the share of IRS-reported tax returns are highly correlated: regressing the PlaceIQ share on the IRS share yields an R^2 exceeding 0.8. At the county level, the correlation is considerably weaker, yielding an R^2 of only 0.15. This reflects in part smaller samples at the county level: if we restrict attention to counties with populations greater than 100,000, the R^2 increases to 0.25, and for county populations greater than 200,000 people, the R^2 rises further to 0.50.

Panel D of Figure 1 examines travel from home to commercial venues by depicting the distributions of trip lengths in our smartphone data and the 2017 National Household Transportation Survey (NHTS). For the PlaceIQ data, we show trips to venues included in the DEX computation.[14] For the NHTS, we show trips within the trip-purpose categories that most closely match DEX venues.[15] The figure depicts two trip-length distributions for each data source, one for people or devices living in block groups within the top quartile of the population density distribution, and one for people or devices living in the bottom quartile. The smartphone and NHTS trip-length distributions are remarkably similar, and both show a greater propensity to make shorter trips in more densely populated areas.

Overall, the patterns documented in Figure 1 suggest the potential of broadly

[14] A trip is from home if the device's previous visit was its home within the previous hour. We estimate driving distance (trip length) as 1.5 times the straight-line distance between the home and venue.

[15] These NHTS categories are "buy goods", "buy services", "buy meals", "other general errands", "recreational activities", and "exercise".

representative smartphone data for use in economic research. That said, we encourage researchers using these data to evaluate the precision and representativeness of their sample in their particular context. To help researchers assess whether our indices are suitably precise for their research application, we publish the underlying number of devices for each index, day, and geographic unit.

3 Exposure Indices

In this section, we describe how we compute the location exposure index, which measures state-to-state or county-to-county movement, and the device exposure index, which measures state- or county-level average exposure of devices to each other within commercial venues.

3.1 Notation and Preliminaries

We use the following notation when defining the LEX and DEX. Let i index devices, j index venues, g index geographic units (counties or states), and t and d index dates. Let $p_{ijt} \in \{0, 1\}$ and $p_{igt} \in \{0, 1\}$ be equal one if device i pinged in venue j or geography g, respectively, on date t. Define $p_{it} \equiv \max_g p_{igt}$ as an indicator that equals one if device i pinged in any geographic unit on date t. Let $r_{igt} \in \{0, 1\}$ be equal one when device i resided in g at date t, where we assign residence based on the geographic unit in which the device spent the most time in residential venues on that date.[16]

Next, we define sets of devices and venues based on these indicators. Let $\mathcal{I}_{j,d} \equiv \left\{i : p_{ijd} = 1\right\}$ and $\mathcal{I}_{g,d} \equiv \left\{i : p_{igd} = 1\right\}$ denote the sets of devices that pinged in venue j or geographic unit g, respectively, on date d. Let $\mathcal{G}_{g,d} \equiv \left\{i : r_{igd} = 1\right\}$ denote the set of devices that reside in geographic unit g on date d. Let $\mathcal{J}_{i,d} \equiv \left\{j : p_{ijd} = 1\right\}$ denote the set of venues where device i pinged on date d.

[16] In the event of a tie, the geographic unit of residence is assigned based on visits to non-residential locations.

3.2 Location Exposure Index (LEX)

The LEX is a matrix that answers the following query: Among smartphones that pinged in geographic unit g' on date d, what share of those devices pinged in geographic unit g at least once during the previous 14 days? We report the LEX as a daily $G \times G$ matrix, in which each cell reports, among devices that pinged on day d in the column location g', the share of devices that pinged in the row location g at least once during the previous 14 days (conditional on pinging anywhere during the previous 14 days). Thus, each element of this matrix is

$$\text{LEX}_{gg'd} \equiv \frac{\sum_{i \in \mathcal{I}_{g',d}} \mathbf{1}\left\{\sum_{t=d-14}^{d-1} p_{igt} > 0\right\}}{\sum_{i \in \mathcal{I}_{g',d}} \mathbf{1}\left\{\sum_{t=d-14}^{d-1} p_{it} > 0\right\}} = \frac{\sum_i \mathbf{1}\left\{i : \left(p_{ig'd} = 1 \text{ \& } \sum_{t=d-14}^{d-1} p_{igt} > 0\right)\right\}}{\sum_i \mathbf{1}\left\{i : \left(p_{ig'd} = 1 \text{ \& } \sum_{t=d-14}^{d-1} p_{it} > 0\right)\right\}}.$$

As an example, if g' is New York County, NY and g is Suffolk County, NY, then $\text{LEX}_{gg'd}$ is the share of devices pinging in Suffolk County on day d that also pinged in New York County over the last 14 days (conditional on pinging anywhere in the US in the last 14 days).

We define the LEX to summarize people's movements with pandemic-related applications in mind. The index describes the share of people in a given location who have been in other locations during the prior two weeks. Thus, if COVID-19 cases surge in county g, $\text{LEX}_{gg'd}$ describes the potential exposure of county g' to the infectious disease via prior human movement from county g to g'. We chose the 14-day period of exposure based on the incubation period commonly cited by public-health authorities during the ongoing pandemic.[17] We chose to focus on all devices pinging in a given location rather than only residents because all human movement is relevant for potential disease exposure. Similarly, LEX_d is not a transition matrix and its columns do not sum to one because a device can

[17] The CDC's COVID-19 FAQ page: "Based on existing literature, the incubation period (the time from exposure to development of symptoms) of SARS-CoV-2 and other coronaviruses (e.g. MERS-CoV, SARS-CoV) ranges from 2–14 days."

visit multiple location during the 14-day period. The temporal frequency and geographic units were selected to protect device user privacy in the context of a public data release. For other research applications, the appropriate length of exposure or geographic units may vary.

Starting in March 2020, there was a general decline in the number of devices generating pings, presumably due to individuals restricting their movements in response to the pandemic. Both the numerator and denominator of $\text{LEX}_{gg'd}$ restrict attention to devices that ping in g' on day d ($i \in \mathcal{I}_{g',d}$), so the LEX captures the locational histories of devices that are "out and about" in geographic unit g' on date d and does not capture the locational histories of devices sheltering-in-place and not generating any pings. This seems the relevant notion of potential exposure in the context of the ongoing pandemic: the index captures non-local exposure associated with "active" devices that are moving around within location g'. For applications that require measuring exposure for the entire population of devices, including those that do not generate pings, we have published the daily number of devices that ping in each county, so that researchers can adjust their computations.

3.3 Device Exposure Index (DEX)

The DEX is a county- or state-level scalar that answers the following query: How many distinct devices does the average device living in g encounter via overlapping visits to commercial venues on each day? To compute the DEX, we first calculate the daily exposure set of device i as the number of distinct other devices that visit any commercial venue that i visits on date t:

$$\text{EXP}_{i,d} = \bigcup_{j \in \mathcal{J}_{i,d}} \mathcal{I}_{j,d}.$$

The DEX is then defined as the average size of the exposure set for devices that reside in geographic unit g on date d:

$$\mathrm{DEX}_{g,d} \equiv \frac{1}{|\mathcal{G}_{g,d}|} \sum_{i \in \mathcal{G}_{g,d}} |\mathrm{EXP}_{i,d}|.$$

Note that the DEX values are necessarily only a fraction of the number of distinct individuals that also visited any of the commercial venues visited by a device, since only a fraction of individuals, venues, and visits are in the device sample.

