The Returns to Face-to-Face Interactions: Knowledge Spillovers in Silicon Valley*

David Atkin[†], Keith Chen[‡] and Anton Popov[§]

June 2022

Abstract

The returns to face-to-face interactions are of central importance to understanding the determinants of agglomeration. However, the existing literature studying patterns of geographic proximity in patent citations or industrial co-location has struggled to disentangle the benefits of face-to-face interactions from other spatial spillovers. In this paper, we use highly granular smartphone geolocation data to measure face-to-face interactions (or meetings) between workers at different establishments in Silicon Valley. To study the degree to which knowledge flows result from such interactions, we explore the relationship between these meetings and the citations among the firms these workers belong to. As firms may organize meetings with those they wish to learn from, we isolate causal impacts of face-to-face meetings by instrumenting with the meetings between workers in adjacent firms that belong to unconnected industries. Our IV approach estimates substantial returns to face-to-face meetings with overidentification tests suggesting we are capturing the returns to serendipity that play a central role in the urban theories of Jane Jacobs.

Keywords: Knowledge Spillovers, Face-to-face Interactions, Serendipity, Agglomeration

*We thank Jeff Luan, Ryne Rohla, Yulu Tang, Thyra Tuttle, Vanessa Wong, and Luqiing Zhou for excellent research assistance. Dave Donaldson, Giles Duranton, Ben Faber, James Fenske, John Friedman, Jason Garred, Jessie Handbury, Gabriel Kreindler, Enrico Moretti, Petra Moser, Jesse Shapiro, Joe Shapiro, Meredith Startz and numerous seminar participants provided valuable comments. This research is covered by UCLA IRB number 21-001328 and was determined Exempt Category 4 (Secondary Use Research) by the MIT institutional review boards.

[†]Department of Economics, MIT and NBER. atkin@mit.edu

[‡]Anderson School of Management, UCLA. keith.chen@anderson.ucla.edu

[§]Unaffiliated. popov@mit.edu

1 Introduction

Measuring the returns to face-to-face interactions is central to understanding the determinants of agglomeration, why cities are the driving force of economic growth, and how firms and cities should be structured. Beyond the economic geography and urban economics literatures (e.g. see Moretti, 2012), the importance of face-to-face interactions also potentially sheds light on why despite substantial reductions in trade and communication costs, the world may not be getting flatter (e.g. see Leamer, 2007), as well as many other questions related to organizations, growth, productivity and labor markets. Finally, the value of in-person meetings is also key to understanding the repercussions of the transition to working from home brought about by the Covid-19 pandemic, and how far the pendulum should swing back post pandemic.

In the context of innovation, the classic work of Saxenian (1996) makes the case that frequent face-to-face interactions, and the knowledge flows that resulted, were a large part of what made Silicon Valley the dominant technology hub it is today (rather than the less-interactive Route 128 corridor in Massachusetts). Her argument is perhaps best encapsulated by the memorable Tom Wolfe (1983) quote: *"Every year there was some place, the Wagon Wheel, Chez Yvonne, Rickey's, the Roundhouse, where members of this esoteric fraternity, the young men and women of the semiconductor industry, would head after work to have a drink and gossip and brag and trade war stories about phase jitters, phantom circuits, bubble memories, pulse trains ... "*

Despite their potential importance, measuring the returns to face-to-face interactions, and more generally opening up the black box that is knowledge spillovers, has proved challenging. Both because *"knowledge flows are invisible, they leave no paper trail by which they may be measured and tracked"* (Krugman, 1991), and because the interactions themselves go unrecorded. This paper attempts to make progress by leveraging newly available geolocation data from smartphones to create a digital trail of interactions between workers that can be matched to the citation behavior of the firms they work in.

The importance of knowledge spillovers in driving agglomeration dates back to at least Marshall (1920), who argued that spillovers of ideas were one of three externalities (alongside labor and supplier pooling) that lead firms to co-locate. As knowledge spillovers are central to modern theories of growth, three large empirical literatures try to measure their magnitude or distinguish them from other types of externalities. Jaffe et al. (1993) initiated an enormous body of work that relates patent citations to geographical proximity as a test for the existence of knowledge spillovers. Ellison et al. (2010) and subsequent papers explore associations between industry co-location and R&D or patent matrices to tease out the relative importance of knowledge spillovers vis-à-vis other agglomeration forces. Finally, Glaeser et al. (1992) and others test whether within or across industry concentration matters to distinguish the theories of Marshall, Arrow and Romer (where knowledge spills over to other firms within an industry) from those of

Jacobs (1969) (where spillovers occur across industries).

What these three literatures have in common is that, at best, they only identify broad mechanisms operating through proximity. Our rich data on interactions allow us to open the black box of knowledge spillovers and isolate a particular channel: face-to-face meetings. To do so, we first link worker interactions—measured by the probability that a worker from one establishment "meets" a worker of another establishment by being in the same place at the same time—with patent citations between their employers, an observable proxy for knowledge flows.

To calculate these meeting probabilities, we combine smartphone geolocation data with maps of building rooftops for all patenting firms in Silicon Valley, assigning workers to establishments based on where they spend a large fraction of their waking hours. To assign firm-level citations to establishments, we scrape citation data from recent patent applications and use the inventors' hometowns coupled with the housing locations of workers to probabilistically assign citations across multi-establishment firms. The resulting dataset of establishment-to-establishment worker meetings and citations reveals a strong positive relationship between face-to-face interactions and knowledge flows, even after conditioning on rich controls for the physical distance between establishments.

Of course, interpreting such an association is difficult since people organize meetings with others they wish to learn from (i.e. worker meetings are endogenous). Thus, we instrument our measure of face-to-face meetings between workers at establishments i and j with meetings between workers at adjacent establishments i' and j' that share a similar "meetings geography"—the relevant amenities, housing and transportation that separate workers at different firms. To address remaining confounders related to endogenous establishment location decisions, we control for firms choosing locations based on proximity to certain firms or types of worker by conditioning on the physical distance that separates i and j, and differences in the attributes of workers at i and j. Furthermore, we restrict our instrument to only $i'j'$ pairs whose industries neither cite nor supply each other, with the remaining meetings being primarily serendipitous meetings driven by the meetings geography of the city.[1] Finally, to deal with firms choosing their locations to maximize chance meetings with firms they want to learn from—or with amenities springing up to serve ij meeting demand—we exclude meetings occurring less than 5km from either establishment that may be salient in these location decisions.

Implementing this approach, we find that face-to-face meetings significantly increase citations between establishments, with the strength of the effect twice the impact of physical distance on citations. Eliminating a quarter of face-to-face meetings in Silicon Valley would reduce the number of citations by approximately 8 percent, similar in magnitude to the 3 to 8

[1] Specifically, the worry is that the similarity of establishments i and j, or their workers, is correlated with the similarity between i' and j', and that workers at more similar establishments both cite each other more and meet each other more. By excluding $i'j'$ pairs whose industries ever cite each other, this ij similarity is either along dimensions unrelated to innovation or is not spatially correlated.

percent reduction in citations Jaffe et al. (1993) find from being in a different city. These estimates are the first contribution of the paper: providing evidence for the impact of face-to-face interactions on knowledge flows.

Our estimates also shed light on the potential impact on knowledge flows if Silicon Valley firms allowed remote work on a permanent basis, a trend that has been greatly accelerated by the pandemic. Allowing the frequency of meetings to depend on others' work-from-home decisions, a back-of-the-envelope calculation finds that if one quarter of office workers worked from home instead, face-to-face meetings would fall by 17 percent and citations by 5.2 percent.

Given the nature of our instrument—specifically that it exploits plausibly exogenous differences in the meetings geography between establishment locations—our estimated returns should be interpreted as local average treatment effects (LATEs) of the types of face-to-face meetings induced by variation in the meetings geography of a city. Thus, our estimates combine two mechanisms with potentially very different returns. The first are knowledge flows that directly result from the chance meetings between workers that our instrument—constructed as it is from meetings between workers in unrelated industries—is primarily picking up. We think of these returns to serendipity as "Jacobs spillovers" despite them being different in nature to the inter-industry spillovers Glaeser et al. (1992) associate with Jane Jacobs. We do so because we believe that they capture an idea that Jacobs is perhaps best known for (e.g. Jacobs, 1961, Jacobs, 1969 and Jacobs, 1984) as is made clear by the following quote by Glaeser (2009) himself:

> *"Jacobs' greatest insight was that cities succeed by enabling people to connect with one another. ... Many of the finest achievements of human civilization occurred because smart people learned from one another in cities. As Jacobs understood better than anyone else, the chance encounters facilitated by cities are the stuff of human progress."*

The second mechanism contributing to our LATE estimate is the knowledge flows resulting from planned face-to-face meetings, with our instrument potentially serving as a cost shifter for organized meetings. For example, a worker from i may decide to arrange an in-person meeting with a worker from j, perhaps instead of a phone call, because there is a popular coffee shop located halfway between their offices.