We have defined the DEX to summarize social contact with pandemic-related applications in mind. The index captures overlapping visits to venues on the same day, which is relevant for potential virus exposure. We chose to define overlapping visits as visits to a venue on the same day rather than during the same hour based on both sample size and the concern that SARS-CoV-2 can persist in circulating air and on surfaces for multiple hours. The geographic units were selected to protect user privacy in the context of a public data release.

Note that devices sheltering in place would drop out of the sample used to compute the DEX if they did not generate any pings. As a result, the DEX may underestimate the reduction in exposure following the COVID-19 outbreak. We therefore implement a simple adjustment of the $\mathrm{DEX}_{g,d}$ denominator as one means of addressing the potential sample selection problem associated with devices sheltering-in-place. Define a counterfactual set of pinging devices $\mathcal{G}^*_{g,d}$ such that any device in $\mathcal{G}^*_{g,d}$ but not in the observed $\mathcal{G}_{g,d}$ is sheltering in place with $|\mathrm{EXP}_{i,d}| = 0$. The adjusted DEX is

$$\mathrm{DEX}^{\mathrm{adjusted}}_{g,d} = \frac{|\mathcal{G}_{g,d}|}{|\mathcal{G}^*_{g,d}|}\mathrm{DEX}_{g,d}.$$

We assign the counterfactual set $\mathcal{G}^*_{g,d}$ to be the largest number of devices observed on any day from January 20, 2020 to February 14, 2020 in geographic unit g, so that:

$$\widehat{|\mathcal{G}^*_{g,d}|} = \max_{d \in [20 \text{ Jan } 2020, 14 \text{ Feb } 2020]} |\mathcal{G}_{g,d}|$$

Given that $|\widehat{\mathcal{G}^*_{g,d}}|$ is an upper bound, $\text{DEX}^{\text{adjusted}}_{g,d}$ likely overestimates the drop in exposure following the COVID-19 outbreak. On the other hand, as noted above, the unadjusted $\text{DEX}_{g,d}$ likely underestimates the drop in exposure.[18] Together, these series should offer useful bounds. As mentioned before, even absent a pandemic there is meaningful variation in the number of devices in the sample that affect the DEX.

For devices that have a home assigned, we compute DEX values by the demographic characteristics of their residential block group. We only report these demographic DEX values at the state level, due to sample size and privacy considerations.

DEX by income Within each state g, we partition all census block groups into four median income quartiles with an equal number of block groups. We index these quartiles by $q \in \{1, 2, 3, 4\}$. Within each state g on each day d, we denote by $\mathcal{G}_{g,q,d}$ the set of devices i that have a home in a block group within quartile q.[19] The DEX by income is then:

$$\text{DEX-income}_{g,q,d} = \sum_{i \in \mathcal{G}_{g,q,d}} \frac{\text{EXP}_{i,d}}{|\mathcal{G}_{g,q,d}|}$$

DEX by education The DEX by education is the same as the DEX by income, except that the four quartiles are based on the college share within each block

[18]In practice, while the average absolute difference between the state-level unadjusted and adjusted DEX values is 7 percent, the two indices have a correlation coefficient of 0.996 in levels and 0.992 in first differences. Figure B.2 shows that the population-weighted median values of the unadjusted and adjusted DEX track each other closely over time. The adjusted DEX should not be used when $|\mathcal{G}_{g,d}| > |\mathcal{G}^*_{g,d}|$, which will occur as social contact resumes and devices stop sheltering in place.

[19]Note that the residential block group is not necessarily within geographic-unit-of-residence g. This allows for cases where a device leaves their assigned home to shelter in place somewhere else. For example, if a device's home is in a block group in New York corresponding to the bottom income quartile, and it moves to Pennsylvania to shelter in place, that device is still assigned to the first income quartile but its state of residence changes to Pennsylvania.

group.[20]

DEX by race/ethnicity We report DEX values by racial/ethnic categories available in the Census of Population. For each $r \in \{\text{Asian, Black, Hispanic, White}\}$, we report a weighted average of device-level exposure,

$$\text{DEX-race}_{g,d,k} = \sum_{i \in \mathcal{G}_{g,q,d}} \frac{w_{i,r}\text{EXP}_{i,d}}{\sum\limits_{i \in \mathcal{G}_{g,q,d}} w_{i,r}},$$

where $w_{i,r}$ is the residential share of race/ethnicity r in device i's block group.[21]

4 Tracking activity during the 2020 pandemic

We now use the LEX and the DEX to document pandemic-induced reductions in activity during 2020 and explore how they vary across people and places.

4.1 Reduced movement between US counties

To illustrate the movement detail captured by the county-to-county LEX, in Figure 2 we plot the fraction of devices that pinged in Manhattan (New York County), one of the early US epicenters of the pandemic. The maps depict the share of devices in each US county that had pinged in Manhattan during the previous two weeks on the last Saturday of February, March, April, and May 2020. The February panel shows a clear role for physical distance, as counties closer to Manhattan typically have a larger share of devices that have been in Manhattan during the previous two weeks, but it also makes clear that physical distance and county-to-county

[20] The college share is the share of adults 25-65 years old with at least a four-year college degree.

[21] To be precise, the categories "Asian," "Black," "Hispanic," and "White" are shorthand for non-Hispanic Asian, non-Hispanic black, all Hispanic, and non-Hispanic white residents. These four categories are sufficiently large to be reported for many geographic units. In a few states, the number of recorded devices is low for some of these four racial/ethnic groups. We only report the DEX-race for a given racial/ethnic group in states where the weighted number of devices for that group is at least 1,000 devices every day from January 6 to 12, 2020.

movements are distinct. These measures of county-to-county movement should be more useful than physical distance in applications describing person-to-person economic linkages and disease spread.

The LEX reveals a swift decline in travel between New York County and other counties over the course of March 2020. While Figure 2 suggests a broad decline in the share of devices that had been in New York County during the previous two weeks, the decline appears greater in counties farther from New York City. Movements connected to New York County became more spatially concentrated by late April. A modest increase in inter-county travel is visible by late May.

Figure B.3 provides a contrasting example, depicting counties' exposure to Houston, Texas (Harris County). In that case, although there is a sizable decline in the shares of devices on the east coast that have recently been in Houston, travel from Houston to southern and southwestern counties shows little to no decline. Because the county-to-county LEX matrix reports more than 4 million values for each day, maps like those in Figure 2 and B.3 offer only a glimpse of the movement patterns captured by these data.

To summarize daily LEX values for the entire United States, Figure 3 depicts changes in state-level LEX values by physical proximity. We group pairs of states based on the distance between them and compute the daily mean value of $LEX_{gg'd}$ for each group. For example, the shortest-distance group consists of all states g and g' such that the distance between the population-weighted centroids of g and g' are less than 100 miles apart. The longest-distance group consists of state pairs with population-weighted centroids more than 1,500 miles apart. To illustrate relative declines, Figure 3 depicts the mean daily LEX value for each distance-defined group of state pairs relative to its value on March 1, 2020.