Our second contribution is to propose and implement a test of whether our IV estimates measure these "Jacobs spillovers"—those coming through serendipitous meetings—rather than a LATE combining the returns to both these and planned meetings. If we have two valid IVs for face-to-face meetings that load differentially on chance and planned meetings, and the IV regressions of citations on ij meetings estimate the same coefficient with either IV, either the returns to chance and planned are equal or there is no first stage for one meeting type. An over-identification test does not reject that the returns are the same whether we use just the $i'j'$ meetings occurring during the workday or those occurring at all other times. Since both IVs rely

on the meetings geography captured by $i'j'$ meetings they are certainly correlated with chance meetings. And if there is a first stage for planned meetings, we would expect the loadings of the two IVs to differ from the loadings for the chance meetings first stage (since, if any meetings geography affects the cost of planned meetings it is the workday geography). Therefore, we conclude that either the returns to chance and planned meetings are the same or our instruments do not affect planned meetings. In either case our LATE estimates capture the returns to chance interactions and we interpret this result as evidence in support of Jane Jacobs' insight that serendipity plays a vital role in generating agglomeration externalities in urban areas.[2]

Our paper relates to multiple literatures in urban economics, economic geography, and innovation. As discussed above, we connect closely to three literatures that: provide evidence for knowledge spillovers from the geographic localization of patent citations (e.g. Jaffe et al., 1993; Thompson and Fox-Kean, 2005, or Moretti, 2021 for a related approach); explore the relative importance of different agglomeration forces by studying industry co-location (e.g. Rosenthal and Strange, 2001; Arzaghi and Henderson, 2008; Ellison et al., 2010); and distinguish between inter- and intra-industry spillovers (e.g. Glaeser et al., 1992; Duranton and Puga, 2001).

There is also a literature emphasizing the importance of face-to-face interactions in various contexts (e.g. Storper and Venables, 2004, Charlot and Duranton, 2006, Leamer, 2007 and Startz, 2021), although these papers do not seek to make causal claims about the knowledge flows resulting from such interactions.[3] More similar to our paper in this regard, Catalini et al. (2020) and Pauly and Stipanicic (2022) use new airline routes and the introduction of the jet engine to study the effect of reduced travel costs on academic coauthorship and patenting, respectively. Catalini (2018) explores the former outcome using random office reallocations at a French university. Perhaps most closely related, Andrews (2020) exploits the differential timing of prohibition across US counties to argue that bar talk drives innovation, showing prohibition led to both a decline and change in the direction of innovation. Finally, a large body of theoretical work in macroeconomics studies growth and the diffusion of knowledge by assuming knowledge flows through chance meetings (e.g. Jovanovic and Rob, 1989 or Alvarez et al., 2013 for a model of trade and idea diffusion). Our results provide support for such a modeling assumption.

The paper proceeds as follows. Section 2 describes the firm, patent and geolocation databases we utilize in our analysis, as well as defining our meetings measure. Section 3 outlines our empirical strategy and introduces our instrument. Section 4 reports the results of our regressions, while Section 5 disentangles the role of serendipity. Finally, Section 6 concludes.

[2]Although we use the term spillovers to describe the knowledge flows induced by serendipity, there would be no externality if, upon meeting by chance, a worker demands payment for passing on knowledge (of course, there may still be inefficiencies arising from search externalities). While we are not aware of any evidence, anecdotal or otherwise, that such compensation is commonplace, our data do not allow us to rule out that these knowledge flows are fully compensated (an issue common to essentially all the empirical literature exploring knowledge spillovers).

[3]Two recent papers use cellphone data like we do but to study the relationship between call patterns and either job referrals (Barwick et al., 2019) or physical distance (Büchel and Ehrlich, 2020).

2 Firms, Patents and Meetings Data

We focus our analysis on workers and firms in Silicon Valley. This context is an important one in which to measure the returns to face-to-face interactions. First and foremost, it is a major engine of US growth and *the* major engine of US (and global) innovation. This leadership is most apparent from the fact that 20 percent of US patent applications with US-based inventors filed during the period covered by our patent sample (January 2007 to March 2019) had at least one Silicon Valley-based inventor—despite Silicon Valley containing less than 3 percent of the US population. Recent commentary has speculated that this dominance may be under threat from a movement to working from home and resulting reductions in serendipitous meetings, a question we return to later in this paper.[4]

In terms of external validity, without repeating the analysis elsewhere we can only speculate about how our findings extend to other technology hubs. The culture of Silicon Valley is particularly known for the importance and frequency of face-to-face interactions (e.g. see Saxenian, 1996). In that sense, our results may overestimate the returns to face-to-face interactions in other technology clusters. That said—as evidenced by a recent article attributing the buoyancy of the biotech hub in Cambridge, MA to the serendipitous interactions the cluster facilitates[5]—these forces are likely common to many locations in the US and abroad.

We define our "Silicon Valley" sample by restricting attention to establishments located within 50 miles of Stanford University (which includes the city of San Jose to the south-east, and San Francisco and Oakland to the north-west). Ultimately, we will end up with a sample of 18,360 patenting establishments and 51,580 workers at these establishments (with 4,137 of these establishments recently citing another Silicon Valley establishment and 6,127 recently cited). The next three subsections describe how we come to these numbers by locating firms, inferring workers, counting meetings, and tracking citations.

2.1 Firms and Establishments

Throughout the paper we will refer to three firm-related entities. The smallest is an *establishment*: a single office location of a given firm. A *firm* is a collection of one or more establishments with the same owner, and is the entity which patents (i.e. firms apply for patents not establishments). A *building* is a unique and contiguous physical entity (for our purposes, a rooftop identified from satellite imagery), that may contain more than one establishment if multiple firms operate in the same building.

Our establishment lists were compiled by Orbis, whose database contains details on more

[4] E.g. see LeVine, S., "How Remote Work Could Destroy Silicon Valley," *Marker by Medium*, 2020, July 13 and Ray, T., "Steve Jobs said Silicon Valley needs serendipity, but is it even possible in a Zoom world?", ZDNet, 2020, June 24.

[5] See Kirsner, S., "The suburbs are cheaper, but they don't have what Kendall Square has for biotechs: serendipity," *Boston Globe*, 2020, January 27.

than 60 million US establishments. In order to focus our attention on firms who may potentially apply for a patent, we take the union of two sets; Silicon Valley establishments that belong to a firm that has ever patented according to Orbis (13,273 establishment matches), and every firm with at least one inventor who lives in Silicon Valley according to the patent application data described below (11,814 matches).[6] The locations of our full sample of 18,360 establishments belonging to 13,054 firms—the union of the two sets above—are shown in Figure 1a.

We assign the 18,360 establishments to 9,049 buildings by matching addresses to Microsoft's rooftop shapefile created from satelite images.[7] These shapefiles allow us to assign smartphone pings to the buildings in which they fall, a crucial step in our identification of workers below. To provide a sense of the data, Figure 1b shows buildings containing patenting firms for Palo Alto.

2.2 Meetings Data

We construct our measure of face-to-face interactions between workers of different establishments using smartphone geolocation data. The basic building blocks are smartphone location pings from approximately 50 million handsets (about one fifth of US smartphone users) collected by the firm Safegraph between September 2016 and November 2017. Smartphone operating systems (e.g. Android and iOS) report the estimated physical location of a phone every 5–10 minutes and more frequently if driving. This location estimate is typically accurate to within 20m and, subject to user permissions, is shared with open or backgrounded apps.[8] Safegraph purchases and collates these location data from popular apps. Each ping reports a unique device identifier, a timestamp, a latitude and longitude, and an accuracy estimate.

As smartphones may be turned off, have no reception, or have the relevant apps neither open nor backgrounded, we rarely capture a continuous set of pings throughout each day. For the subsample of phones we identify as belonging to workers below, we observe pings for a mean of 14.3 hours (median 15.6) conditional on observing the phone that day.

Several papers have found Safegraph data to be demographically representative at the national level (see Athey et al., 2020; Chen and Pope, 2020). Here, we examine whether they appear demographically representative within our Silicon-Valley subsample by testing whether the imputed demographics of our full smartphone sample match census-reported demographics at higher levels of spatial aggregation. Table 1 compares county-level (or Silicon Valley-level) demographics to aggregates of census block-group-level demographics, using the distribution of our smartphones across block groups to aggregate demographics up to counties (or to Silicon

[6] If there is no match between a patenting firm and a Silicon Valley firm listed in Orbis, we match to the name of the Silicon Valley firm's global ultimate owner in Orbis as many subsidiaries patent under their owner's name.

[7] A shapefile is a set of vertices of buildings' rooftops, coded as latitudes and longitudes of each vertex. We match 97 percent of establishment addresses to buildings. Data downloaded from https://github.com/microsoft/USBuildingFootprints.

[8] The median accuracy of 95 percent confidence intervals across pings is 19.6m. Unlike today, most users granted location privileges at the time of our sample both because of lower salience regarding privacy and because the Android operating system first allowed users to grant apps 'only while using' location privileges starting in 2019.

Valley). The close fit between the actual demographics and the aggregates provides some reassurance that our smartphone sample is representative of Silicon Valley residents.[9]

2.2.1 Identifying Workers

Before using the geolocation data to measure meetings between workers, we must link smartphones to establishments and thus firms. We do so by assigning smartphones to buildings containing our patenting establishments. We define a *worker* at establishment i as a smartphone device which leaves pings in at least 20 different hours in establishment i's building in a particular month.[10] If there are multiple establishments in the same building, we use all smartphones assigned to that building to generate the various establishment-level meetings measures.[11]

Applying this definition, we locate 51,520 workers in patenting establishments in Silicon Valley in the thickest month of our sample, September 2017. Figure 2 shows the number of workers we identify by sample month, as well as the number of pings for these workers.

To asses whether our methodology for identifying workers is effective, Appendix B correlates our worker counts with the LEHD Origin-Destination Employment Statistics (LODES). These data provide employment counts for 20 industries across 1,942 Silicon Valley census block groups. Conditioning on industry and block group, LODES counts explain 23 percent of the spatial variation in our worker counts despite only capturing a subset of smartphones (one fifth) and pings, being unable to separate workers in the same building, and ignoring workers at non-patenting firms. Focusing just on industries where a large share of firms patent (manufacturing, and professional, scientific, and technical services) we can explain 50 and 72 percent respectively. Taken together, we believe that our smartphone data matched to establishment rooftops do a good job identifying workers at patenting firms in Silicon Valley.

2.2.2 Constructing Meetings Measures

Having identified workers by establishment, we use the ping data to construct measures of "meetings" between workers of different establishments. As will become clear when we introduce our regression specification in Section 3.1, we do not claim to perfectly capture all the actual meetings that occur between workers. Rather, our data record whether two smartphones are located close to each other at the same time, which we use to measure the likelihood that workers of establishment i meet workers of establishment j. We note that this likelihood includes both meetings that arise through serendipitity and those that are planned, with Section 5 devoted to distinguishing the two.

[9] We might expect Silicon Valley workers to be heavier users of mobile apps and so more likely to be sampled. At the same time, they may be more savvy and deny location privileges to apps.

[10] We assign workers to the building in which they spent the most time that month if they are in multiple buildings for more than 20 hours. Workers may be at different firms in different months. Appendix C explores these transitions.

[11] This misattribution will generate measurement error in the (scale-independent) meetings measures we describe below but should not attenuate the IV estimates we describe in Section 3. When worker counts are needed we split workers evenly across establishments within a building.

We start by dividing Silicon Valley into a 7-digit *geohash* grid (henceforth, geo7), where each box is approximately 152-by-152 meters, about the size of a city block. We define each worker as visiting a geo7 g in half-hour-day-month-year h if they spend at least 10 minutes in geo7-box g during time-period h.[12] A "meeting" is defined as two workers who work in different buildings both being in the same g at the same time h, with all establishments in multi-establishment buildings assigned the same value as we discuss below. To calculate the total number of these meetings, we simply multiply e_{igh}, the number of workers of building i who visit geo7 g at time h, by the equivalent number for building j, e_{jgh}, and sum over all locations and time periods:

$$TotalMeetings_{ij} = \sum_{hg} e_{igh} e_{jgh}.$$

Thus, if three workers from building i are at a coffee shop at the same time as two workers from j, that would contribute six meetings to that firm-pair's total.

While both simple and appealing, this meetings measure is subject to the concern that we will record many more meetings for firms where: a) a larger share of workers use the apps our data come from; and b) smartphones ping more frequently (which is a function of the reception in the office, app permissions, whether apps are backgrounded, etc.). To create a measure immune to missing pings, we calculate $PotentialMeetings_{ij}$, the number of times workers from i and j were both present in our database at the same time and so could have potentially met:

$$PotentialMeetings_{ij} = \sum_{h} \left((\sum_{g} e_{igh})(\sum_{g} e_{jgh}) \right).$$

Our main meetings measure, $TotalMeetingProbability_{ij}$ (abbreviated to $TotalMP_{ij}$), is the ratio of observed to potential meetings:[13]

$$TotalMP_{ij} = \frac{TotalMeetings_{ij}}{PotentialMeetings_{ij}}. \tag{1}$$

This ratio has a clear probabilistic interpretation. For a time period h where we observe workers from both i and j, $TotalMP_{ij}$ is the observed likelihood that a randomly chosen worker from i is in the same place as a randomly chosen worker from j. Our goal is to test whether citation behavior is related to variation in this probability. In other words, do firms that meet more also cite each other more?

A couple of comments are in order. First, our meeting probability measure corresponds to a particular mechanism through which knowledge flows occur. $TotalMP_{ij}$ measures the probability that a particular inventor on a patent originating from establishment i meets an inventor on a patent originating from establishment j. Therefore, our meetings measure is appropriate

[12] If two consecutive pings indicate that a worker traveled from geo7 g to g', we split the time between g and g'.

[13] By separately summing total and potential meetings over all h before taking the ratio, we are calculating a weighted average of half-hour probabilities with higher weights given to hs where we observe more smartphones.

if citable knowledge primarily flows between firms when inventors meet other inventors.

Second, the fact that our measure is dimensionless helps deal with the issue that a building may contain multiple establishments and we have no way to distinguish which establishment within a building a smartphone user works at. The probability measure above does not require us to take a stand on how workers are split across establishments within a building, we merely need to assume that the probabilities are the same for the different establishments in the same building. Violations of this assumption will of course generate measurement error but the IV strategy we introduce in Section 3 will be robust to such measurement error as long as it is uncorrelated with errors at establishments in nearby buildings.

Finally, note that our measure does not scale with the number of workers at an establishment. One implication of this property (and of the mechanism it captures), is that we are restricting the effects of meetings on citations to be independent of establishment size. If knowledge flows occur not only when inventors meet each other but also when any of their colleagues meet, the same $TotalMP_{ij}$ will result in larger flows for a pair of large establishments compared to a pair of small ones. Whether there are size effects of this type is ultimately an empirical question and one that we explore in Section 4.2.

To illustrate the construction of our meetings measures, Figure 3 plots the pings of workers at the headquarters of two sizable firms in Silicon Valley, Apple and Google.[14] Green tiles indicate locations frequented by Apple workers, orange tiles by Google workers, and brown tiles indicate overlap (and the darkness of shading indicates intensity). Although informative, $TotalMP_{ij}$ relies only on coincidences of workers—workers in the same place at the same time. Figure 4a shows these coincidences, with reds denoting more meetings in that location and blues fewer. These meeting locations include the establishments themselves (e.g. Apple workers visiting Google headquarters) but also shopping malls, parks, schools, restaurants, golf clubs, doctors' offices, airports, and apartment complexes, as we mark on the figure.

The resulting distribution of $TotalMP_{ij}$ is highly skewed. For our regression analysis, we take the logarithm of $TotalMP_{ij}$ which has a bell-shaped distribution (Appendix Figure A.2, left panel). However, 64 percent of our ij establishment pairs have a zero probability because the numerator of equation (1), $TotalMeetings_{ij}$, equals zero although the denominator, $PotentialMeetings_{ij}$, is positive. Not only is the zero problematic for our log transformation, but it also obscures potentially informative variation in the denominator.[15] We address both issues by adding a small positive number to all numerators.[16] The right panel of Appendix Figure

[14]We use the now-old Apple headquarters at 1 Infinite Loop, Cupertino. Their new headquarters opened in September 2017 but the vast majority of employees only moved after the end of our sample period (November 2017).

[15]Suppose two ij pairs both have no total meetings but in the first case there are many more potential meetings. If we had perfect coverage of workers and their pings, the first case would likely reveal a small meeting probability while the second could have a large probability.