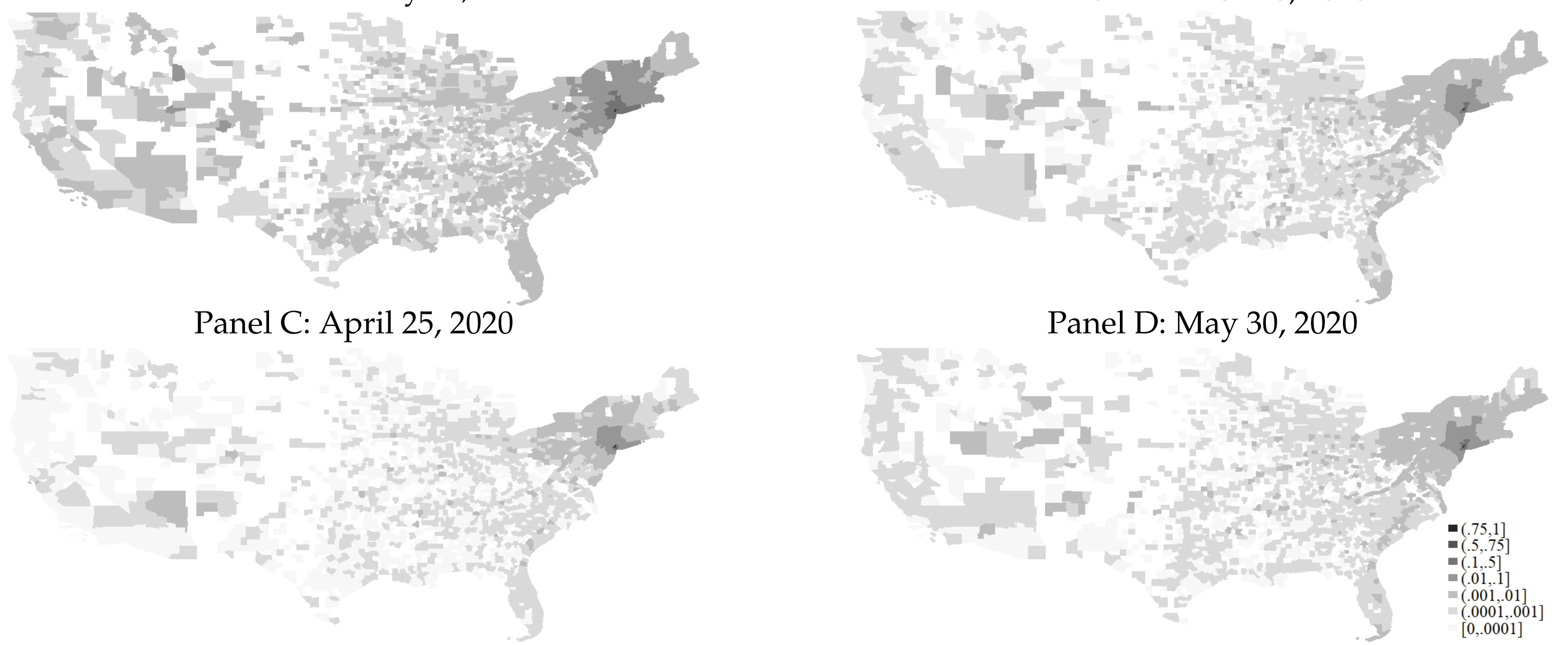
Figure 2: County-Level Exposure to New York County (Manhattan)

Panel A: February 29, 2020

Panel B: March 28, 2020

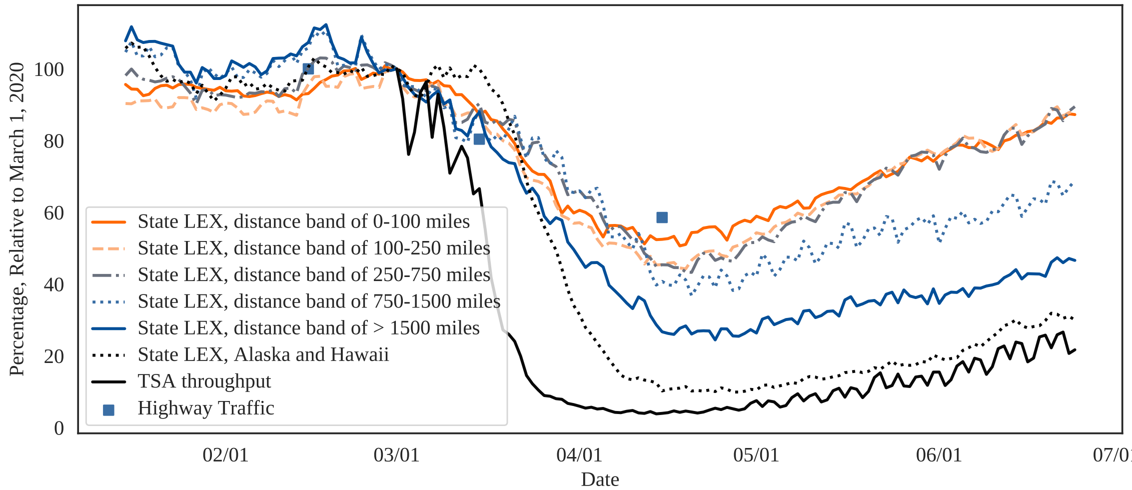
Panel C: April 25, 2020

Panel D: May 30, 2020



Notes: Each panel of this figure depicts, for each of 2,018 counties, the share of devices pinging in that county that had pinged in New York, New York during the previous 14 days. The four panels depicts this for four Saturdays in early 2020. Using the notation of Section 3, the three panels depict $LEX_{36061,g',d}$ for d equal to February 29, March 28, April 25, and May 30 of 2020, where 36061 is the FIPS code for New York County.

Figure 3: State-level LEX values by distance between states



Notes: This figure depicts average LEX values for pairs of states grouped by the distance between their population-weighted centroids. Each series is normalized relative to its value on March 1, 2020. The TSA throughput series reports the number of travelers passing through TSA checkpoints on each day.

Although the average LEX value declines for all state pairs through late April, pairs of states that are farther apart tended to exhibit larger relative declines. By mid-April, state-level LEX values at all distances were down 40 percent relative to their earlier levels. For comparison, monthly total vehicle-miles traveled, a measure that reflects both intrastate and interstate travel, fell by about 40 percent from February to April.[22] The steepest decline observed is for state pairs that include Alaska or Hawaii where across-state movements depend heavily on air travel.[23] This decline, which was down about 90 percent by mid-April, closely tracks the decline in daily checkpoint totals at US airports reported by the Transportation Security Administration (TSA) two weeks earlier, as the LEX captures inter-state movements using a fourteen-day window. Inter-state travel at all distances began

[22]We computed this figure using monthly seasonally adjusted vehicle-miles-traveled estimates from the Federal Highway Administration (series TRFVOLUSM227SFWA at `https://fred.stlouisfed.org`). Note that total distance traveled and the notion of exposure captured by the LEX are distinct concepts.

[23]Alaska and Hawaii are both at least 1,400 miles from every other US state.

to rise in late April 2020.