[16]We choose this number using an iterative procedure that ensures that the mean of the previously zero-numerator observations lies at the 10th percentile of the full post-addition distribution. This procedure results in

A.2 shows that the resulting distribution has a broadly similar shape but is more compressed compared to the distribution without the zero-numerator values. Section 4.3 explores robustness to alternative transformations.

We are now in a position to illustrate the variation in $\ln TotalMP_{ij}$, the meetings measure at the center of our analysis. Figure 5a plots the variation for a single establishment i, Apple's headquarters, with all other establishments j. For every j there is a underlying map that contains potentially many coincidences with Apple workers, like the one for Apple and Google above, and we simply mark on j's establishment location the single value of $\ln TotalMP_{ij}$ that summarizes these meetings. Reds represent a higher meeting probability and blues a lower one. Intuitively, workers have a higher probability of meeting workers at nearby establishments. However, much of the variation is not explained by distance. For example, Apple workers are much more likely to meet with workers at firms in the central business district of San Francisco than firms in South San Francisco, and more likely to meet firms in some towns in the South Bay than others. These differences are driven both by planned meetings between workers and by the geography of amenities, infrastructure, and housing in the region—with the latter variation behind the instrument we introduce in Section 3.3.

2.3 Patent Citation Data

The final ingredient in our analysis is information on which firms cite which other firms. Following a long literature in economics, these citations will be our observable proxy for knowledge flows (e.g. see Jaffe et al., 1993 and Thompson and Fox-Kean, 2005). While imperfect, Jaffe et al. (2000) and Roach and Cohen (2013) show that these citations are related to inventors' perceptions of knowledge flows from surveys.

We build our patent citation dataset from the US Patent and Trademark Office (USPTO) databases of patent applications and granted patents. While firm addresses may be misleading (e.g. almost all patents by Palo Alto-based Hewlett Packard were filed by their Texas subsidiary), both patent and patent application files include the home town of each inventor. We therefore restrict attention to files where at least one inventor lives within 50 miles of Stanford.

We draw on the less-analyzed patent application database (rather than the granted patent database) for three reasons. First, additional citations are often added later in the patenting process. This can occur as part of a back and forth with the examiner (and be added by either the applicant or examiner), because the scope of the patent is narrowed, or if new knowledge becomes available. By limiting attention to applications, our citation data provides a more accurate measure of the knowledge that was known by the inventor when the innovation was made. Second, since we are not focused on the novelty of the innovation but instead on whether knowledge flowed, applications provide a more complete picture of these flows than the sub-

adding 0.0561 meetings to the $TotalMeetings_{ij}$ numerator in equation (1).

set of successful applications (i.e. granted patents). Finally, applications are published more quickly than patents, providing us complete visibility on applications made between March 2017 and May 2018 (the time period of our meetings data lagged six months to allow a lag between a knowledge flow and filing a patent).[17] That said, the nature of the cross-sectional regression specification we introduce in Section 3.1 means that this timing assumption will not be particularly consequential for our results.

Extracting citations from patent applications poses a number of issues. Unlike for granted patents—where there is a references cited field in the USPTO full text files—the equivalent files for applications do not include such a field.[18] However, most of these citations are found in the European Patent Office (EPO) database, which receives data from national patent offices including the USPTO. If a patent is filed internationally, as most Silicon Valley patents are, other countries' patent offices record citations in a machine readable form. Thus, we draw on the EPO citation data attached to the foreign counterpart applications to US applications.[19] We supplement these data by scraping 'in-text' citations contained in the patent text itself but which may not appear in the citation lists (comprising 17.2 percent of citations in our sample).

Our final sample contains 16,223 applications from Silicon Valley inventors, citing 41,199 Silicon Valley patents or applications, with 120,530 total citations. We match 104,102 of these to our firm sample to obtain 1,679 citing firms (with 4,126 establishments) and 2,808 cited firms (with 6,562 establishments). Appendix Figure A.3 displays a random sample of citation links between firms, with thicker arrows representing more citations and the arrowhead indicating the direction of flow. Firms throughout Silicon Valley cite each other, with particularly sizable flows for a few key firms and a particularly dense citation web among firms in the South Bay.

These citation measures are at the firm-to-firm level while our meetings are at the establishment-to-establishment level (recall that multi-establishment firms typically use one address to file all their patents). We rely on the inventor's hometown to inform us as to where the invention was developed. Specifically, using our geolocation data we calculate the probability that an inventor worked at a specific establishment from the empirical likelihood that a worker at the same firm living in the same town works at each of the firm's Silicon Valley establishments.[20] To generate a citation measure consistent with our meetings measure—which

[17] The median patent is granted three years after submission while applications are made public in 6–18 months. It is reasonable to think that an inventor working on a new patent application may learn something useful related to a patent filed several years ago. Thus, on the cited side we include both granted patents and patent applications made from January 2007 to May 2018.

[18] Instead the USPTO provides viewable images of the information disclosure statement (IDS) that contains citations to all prior knowledge but purposefully scrambles this information to ensure it is not machine readable.

[19] We use "simple (DOCDB) family" citations that include citations in other documents covering the same invention (e.g. patent continuation applications or divisional patent applications). Of the universe of Silicon-Valley UPTO applications, 80 percent have a non-zero number of citations in the EPO database. The EPO distinguishes examiner-added citations and applicant-added citations, a distinction we explore in Section 4.3.

[20] In cases where we do not observe any of the firm's workers living in an inventor's town, we calculate probabilities by forming weights that sum to 1 from the inverse of the distance between each establishment and the center of an

estimates the probability that a particular inventor at one establishment meets a particular inventor at another one—we treat each inventor-to-inventor pair as its own citation and allocate inventors to establishments using the empirical probabilities described above.[21]

As with the meetings measures, the resulting $PatentCitations_{ij}$ data are sparse.[22] Given this skewness, and our desire to identify knowledge flows from both the extensive and intensive margins, we transform citations using an inverse hyperbolic sine (IHS) transformation that is a logarithmic function for large numbers but takes the value of zero when citations are zero. The histograms of the raw and transformed number of citations are shown in Appendix Figure A.4. For robustness, Section 4.3 considers alternative citation allocation rules and transformations.

3 Empirical Strategy

The cornerstone of the paper is a regression of knowledge flows, as captured by patent citations, on the probability of face-to-face meetings, as captured by geolocation data. In the next four subsections, we describe our specification, discuss identification concerns, construct an instrument to address these concerns, and discuss the interpretation of the resulting estimates.

3.1 The Relationship Between Patent Citations and Face-to-Face Meetings

To understand the link between our measures of knowledge flows and face-to-face meetings, we start with the following ordinary least squares specification:

$$\text{arcsinh}PatentCitations_{ijt} = \beta \ln TotalMP_{ij,t-1} + \Gamma X_{ij} + \delta_i + \delta_j + \varepsilon_{ijt} \tag{2}$$

where $PatentCitations_{ijt}$ are total citations between inventors at establishments i and j (described in Section 2.3) and $TotalMP_{ij,t-1}$ is the probability that a worker at i meets a worker at j (described in Section 2.2). As discussed above, we take log-like transformations of these two variables, both to normalize the skewed distributions and to ease interpretation.[23]

As any knowledge gleaned from a face-to-face meeting does not instantly become a patent application, our baseline specification regresses citations on the meeting probability lagged six months, $\ln TotalMP_{ij,t-1}$ (where the $t-1$ subscript denotes a 6 month lag). Thus, we run a cross-sectional regression with patent citations appearing between March 2017 and May 2018 as the dependent variable, and meetings occurring between September 2016 and November 2017 as

inventor's home town. We do not attempt to identify the homes of named innovators, and hence their meetings, both for privacy reasons and because we only have smartphone data on a fraction of workers.

[21] To illustrate our approach, imagine two inventors at firm A cite one inventor at firm B, generating two inventor-to-inventor citations. Firm A has establishments A1 and A2, and firm B has only B1. If firm A's two inventors have a probability of 0.8 and 0.4 of working at A1, respectively, we will assign $(0.8+0.4)=1.2$ citations to the establishment pair (A1, B1), and $(0.2+0.6)=0.8$ citations to the pair (A2, B1).

[22] For the vast majority of ij pairs, i does not cite j during our sample period. But this leaves non-zero citations for 378,689 pairs of establishments out of 218m (with a mean of 0.22 and a max of 804 citations for these pairs).

[23] Recall we use the IHS of citation counts which are whole numbers (or fractions for multi-establishment firms). For meeting probabilities, which are small ratios, we add 0.056 to all $TotalMP_{ij,t-1}$ numerators before logging.

the independent variable. However, given the serial correlation of both citations and meetings—and the heterogeneity in the time taken for a knowledge flow to become a citation—we are not able to link a specific set of meetings to a specific knowledge flow. Instead, we can only ask whether establishments whose workers meet each other more often cite each other more.[24]

Our baseline specification additionally includes establishment i and j fixed effects to control for the fact that some establishments cite or are cited more, as well as controls at the ij level, for example whether i and j are in the same industry. Given that our ij citation flows derive from firm–level not establishment-level data, standard errors are clustered at the IJ pair level where I is the firm establishment i belongs to and J is the firm establishment j belongs to.

This regression reveals whether our face-to-face meeting measures are associated with citation behavior. What the OLS regression cannot provide is the causal relationship between the two. Most obviously, if R&D workers in establishment i are working on an idea and come across a relevant patent from workers at j, they may contact them and arrange a meeting to learn more. In this case the citation generates the meeting and we have reverse causality. The endogenous location decisions of firms and workers add further challenges. We now describe these identification concerns and how we address them.

3.2 Identifying the Impact of Face-to-Face Meetings on Patent Citations

For ease of exposition, we first discuss identification concerns that arise even if establishments and workers were assigned randomly across locations before turning to the more general case.

Case 1: Patenting establishments and workers randomly assigned

In this scenario, the main threat is the reverse causality outlined above. Firms working on similar things may both cite each other (a positive error ε_{ij} in specification 2) and organize meetings to learn from each other or attend the same events (generating higher $TotalMP_{ij}$).[25]

We address this issue through an IV strategy. We instrument $TotalMP_{ij}$ with $TotalMP_{i'j'}$, the meeting probability between workers at establishments neighboring i, labeled i', and establishments neighboring j, labeled j'. The instrument is highly relevant since most meetings picked up by our geolocation data are chance meetings driven by the meetings geography that separates i from j—i.e. the layout of amenities, transportation and housing—and this geography is very similar to that which separates i' and j'. The instrument is also likely exogenous, at least when considering this endogeneity concern under the assumption that establishments and workers locate randomly. $TotalMP_{i'j'}$ derives from behaviors of workers at i' and j'. Thus, i and j working on similar technologies will not induce workers at i' and j' to meet more frequently. For the same reason, this IV strategy addresses concerns that similar types of workers

[24] For completeness, we also report specifications using a pure cross-section as well as alternative lag structures. Testing the validity of the timing assumptions is not feasible given the short 15 month panel.

[25] Given we are de facto running cross-sectional regressions, we drop the t subscripts for clarity.

at both i and j, for example those with PhDs, may both cite each other more and hang out in the same places. Note that these arguments do not rest on us assuming all $i'j'$ meetings occur by chance. If i' and j' do meet for work-related reasons, these meetings would have to be correlated with ij citation behavior to bias our estimates. Below we refine our instrument to deal with patterns of spatial correlation that may generate such biases.

A second concern that arises even in this simple case is that amenities may be endogenous to ij meeting demand. For example, if many workers from i and j meet regularly to share ideas (a high ε_{ij}), a coffee shop may open between them to profit from these customers. If this coffee shop subsequently attracts workers from i' and j', $TotalMP_{i'j'}$ will rise. While this concern may be serious in scenarios where i and j are located close to each other, it is unlikely to generate endogeneity concerns for amenities located far from both i and j (since ε_{ij} shocks constitute a very minor share of sales for these amenity providers and thus should not drive location decisions). Therefore, to address this concern we restrict our instrument to only include $i'j'$ meetings occurring more than 5km from either establishment i or j—i.e. meetings at amenities whose location decisions were unlikely to be influenced by $TotalMP_{ij}$.

Finally, measurement error is a serious concern given that our meetings measures are noisy estimates of the actual meeting probabilities for workers at i and j—both because of our incomplete coverage of workers and pings and because we only measure smartphone proximity. This measurement error will attenuate our estimates but the noise in our $i'j'$ meetings should be uncorrelated with that in our ij measures and thus serve as a suitable instrument.

Case 2: Patenting establishments and workers locate endogenously

Endogenous location choices generate a number of additional identification concerns. We mitigate these concerns through rich controls and further refinements to the IV strategy.

The most obvious source of bias is that firms working on similar things locate close to each other, anticipating that such proximity will reduce costs related to physical distance. A firm that supplies or collaborates with Intel may both cite them and want to locate close by (i.e a high ε_{ij} leads i and j to locate nearby leading to a high $TotalMP_{i'j'}$). Such endogeneity concerns are explicitly captured by flexible controls for as-the-crow-flies ij distance, ij driving distance and ij travel time.[26] The identifying variation that remains compares pairs of firms that are equally far apart in terms of physical distance but differ in the meetings geography separating them.

A more extreme version of this concern is that firms are savvy enough to chose locations based on the probability of chance meetings with firms they want to learn from. For example, a firm wishing to foster serendipitous interactions with Intel may locate on the amenity-rich side of Intel's offices. In this scenario, variation in $TotalMP_{i'j'}$ conditional on physical distance will still be correlated with ε_{ij}. We conjecture that any such bias is small for three reasons.

[26]We include cubics of; the Euclidean distance between the business addresses of i and j, ij driving distance, and ij driving duration (the latter two using OpenStreetMap's API based on speed limits of each road segment).

First, firms choosing locations need to know how the ij probability of chance meetings varies by location (beyond variation due to physical distance, travel time etc. and conditional on i and j fixed effects), information that is hard to obtain absent geolocation data. Second, in the tight Silicon Valley real estate market there needs to be enough vacant and affordable spaces such that differences in the probability of ij chance meetings affect choices. Third, these location decisions have to have been made recently or the probabilities of chance meetings need to have remained fairly constant over time despite changes in infrastructure and amenities.

Ideally, we would mitigate this concern by exploiting these changes in amenities over time (i.e. we would compare ij citations before and after new amenities opened that altered the ij meetings geography). However, our 15 month panel is too short to implement such a strategy. Instead we rely on the restriction described above that removes $i'j'$ meetings that occur less than 5km from i or j from our instrument. The logic is that while firms may try to chose locations that maximize meeting probabilities with specific firms (even after conditioning on distance and fixed effects), only the meetings geography close to the establishment is likely to be salient and/or pivotal. In other words, a firm will not choose a location based on the probability their workers will bump into Intel workers many miles from the office.

The final source of bias for our IV approach is that workers with similar characteristics or preferences both cite each other more and congregate in the same places. While our IV guards against this bias by not using ij meetings but the meetings of neighbors i' and j', concerns remain if these worker types are spatially correlated. For example, if Apple and Google workers tend to have postgraduate degrees or like hip coffee shops, workers at firms next door may also share these characteristics—either because highly-skilled firms locate close to each other or close to specific types of worker, or because workers choose jobs near amenities they like. Specifically, our IV estimates will be biased if three relationships hold simultaneously: (1) Similar types cite each other more or less ($\varepsilon_{ij} = \gamma similarity_{ij} + u_{ij}$ with $\gamma \neq 0$, where $similarity_{ij}$ is the similarity of worker types at i and j); (2) similar types congregate in the same places conditional on ij distance ($E[similarity_{i'j'} TotalMP_{i'j'} | X_{ij}] \neq 0$); and (3) similarity is spatially correlated ($E[similarity_{ij} similarity_{i'j'} | X_{ij}] \neq 0$). We address this concern both through additional controls and by removing the more problematic variation in our instrument.

Most directly, we control for the similarity of the workers employed by i and j. Specifically, we calculate distances between i and j in demographics space for 11 measures of income, race, educational attainment, and labor force participation. We first construct the demographics of an establishment by averaging values of each measure from the census across the census block groups where the establishment's workers live.[27] We then use cubic polynomials of the differ-

[27] These 11 census-block-group level variables are: median income; shares by race (asian, white, black, and hispanic); shares by education (high school degree, some college, associate degree, bachelor degree, and graduate degree); share unemployed; and share not in labor force. Every group has an excluded category. We take weighted

ence between i and j along each demographic dimension to flexibly control for the similarity of i and j's workers. Continuing the Apple-Google example, we control for the fact that both companies have a high share of workers with graduate degrees who may be more likely to bump into each other and cite each other. If the inclusion of these (observable) controls has only a small impact on our coefficient of interest, β, an argument along the lines of Altonji et al. (2005) suggests that any bias due to similarity in other (unobservable) characteristics is likely small.