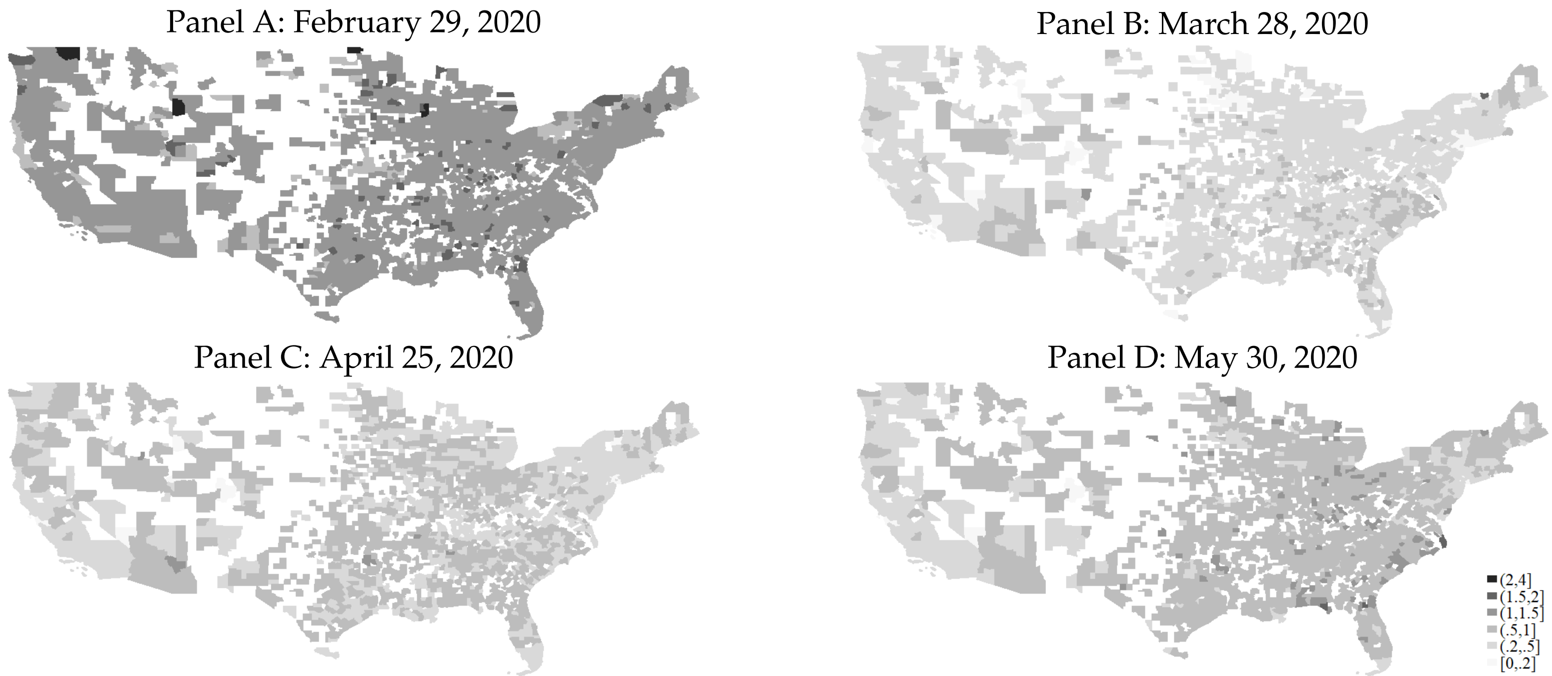
4.2 Reduced overlapping visits to venues

Figure 4 maps the county-level DEX on the last Saturday of February, March, April, and May 2020 relative to its level on Saturday, February 1. Panel A shows a rise in activity nationwide in late February. The median county saw a 20 percent increase in the DEX between February 1 and February 29. A similar relative uptick in activity in February 2019 suggests this increase is likely regular seasonal variation rather than a pandemic-induced shift. Panel B shows a reduction in activity over the subsequent four weeks in all but two counties. On March 28, the median county's DEX was just 35 percent of its February 1 level.[24] Panel C shows that by late April, activity had increased across much of the country, though even in late May (Panel D) it remained lower than it was in early February, by more than a factor of two in the greater New York City area, California, Washington, and the southern tip of Florida. The counties that saw outsized growth in activity by late May are often summer vacation destinations, such as Dare County, NC (containing the Outer Banks) and Bay County, FL (containing Panama City).

Some of this variation might be explained by policy differences. Appendix Figure B.5 depicts the evolution of the county-level DEX around policy events after controlling for county and time fixed effects. As in Brzezinski, Deiana, Kecht, and Van Dijcke (2020), we find that some of the DEX decline coincided with the timing of shelter-in-place orders, after which the DEX dropped by approximately 20 percent. A similar event study suggests a more modest and gradual increase in activity following the re-opening of non-essential businesses, with the DEX increasing by less than 10 percent relative to its pre-opening level a week after the event. We note that given how many forces are simultaneously impacting people's movement during the pandemic, these simple regressions are necessarily only suggestive.

[24]Figure B.4 plots the population-weighted median and interquartile range of the DEX over time.

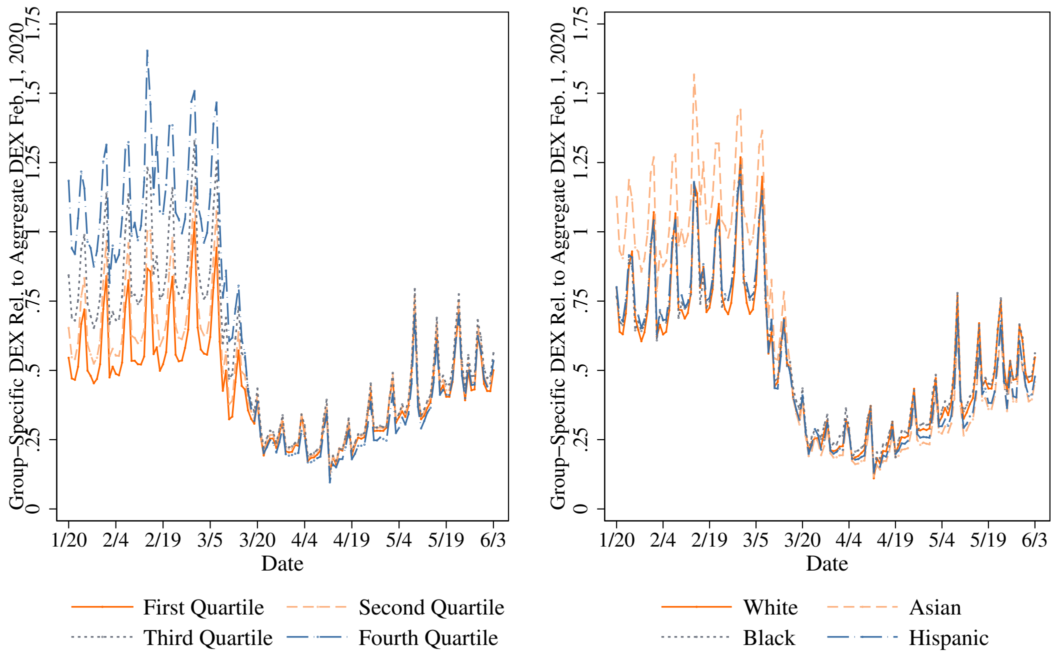
Figure 4: County-Level Device Exposure (DEX) Normalized to Level on February 1, 2020



Notes: This figure shows the county-level average device exposure (DEX) on a given date (February 29 in Panel A, March 28 in Panel B, April 25 in Panel C, and May 30 in Panel D) as a percentage of its level on February 1, 2020.

Figure 5: DEX values by block-group demographics

(a) by educational attainment (b) by racial/ethnic demographics



NOTES: These plots depict the state-level DEX by demographic groups. For each state, the demographic DEX time series is divided by the level of the aggregate DEX on February 1, 2020. The depicted series is a device-weighted average over all states. Panel A depicts this series for DEX by education and Panel B depicts this series for DEX by race/ethnicity as defined in Section 3.

Figure 5 reveals variation in the reduction in activity across educational attainment and race. Panel A depicts each DEX-education quartile relative to the aggregate DEX on February 1. Prior to the onset of COVID-19 in the U.S., residents of block groups with more college graduates were more exposed to other devices than average.[25] In March, exposure fell for residents of all block groups, but residents of block groups with more college graduates exhibited a proportionately greater decline. As a result, by the end of March 2020, there was little discernible difference in device exposure across neighborhoods with different shares of college graduates. After converging, device exposure remained at low levels through

[25]This is consistent with the finding that devices from higher-income neighborhoods visit more places (Chen and Pope, 2020).

April and May, at roughly one-third of the pre-pandemic average. This represents a 70 percent decline in the DEX for devices residing in block groups above the median college-graduate share and a 55 percent decline in the DEX for devices in below-median block groups.

Panel B of Figure 5 depicts device exposure by racial/ethnic demographics. Prior to the pandemic, devices living in block groups with more Black, Hispanic, and White residents had similar levels of exposure, while devices living in block groups with more Asian residents had higher DEX values. From mid-March onwards, all four demographic groups exhibited similarly low exposure levels.

The limited variation in device exposure across different demographic groups after March 15 may imply a limited role for heterogeneous exposure rates in explaining differences in these demographic groups' infection and mortality rates during the pandemic. Researchers investigating these questions could combine these local measures of social contact by demographic traits with other observed demographic differences that may explain disparate outcomes.

These initial applications of our indices demonstrate the potential of smartphone movement data to quantify movement and social contact with high frequency and spatial precision. We have also articulated a number of caveats relevant for researchers using such data. We hope that our publicly available indices will support deeper and varied investigation of human movement during the ongoing pandemic.

References

BRZEZINSKI, A., G. DEIANA, V. KECHT, AND D. VAN DIJCKE (2020): "The covid-19 pandemic: government vs. community action across the united states," Discussion paper, CEPR.

CHEN, M. K., AND D. G. POPE (2020): "Geographic Mobility in America: Evidence from Cell Phone Data," Working Paper 27072, National Bureau of Economic Research.

COUTURE, V., C. GAUBERT, J. HANDBURY, AND E. HURST (2020): "Income Growth and the Distributional Effects of Urban Spatial Sorting," .

ENGLE, S., J. STROMME, AND A. ZHOU (2020): "Staying at home: Mobility effects of Covid-19," Discussion paper, Covid Economics: Vetted and Real Time Papers.

FARBOODI, M., G. JAROSCH, AND R. SHIMER (2020): "Internal and External Effects of Social Distancing in a Pandemic," mimeo.

GREENSTONE, M., AND V. NIGAM (2020): "Does social distancing matter?," Discussion Paper 2020-26, University of Chicago, Becker Friedman Institute for Economics.

GUPTA, S., T. D. NGUYEN, F. L. ROJAS, S. RAMAN, B. LEE, A. BENTO, K. I. SIMON, AND C. WING (2020): "Tracking public and private response to the covid-19 epidemic: Evidence from state and local government actions," Discussion paper, National Bureau of Economic Research.

MALONEY, W. F., AND T. TASKIN (2020): "Determinants of social distancing and economic activity during COVID-19: A global view," Discussion Paper 9242, World Bank, World Bank Policy Research Working Paper.

MONTE, F. (2020): "Mobility Zones," mimeo.

RAIFMAN, J., K. NOCKA, D. JONES, J. BOR, S. LIPSON, J. JAY, AND P. CHAN (2020): "COVID-19 US state policy database," Boston, MA: Boston University.

SQUIRE, R. F. (2019): "Quantifying Sampling Bias in SafeGraph Patterns," https://colab.research.google.com/drive/1u15afRytJMsizySFqA2EPlXSh3KTmNTQ.

XIAO, K. (2020): "Saving Lives versus Saving Livelihoods: Can Big Data Technology Solve the Pandemic Dilemma?," Discussion paper, Available at SSRN 3583919.

YILMAZKUDAY, H. (2020a): "COVID-19 and Unequal Social Distancing across Demographic Groups," Discussion paper, Available at SSRN 3580302.

——— (2020b): "COVID-19 Deaths and Inter-County Travel: Daily Evidence from the US," Discussion paper, Available at SSRN 3568838.

Appendix – For Online Publication

A Data appendix

A.1 Smartphone visits data

Each observed visit consists of a device, a venue, a timestamp, and an attribution score. PlaceIQ's attribution scores are larger when a device is more likely to have been within a venue, based on the number and density of pings, data source of pings, and proximity of the pings to the polygon defining the venue. We retain all visits with an attribution score greater than a threshold value recommended by PlaceIQ based on their experience correlating their data to a diverse array of truth sets, including consumer spending data and foot-traffic counts. PlaceIQ also reports a lower bound for the visit's duration based on the time between consecutive pings at the same venue.

We also clean the visit data to remove simultaneous visits. For instance, when two venues are in close proximity to one other, a single visit event may have an attribution score for both venues that exceeds the threshold value recommended by PlaceIQ. We retain only the visit to the venue with the highest attribution score. In other cases, the polygons of two different venues overlap.[26] When two polygons overlap, we retain polygons with an identified business category over those lacking a category.

Table A.1 summarizes the smartphone movement data after this cleaning for days between January 20 and March 1, 2020. On the average day, there were 176 million visits produced by 33 million devices visiting 40 million residential and non-residential venues. The average device appears in the data for 25 days between January 20 and March 1, but a notable number appear on only one day.

[26] This could happen, for instance, if the basemap contains one polygon representing a business establishment and a second polygon representing both that building and the accompanying parking lot.

After we apply the device selection criteria we use when computing the LEX and DEX indices (devices that pinged on at least 11 days over any 14-day period from November 1, 2019 through the reporting date), there are 152 million visits from 23 million devices visiting 37 million venues on an average day. The selected devices appear in the data between January 20 and March 1 for 35 days on average.