Our second and complementary approach to dealing with spatially correlated worker similarity is to restrict the variation in our instrument to meetings between workers at $i'j'$ pairs whose industries have never cited each other nor supplied each other. Put another way, we instrument ij meetings with meetings between workers at neighboring firms who are distant from each other in both citation space and input-output space. These restrictions are severe with 55 percent of potential $i'j'$ pairs eliminated. By removing the $i'j'$ meetings most likely to be work-related and hence planned, these remaining $i'j'$ meetings primarily capture serendipitous meetings driven by the meetings geography of the city—the types of meetings that generate what we term "Jacobs spillovers".

More formally, since i' and j''s industries don't cite each other, one of two things must be true. Either the types of similarity related to innovation are not spatially correlated for these $i'j'$ pairs (i.e. $\varepsilon_{ij} = \gamma similarity_{ij} + u_{ij}$ but $E[similarity_{ij} similarity_{i'j'} | X_{ij}] = 0$); or the dimensions on which ij and $i'j'$ are similar are unrelated to innovation (i.e. $E[similarity_{ij} similarity_{i'j'} | X_{ij}] \neq 0$ but $\gamma = 0$). In both cases, any endogeneity bias is eliminated. Continuing the Apple-Google example, rather than instrumenting with meetings between the many skill-intensive firms close to Apple and close to Google, we only use meetings between two firms in sufficiently different industries such that either their workers are not similar (e.g. high-skilled scientists at a medical equipment designer and low-skilled cooks at a food service provider) or the dimensions of similarity are irrelevant for citation behavior (e.g. a desire to work close to a Caltrain stop).

One remaining concern is that even if the dimension of similarity is irrelevant for citation behavior, it may still affect worker transitions and these transitions may affect citations (e.g. if workers bring with them knowledge of their previous employer's patents). Indeed, Appendix C shows that ij transitions respond to our meetings measures. To tackle this concern directly, we will explicitly control for ij transitions measured through the (inverse hyperbolic sine of the) number of our workers that move between i and j during our sample period.

3.3 Constructing the Instrument

Recall that our instrument for $\ln TotalMP_{ij}$ is constructed from the meetings between the neighbors of i and j, removing both meetings occurring less than 5km from either i or j and

averages $\bar{x}$ with weights equal to the number of workers in each block group and then calculate absolute ij differences, normalized to lie between 0 and 1: $|\overline{x_i} - \overline{x_j}| \,/\, \max_{i,j} |\overline{x_i} - \overline{x_j}|$.

those between establishments in industries that cite or supply each other. To implement, we draw 1km donuts $D(i)$ and $D(j)$ around establishment i and j. We then form the set $D(i,j)$ from all pairs of our sample establishments $i' \in D(i)$ and $j' \in D(j)$ that fulfill three criteria that eliminate the more problematic meetings variation: (1) i' and j' do not have a primary or secondary 4-digit NAICS or NACE industry category in common; (2) no firm in the industry of i' in our patent sample has ever cited or been cited by *any* firm in the industry of j';[28] and (3) the industry of j' does not buy more than 1 percent of the production of the industry of i' or supply more than 1 percent of its inputs.[29]

Figure 4b illustrates this IV strategy by plotting the meetings between the admissible pairs formed by establishments in the donuts around Apple and Google. We also mark with circles around Apple and Google the $i'j'$ meetings that are excluded when we focus only on meetings occurring more than 5km away from either establishment. Compared to the actual Apple-Google meetings in Figure 4a, Figure 4b is much more sparse as Apple and Google have large campuses and so few other establishments in their 1km donuts. Of more importance for our identification strategy, the locations of the meetings also differ between figures. Many Apple-Google meetings occur at work-meeting locations such as each other's offices, another firm's office, a university campus, or the San Jose McEnery Convention Center. In contrast, once we exclude the meetings within the 5km rings (that include Google offices and a potentially endogenously located coffee shop), the $i'j'$ meetings occur entirely at amenity locations (restaurants, shopping malls, fairgrounds, and apartment complexes)—plausibly capturing the serendipitous meetings driven by the meetings geography of the city.

Using the total meetings data from the $i'j'$ establishment pairs in the set $D(i,j)$ defined above, we calculate an analogous measure to $TotalMP_{ij}$ defined in equation (1) above:

$$TotalMP_{i'j'} = \frac{\sum_{hg, i' \in D(i,j), j' \in D(i,j)} w_{i'} w_{j'} e_{i'gh} e_{j'gh}}{\sum_{h, i' \in D(i,j), j' \in D(i,j)} \left(w_{i'} w_{j'} (\sum_g e_{i'gh})(\sum_g e_{j'gh}) \right)}. \tag{3}$$

As before $e_{i'gh}$ is the number of workers of establishment i' in geo7 g during the half-hour period h. The numerator sums total meetings over all donut pairs while the denominator sums potential meetings over all donut pairs. As i's close to i at the center of the donut likely serve as more relevant instruments than i's on the edge of the donut, we weight $i'j'$ pairs using the product of $w_{i'} = w(distance_{ii'})$ and $w_{j'} = w(distance_{jj'})$ with $w(\cdot)$ being the Epanechnikov kernel that places weight zero at the edge of the donut and the highest weight in the middle. To eliminate any meetings occurring near i or j, we restrict the summation over g in the numera-

[28] Since we cannot distinguish workers from different establishments in the same building, we apply this restriction at the building level. Omitting industry pairs that ever cite each other also guards against misclassification of establishments to industries. If, for example, workers at i' and j' are working on similar technologies despite being assigned to very different industry codes (e.g. Google and Volvo both working on electric cars), then firms in the automotive and IT industries would likely cite each other and so not be admissible $i'j'$ pairs.

[29] We use the 2012 US 4-digit NAICS use-based input output tables (the most recent 4-digit table).

tor of equation (3) to locations more than 5km from both i and j, and restrict the denominator to individuals observed more than 5km from their workplace in that half-hour h. We label the resulting measure $TotalMP_{i'j'}^{>5km}$.

Although much less severe than for $\ln TotalMP_{ij}$ since we are summing over multiple $i'j'$ pairs, in 23.9 percent of cases the numerator of $TotalMP_{i'j'}$ is zero yet the denominator still provides valuable information. Thus, we carry out the same procedure described in Section 2.2 and add a small number to all numerators.[30] However, $TotalMP_{i'j'}$ is still undefined when one or both of the donuts is empty. Given the identification assumptions implicit in the construction of the $i'j'$ measures above, if only one donut is empty we can still construct valid measures using meetings between i and $D(j)$ pairs, or between $D(i)$ and j pairs, after first removing those pairs that do not fulfill the three criteria above. For the 7.25 percent of ij pairs where $D(i,j)$ is empty but these alternative measures are non missing, we use the average of the (i,$D(j)$) and ($D(i)$,j) measures. For robustness, we report results using only the ($D(i)$,$D(j)$) measures.

Figure 5b plots $\ln TotalMP_{i'j'}^{>5km}$, again for Apple headquarters i with all other establishments j. Compared to $\ln TotalMP_{ij}$ in Figure 5a, the $i'j'$ meetings captured by our instrument more strongly correlate with proximity to Apple (most evident from the small number of $i'j'$ meetings with workers at San Francisco establishments relative to ij meetings). This pattern is consistent with a simple data generating process whereby workers are more likely to visit locations close to their establishment. Thus, the chance meetings that feature more prominently in $TotalMP_{i'j'}^{>5km}$ are more common for closer establishment pairs. In contrast, planned meetings that feature more prominently in $TotalMP_{ij}$ are less determined by proximity. That said, there is still a large amount of differential variation across locations.

3.4 Understanding and Interpreting the IV Specification

Where does our identifying variation come from? We are implicitly comparing the citation behavior of ij pairs that are equally far apart in both physical distance and worker demographics but have different meetings geographies separating them—measured through meetings of workers at neighboring establishments i' and j' belonging to unconnected industries. For example, Apple, Google and Intel's headquarters (located in Cupertino, Mountain View and Santa Clara) form three vertices of an equilateral triangle. If the meetings geography is such that people working in Cupertino and Mountain View typically live and go out around Palo Alto and those working in Santa Clara do so around San Jose, we would record more $i'j'$ meetings between firms close to Apple and Google compared to those close to the two other pairs.

[30] We add 0.013 to the 5km ($D(i)$,$D(j)$) measures (and 0.014 to both (i,$D(j)$) and ($D(i)$,j) described below) chosen so that the mean of initially zero-numerator observations falls at the 10th percentile of the final distribution. Appendix Figure A.5 shows that the $\ln TotalMP_{i'j'}^{>5km}$ distribution is almost identical before and after this procedure.

What are we estimating? As with any IV regression, the IV coefficient only identifies the local average treatment effect (LATE) in the presence of treatment heterogeneity. Specifically, we capture the effect on ij citations of the additional ij face-to-face meetings $TotalMP_{ij}$ induced by variation in the meetings geography underlying our instrument $TotalMP_{i'j'}^{>5km}$.

We expect our instrument to primarily induce variation in chance meetings between i and j given its construction from meetings between unconnected neighbors of i and j. If this is the only route through which our instrument affects meetings, our LATE will capture the effect of serendipitous meetings on knowledge flows (our "Jacobs spillovers"). However, variation in $TotalMP_{i'j'}^{>5km}$ may also affect planned meetings by changing the costs of organizing face-to-face meetings. For example, a popular coffee shop between i and j may make a worker at i more likely to arrange a meeting with one at j, or turn an online meeting into an in-person one. Thus, our LATE estimate is some combination of the returns to serendipity and the returns to the types of planned meetings that are sensitive to this variation in meeting costs. Given their nature, the returns to these two types of meeting may differ greatly—an issue we revisit in Section 5 where we develop a test for whether the our LATE recovers the returns to serendipity.

What does this LATE estimate tell us? First, a positive and significant estimate allows us to reject the null that face-to-face meetings do not affect knowledge flows as measured through citation behavior. Second, the magnitude is directly informative as to the return to the types of face-to-face meetings induced by increasing how often people bump into each other (i.e. the 'meetings geography' of the city). This is an important object for urban and innovation economists, and the reduced form is of particular interest to urban planners as it provides a mapping from the geography of meetings—which policymakers can affect through infrastructure projects, zoning, public events, or pedestrianized shopping districts—to knowledge flows.

A final caveat. Given our fixed effects, our estimates come from comparing firm pairs with more meetings to those with fewer. Thus, we cannot distinguish whether face-to-face meetings create novel citations rather than simply reallocate citations across firms. We are helped by the fact that firms have a legal requirement to report all 'prior art'—i.e. public information relevant to the patent claim—that they are aware of, so meetings are not simply pushing them towards citing an inventor they know better. If firms follow the law and our effects are driven by serendipitous meetings (something we find support for in Section 5), the former explanation—that we are identifying increases in total knowledge—is more plausible since firms may reorient planned meetings but serendipitous meetings are not under their control. That said, in either scenario face-to-face meetings are conduits for knowledge transfers.

Serendipity and knowledge flows. Two clarifications are valuable when thinking about the types of serendipitous meetings that generate knowledge flows. First, we conjecture that the majority of such meetings do not involve striking up a conversation with a stranger but instead

meeting an acquaintance by chance. For example, while waiting in line at a coffee shop a worker from i may bump into a classmate from college or a former colleague who works at j. This chance meeting may spark a conversation that leads to a transfer of knowledge or a collaboration. In effect, the density and frequency of interactions in Silicon Valley lead to serendipitous meetings that activate links on a rich pre-existing social network.

Second, we note that workers may deliberately go to certain locations—such as the bars listed in the Tom Wolfe quote in the introduction—with the hope of meeting someone who benefits their work, career or social life. Certainly in the parlance of Silicon Valley these meetings are thought of as serendipitous since the worker does not know who they will meet there. Implicitly, we will be exploiting the fact that a popular bar is located between the establishment pair $i_1 j_1$ but not $i_2 j_2$ (and so more meetings occur between $i_1' j_1'$ since they often end up in the same bar than between $i_2' j_2'$ who do not).

4 Estimating the Returns to Face-to-Face Meetings

Before turning to our IV regressions of citations on face-to-face meetings, Table 2 presents the first stage regressions of ij meeting probabilities, $\ln TotalMP_{ij}$, on $i'j'$ meeting probabilities $\ln TotalMP_{i'j'}$. As expected, the first stage is very strong as the meetings geography separating ij is highly correlated with that separating $i'j'$. Put another way, most coincidences of workers occur by chance and the probability of a worker at i bumping into a worker at j is highly correlated with the probability of workers at adjacent establishments bumping into each other. While the first stage weakens somewhat when we restrict the instrument to meetings more than 5km from either establishment (column 2), and falls further when we add cubic controls for ij distance and ij demographic distance (column 3), the first-stage F stats remain very high.

We now turn to our main regression. Column (1) of Table 3 presents the (likely biased) OLS specification in equation (2) that regresses ij citations on the total meeting probability between workers at i and j. We find that meetings are positively and significantly associated with citations. In terms of magnitudes, the coefficient on the meeting probability is small, but that is primarily because few firms actually cite each other (magnitudes are ten times larger if we drop firms that neither cite nor are cited). We turn to dealing with the various confounds discussed in Section 3.2 before further interpreting magnitudes.

To better understand the results that combine both a rich set of bilateral controls and our $i'j'$ meetings instrument, we report each modification to the OLS in isolation before presenting the full baseline specification. Column (2) of Table 3 replicates the OLS in column (1) but controls for the possibility that firms working on similar things may choose to cluster together by incorporating flexible controls for the distance between i and j (cubic polynomials of Euclidean distance, driving distance and driving duration). The coefficient on face-to-face meetings falls

by 8 percent, consistent with endogenous location choices exerting an upward bias. Column (3) extends the OLS to include cubic polynomials of absolute differences in the demographics of workers at i and j—income, race, labor force participation, and shares of population across six education bins. These controls address the fact that ij worker similarity may lead workers to both cite each other more and congregate in the same places. The coefficient falls by 10 percent, again consistent with an upward bias from endogenous location choices. Following the logic of Altonji et al. (2005), the fact that controlling for observable differences in worker characteristics has relatively little effect on our estimates provides some reassurance that unobservable differences are not driving our findings.

Columns (4)–(6) of Table 3 introduce our instrument. Column (4) instruments $\ln TotalMP_{ij}$ with $\ln TotalMP_{i'j'}$—the log meeting probability between establishments adjacent to i and j whose industries neither cite nor supply each other. As discussed at length in Section 3.2, this instrument tackles reverse causality coming from workers meeting with firms they wish to learn from and mitigates the concern outlined above regarding ij worker similarity, both of which exert an upward bias on our OLS. The instrument also addresses measurement error in $TotalMP_{ij}$ which exerts a downward bias. Compared to the OLS, the coefficient *rises* by 54 percent rather than falls, suggesting substantial measurement error in our imperfect meetings measures inferred from smartphone proximity. To separate these sources of bias, column (5) presents a split-sample IV that purely addresses measurement error. We lay a checkerboard of geo7 rectangles over our map and calculate $TotalMP_{ij}$ only using meetings occurring in the white squares of the checkerboard and instrument it with the same object calculated from meetings in the black squares. Classical measurement error in each of these meetings measures will be uncorrelated and so will not bias this IV specification. The coefficient rises by 89 percent, an additional 35 percent compared to column (4)—consistent with the reverse causation and worker similarity issues present in column (5) but addressed in column (4) generating an upward bias as conjectured. Finally, column (6) attempts to deal with concerns that firms choose locations based on the potential for serendipitous meetings or that amenities may be endogenous to meeting demand by refining the instrument to only include $i'j'$ meetings that occur more than 5km from i and j ($\ln TotalMP_{i'j'}^{>5km}$). The coefficient rises slightly compared to column (4), the IV that does not exclude nearby meetings, although the sample also shrinks as we lose establishment observations where no workers were observed more than 5km from their workplace.

Our baseline specification presented in column (7) combines this last instrument with the distance and demographic controls. We find a highly significant coefficient on the log total meeting probability equal to 0.0001. The simplest way to gauge magnitudes is to compare the coefficient on log meetings to (minus) the coefficient on log distance. With the caveat that these are not true elasticities due to the IHS transformation, the coefficient on total meetings in col-

umn (7) implies an elasticity twice the size of that on Euclidean distance. Furthermore, the coefficient on log distance shrinks by two thirds after conditioning on $\ln TotalMP_{ij,t-1}$, suggesting that face-to-face meetings are a primary conduit for the distance effects found in the patenting literature.[31] Given the myriad channels through which knowledge flows may dissipate with distance, the impacts of face-to-face meetings appear substantial.