Table A.1: Summary statistics for cleaned visits and indices samples

	Cleaned visits sample				Indices sample			
	Mean	SD	5th	95th	Mean	SD	5th	95th
Devices	33.43	1.92	31.15	36.58	22.80	0.49	22.05	23.61
Venues	40.46	0.81	39.17	41.51	36.88	0.92	35.35	38.28
Visits	175.85	11.33	154.15	191.12	151.56	11.30	132.59	166.74
Duration	25.81	14.31	1.00	41.00	34.91	9.89	11.00	41.00

NOTES: This table summarizes PlaceIQ data for January 20, 2020 to March 1, 2020 after our cleaning of the visits as described in the text. The counts of devices, venues, and visits are stated in millions per day. Duration is the number of days between a device's first and last appearance in the data (between January 20 and March 1).

A.2 Home assignments

Residential venues are a distinct category in the PlaceIQ data. This allows us to construct a weekly panel of home locations for a subset of devices using the following assignment methodology:

1. For each week, we assign a device to the residential venue where its total weekly visit duration at night (between 5pm and 9am) is longest, conditional on it making at least three nighttime visits to that venue within the week.[27] If a device does not visit any residential location on at least three nights, then on initial assignment that device-week pair has a missing residential location.

[27]Since we only observe minimum duration, there are instances where total duration is 0 across all residential locations. In these cases, we assign the residential venue as the venue a device makes the most nighttime visits.

Table A.2: Venue categories in DEX

Retail	209,274
Restaurants	200,839
Gas Station/Convenience Stores	118,307
Night Clubs/Bars	88,784
Banks	79,150
Shipping	36,745
Hotels	32,303
Home Improvement Stores	27,097
Grocery Stores	25,770
Financial Services	23,238
Pharmacies	22,408
Car Dealerships	20,644
Beauty Stores	15,556
Big Box Stores	11,558
Real Estate Offices	9,732
Gyms	9,289
Car Rental	8,999
Pay Day Loan	6,043
Storage	5,935
Movie Theaters	4,632
Library	1,962
Liquor Stores	1,193

NOTES: This table lists the venue categories that enter the computation of the Device Exposure Index (DEX) and shows the total number of distinct venues on 30 June 2020 in each category. Some venues belong to multiple categories, so the number of distinct venues (about three-quarters of a million) is smaller than the sum of all rows in this table.

2. After this preliminary assignment, we fill in missing weeks and adjust for noisiness in the initial panel using the following interpolation rules:

Rule 1: *Change "$X\ \cdot\ X$" to "$X\ X\ X$"*: If the residential assignment for a week is missing and the non-missing residential assignment in the weeks before and after is the same, we replace the missing value with that residential assignment.

Rule 2: *"$a\ X\ Y\ X\ b$" to "$a\ X\ X\ X\ b$" where $a \neq Y$ and $b \neq Y$*: If a device has a residential assignment Y that does not match the assignment X in the week before or after, we replace Y with X as long as Y was not the residential assignment two weeks before or two weeks after.[28]

3. After step 2's interpolation, for any spells of at least four consecutive weeks where a device is assigned the same residential venue, we assign that venue as a device's "home" for those weeks. Spells of less than four weeks are set to missing.

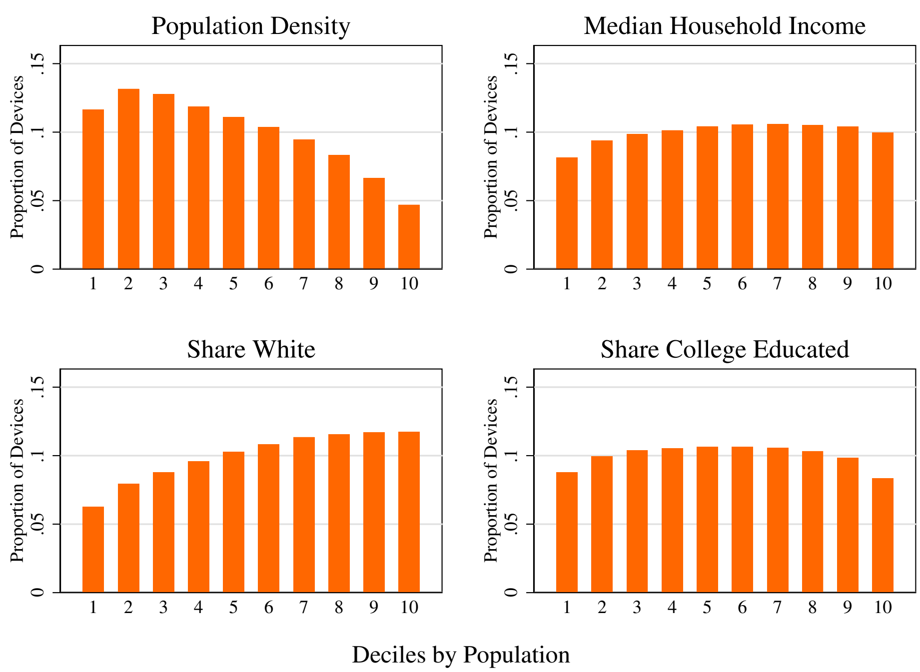
4. If a device has more than one home assignment and the pairwise distance between them is less than 0.1 kilometers, we keep the home that appears for the most weeks.

5. If a device has the same home assignment in two non-consecutive periods and no other home assignments in between, then we assign all weeks in between to that home assignment.

[28]For cases where a device's residential location is bouncing between two places ("Y X Y X X") we are not able to ascertain whether Y or X is more likely to be a device's residence in a given week

B Figures appendix

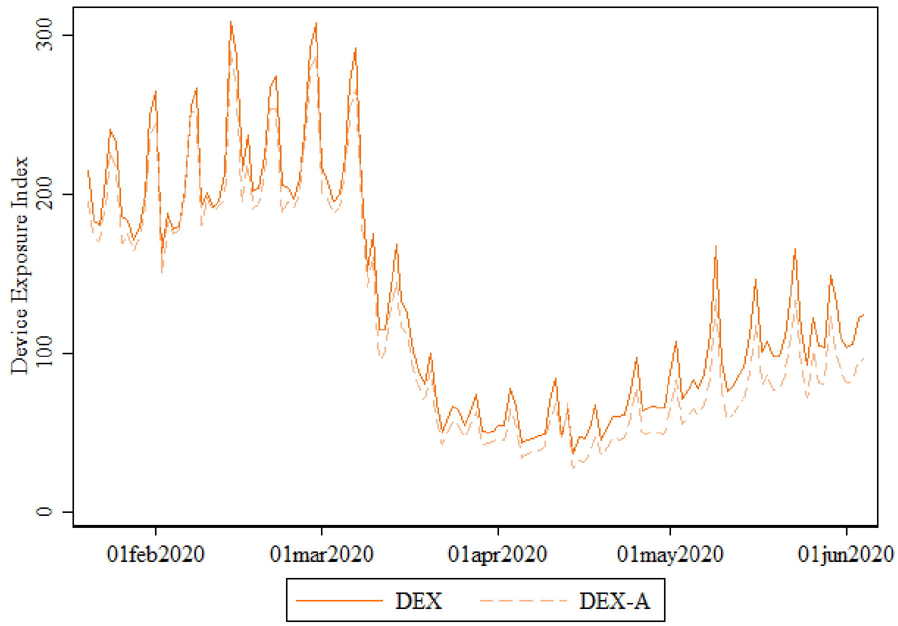
Figure B.1: Balance of devices' residences across block groups by national demographic deciles



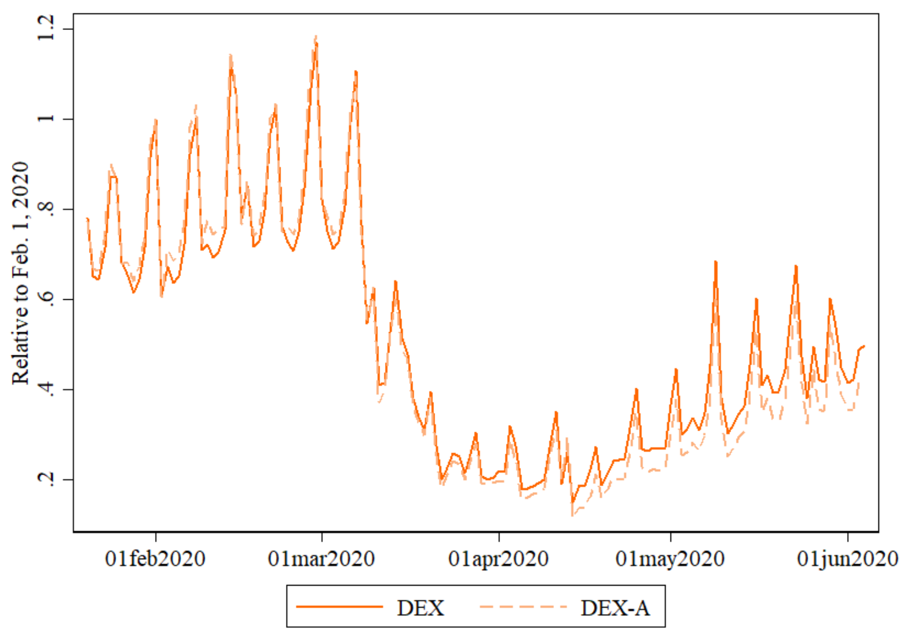
Notes: This figure shows the total share of devices living in census block groups corresponding to the national deciles for each of the four demographic categories.

Figure B.2: DEX and DEX-A over time

Panel A: Raw Device Exposure Index

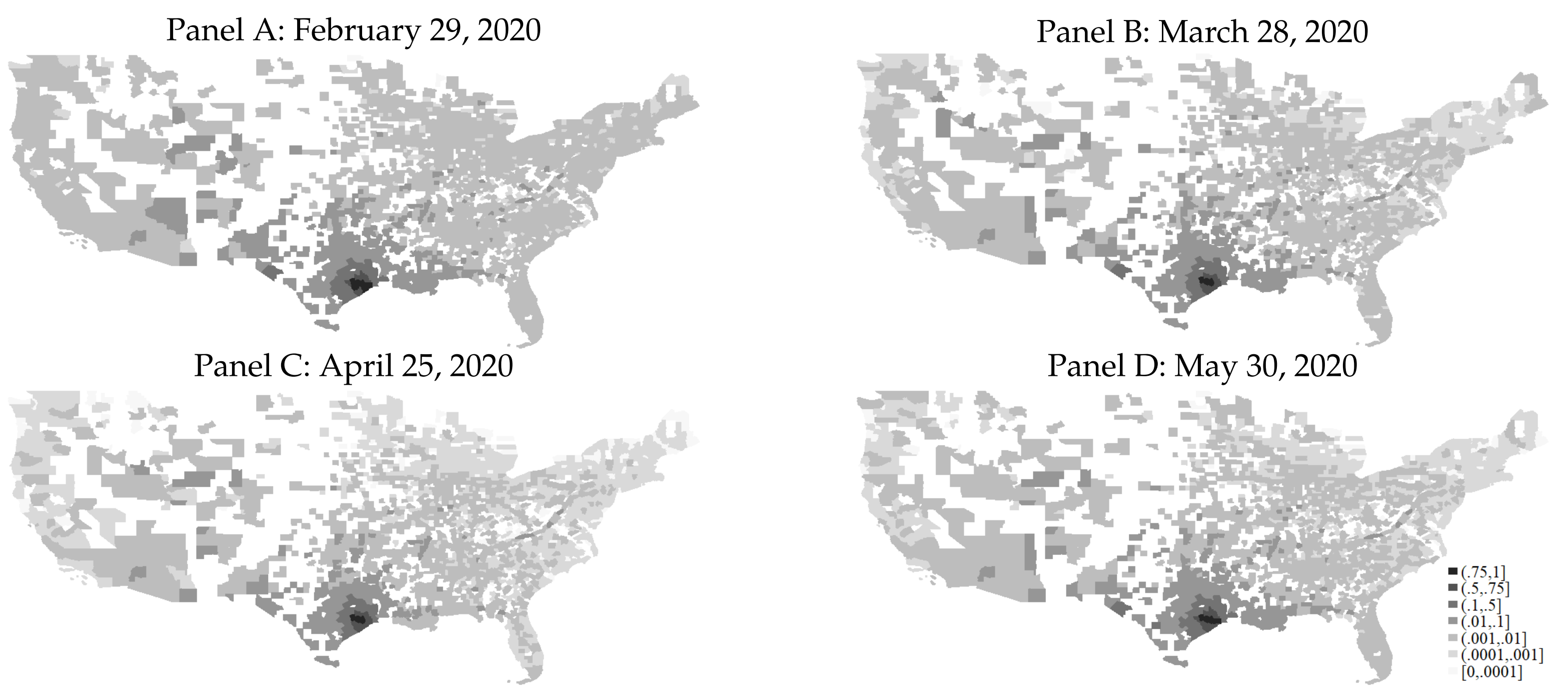


Panel B: Normalized to February 1, 2020



Notes: This figure shows the population-weighted median unadjusted and adjusted device exposure indices (DEX and DEX-A) over time.

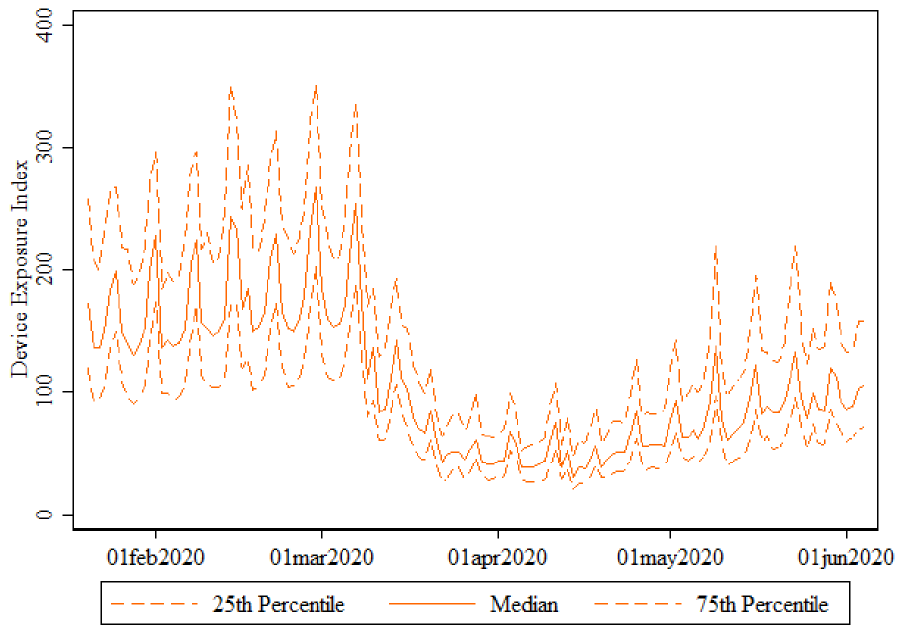
Figure B.3: County-Level Exposure to Harris County (Houston)



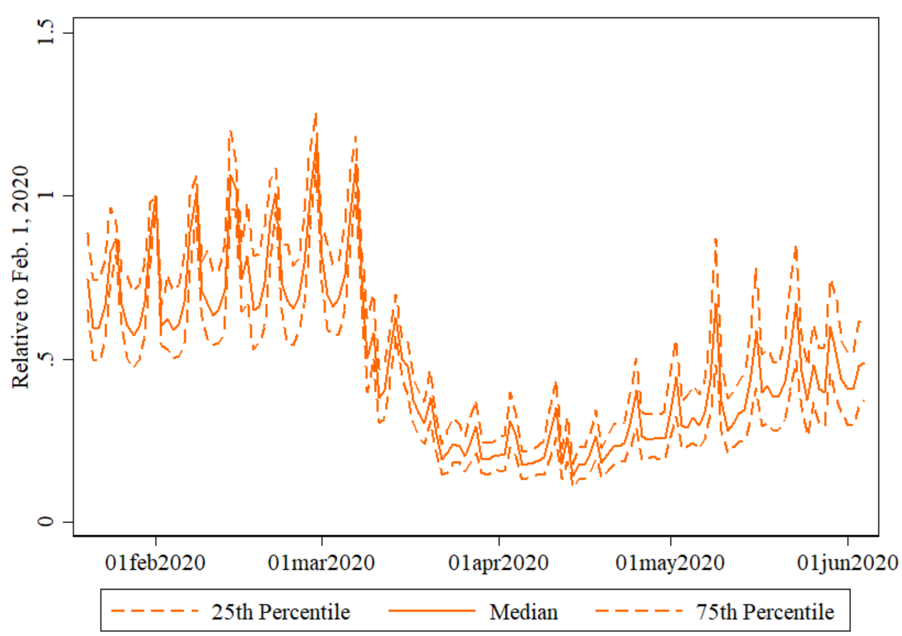
Notes: Each panel of this figure depicts, for each of 2,018 counties, the share of devices pinging in that county that had pinged in Harris County, Texas during the previous 14 days. The four panels depicts this for four Saturdays in early 2020. Using the notation of Section 3, the three panels depict $LEX_{48201,g',d}$ for d equal to February 29, March 28, April 25, and May 30 of 2020, where 48201 is the FIPS code for Harris County.

Figure B.4: Interquartile Range of DEX over time

Panel A: Raw Device Exposure Index



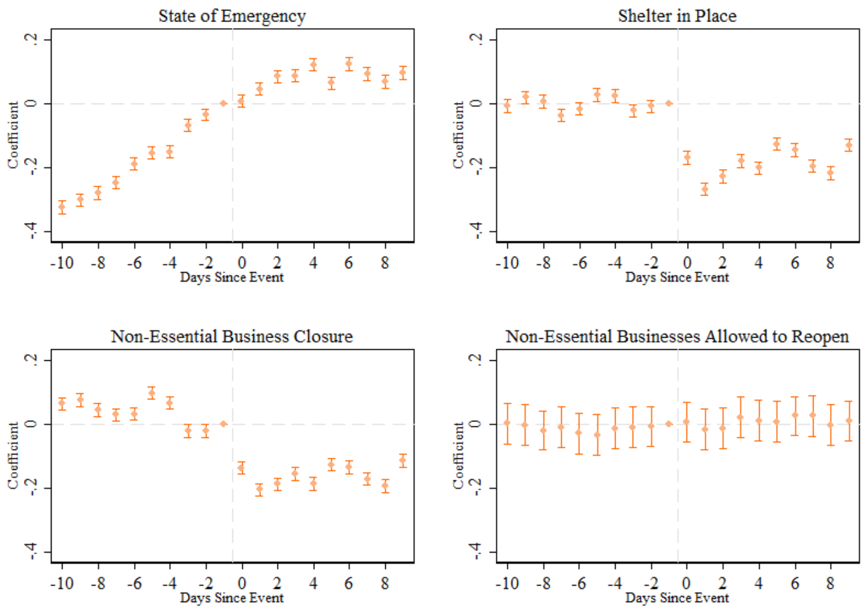
Panel B: Normalized to February 1, 2020



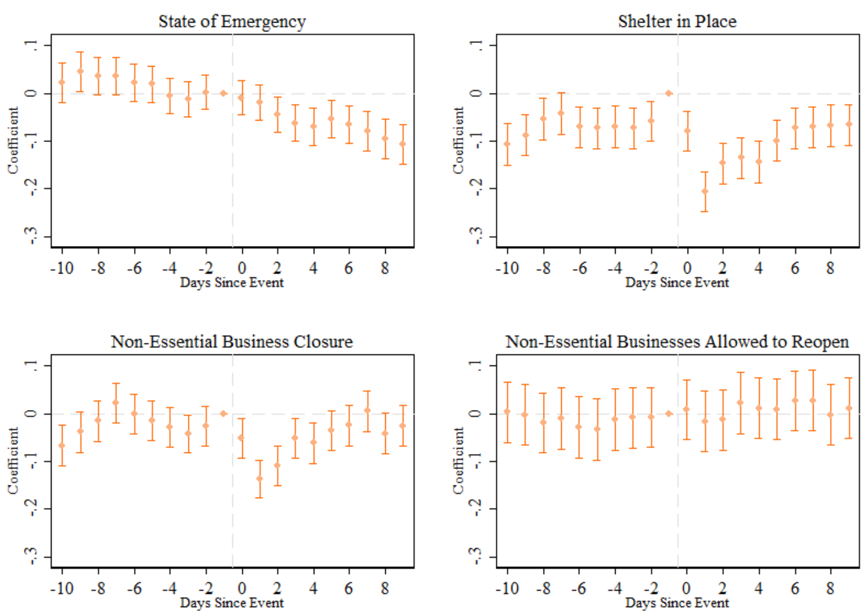
Notes: This figure shows the population-weighted median and interquartile range, of the device exposure index over time.

Figure B.5: Changes in DEX Relative to Lockdown Policies

Panel A: Using All Variation



Panel B: Only Using Cross-State Variation within Commuting Zones



Notes: Each plot in this figure presents the coefficients estimated in a regression of the county-level device exposure index on dummies for the time since a given policy change. In Panel A, these regressions also include county and date fixed effects. In Panel B, the regressions include county and commuting zone-by-date fixed effects. Each plot presents the results for a different state-wide policy, each drawn from Raifman, Nocka, Jones, Bor, Lipson, Jay, and Chan (2020). Each point represents the coefficient on the dummy for a given number of days since the policy was instituted, with the bands reflecting 95% confidence bounds on those estimates.