We can assess magnitudes in other ways. Reducing $TotalMP_{ij}$ by 1 percent reduces citations by 0.27 percent, while reducing $TotalMP_{ij}$ by 25 percent reduces citations by 7.9 percent. For comparison, Jaffe et al. (1993) find that firms are between 3.2 and 7.5 percentage points more likely to cite firms in their own metropolitan area (the San Jose–San Francisco–Oakland CSA in our case).[32] Thus our estimates are broadly comparable in magnitude. Below we explore the citation implications if Silicon Valley's largest firms were to move elsewhere.

4.1 Working from Home and Knowledge Flows

These estimates also provide a way to assess the impacts of the shift to working from home. This trend has accelerated substantially due to the Covid-19 pandemic, particularly in Silicon Valley. For example, in May 2020, CEO Mark Zuckerberg suggested that half of Facebook employees could be working remotely within 5 to 10 years, while firms such as Dropbox, Lyft, Slack, Square and Twitter now allow their employees to work from home permanently. A major concern with such developments is the decline in knowledge flows that might result from fewer face-to-face interactions. Combined with a measure of how working from home affects meetings, our estimates allow us to quantify such concerns.

While we do not have a source of exogenous variation in working from home that could be used to estimate declines in meetings, we perform the following back-of-the-envelope calculation.[33] We first isolate workers who spend working hours at their establishment on some work days and at their home (i.e. the location where they are most often observed at night) on other workdays. We then calculate the probability that these workers meet other workers, with the probability depending on whether they are working from home and whether the person they meet is also working from home. We estimate $p^{home}_{home} = 0.032$, $p^{office}_{home} = 0.044$, $p^{home}_{office} = 0.097$, and $p^{office}_{office} = 0.208$ where p^n_m is the probability a worker of type m meets a worker of type n in a given half hour. Intuitively, office workers are most likely to meet other office workers, home workers are the least likely to meet each other, and the two mixed cases lie in between.

These four probabilities are likely to change if more people permanently work from home (for example, many more people will be buying lunch in residential areas and so there will be

[31] Specifically, column (1) of Appendix Table A.2 runs the OLS regression in (2) (without distance or demographic controls) but replaces $\ln TotalMP_{ij,t-1}$ with log Euclidean distance between i and j and obtains a coefficient equal to -5e-05. Column (2) further includes $\ln TotalMP_{ij,t-1}$ and the coefficient on log distance attenuates by 62 percent.

[32] See Jaffe et al. (1993) Table 3, taking the difference between the row "citations excluding self-cites" and "controls".

[33] The increase in working from home during the pandemic perfectly coincided with a massive reduction in people meeting face to face and so does not provide suitable variation.

more chance encounters). We exploit day-to-day variation in the number of people working from home in our Silicon Valley sample to estimate these elastictities.[34]

With these probabilities and elasticities in hand, we predict that if 25 percent of our office workers permanently work from home, the number of meetings with other workers would decrease by 17.3 percent. Feeding this decline through our estimates from column (5) of Table 3 implies a not-insubstantial reduction in citations of 5.2 percent. If half our office workers worked from home, meetings would fall by 35.1 percent and citations by 11.8 percent. Of course, these calculations assume that workers (and firms) remain in Silicon Valley, and that permanently working from home does not lead to behavioral changes that our elasticities miss. If workers take advantage of the greater flexibility to move to further-away locations that are cheaper or more scenic, face-to-face meetings would likely fall further. If permanent work-from-homers make conscious efforts to meet more people, meetings would fall less.

4.2 Interactions with Establishment Size

As discussed in Section 2.2.2, our meeting probability measure $TotalMP_{ij}$ is scale free as we divide actual by potential meetings. However, the same meeting probability may have different impacts on citation behavior when i and j have many workers than when they are small.

Column (8) of Table 3 explores this heterogeneity. We rerun the specification in column (7) but now adding an interaction between $\ln TotalMP_{ij}$ and $\ln\sqrt{Workers_i}\sqrt{Workers_j}$, the product of the number of workers at establishment i and j.[35] The interaction of our earlier IV with $\ln\sqrt{Workers_i}\sqrt{Workers_j}$ serves as a valid instrument for this additional term under our previous arguments. There is clear evidence of heterogeneity, with a strong positive coefficient on the interaction. Compared to when i and j both have the median number of workers per month observed in our data (0.91), moving both to the 95th percentile (4.3) raises the effect size by 37.8 percent while moving to the 5th percentile (0.18) decreases the effect size by 39.4 percent.

This heterogeneity by firm size, while not enormous, warrants discussion. There are two obvious explanations. The first is that larger firms have already paid fixed costs such as in-house patent lawyers or derive more value from patents as they have the resources to defend and enforce them. Thus, knowledge flowing to large firms is more likely to result in patents being filed. The second is that the mechanics of knowledge flows are such that the absolute number of meetings rather than the meeting probability matters—and there are more meetings between establishments with more workers. For example, rather than only meetings between inven-

[34] We regress each of the four probabilities above on the log number workers observed in that half-hour in the group n they are meeting with, $p^n_m = \alpha^n_m + \beta^n_m \ln N_n + \epsilon$ where N_n is the number of type n worker half-hours. We obtain the following coefficient estimates: $\beta^{home}_{home} = 0.016$, $\beta^{office}_{home} = 0.011$, $\beta^{home}_{office} = 0.028$, and $\beta^{office}_{office} = 0.068$. We then calculate the number of meetings from $N_{home}(p^{home}_{home}+p^{office}_{home})+N_{work}(p^{home}_{office}+p^{office}_{office})$ in each period.

[35] $Workers_i$ is the the total number of unique smartphone-month pairs attached to establishment i divided by the total number of months. The main effect of $\ln\sqrt{Workers_i}\sqrt{Workers_j}$ is absorbed by the i and j fixed effects. We find similar results interacting $\sqrt{Workers_i}\sqrt{Workers_j}$ and including the (no longer absorbed) main effect.

tors at i and j generating knowledge flows (as is implicit in our meetings probability measure), meetings between their colleagues may also matter. We leave discriminating between these two explanations, and further exploration of the mechanics behind knowledge flows, to future work.

The impacts of departing superstar firms. We can use these estimates to predict the reduction in citations if the 20 most-cited firms in Silicon Valley were to move elsewhere, an exercise similar to that for generic spillovers in Moretti (2021). These 20 firms alone account for 33 percent of Silicon Valley citations. To mimic these firms leaving, for every establishment i we reduce their meetings with the establishments $\tilde{j}$ of these 20 most-cited firms to the 5th percentile of all their other $TotalMP_{ij}$. Appendix Table A.3 reports the predicted reduction in citations to each of these most-cited firms using the coefficients in Column (8) of Table 3. On average, these firms would be cited 7.9 percent less although there is substantial heterogeneity across the top 20 firms. Citations fall by less than 1 percent for the least affected firms (Apple, and Pelican Imaging) and more than 30 percent for the most (Stanford), with this heterogeneity depending on the size of the meetings reduction for each $i\tilde{j}$ pair, the number of establishments of the most-cited firms, and the distribution of citations across those establishments.

4.3 Robustness

In this section, we present a number of robustness checks and additional specifications. Table 4 reports these results, with the first column repeating our baseline specification (i.e. column 7 from from Table 3). In all cases we report IV specificatiosn that instrument $\ln TotalMP_{ij,t-1}$ with variants of $\ln TotalMP^{>5km}_{i'j',t-1}$ and include the various controls discussed above.

First, column (2) explores whether our results are robust to controlling for worker transitions by including the IHS of transitions between i and j during our sample period (recall from Case 2 in Section 3.2 these transitions may generate violations of our exclusion restriction). Reassuringly, the coefficient on $\ln TotalMP_{ij}$ is unchanged. Next, we investigate whether our findings are driven by establishments located close to each other for whom the meetings geography may be more salient to firm location decisions (recall our IV already restricts attention to meetings more than 5km from either establishment to mitigate this concern). Column (3) drops all establishments less than 5km apart with coefficient on $\ln TotalMP_{ij}$ rising slightly in magnitude (rather than falling as endogeneity in firm location decisions would suggest). The fact returns do not diminish beyond 5km is also of independent interest, standing in contrast to Arzaghi and Henderson (2008) who find that spatial spillovers in New York's advertising industry—which they attribute to face-to-face networking—tail off rapidly with distance.

Columns (4) and (5) explore different timing assumptions. Recall that meetings are lagged six months in our baseline specification. Given the short length of our time series, and the variability in the time taken to turn a knowledge flow into a citation, our estimates are best seen as cross-sectional regressions asking whether establishments whose workers meet more cite

each other more. Consistent with this interpretation, our results are relatively insensitive to a lag length of one year (column 4) or no lag at all (column 5).

The remaining columns investigate robustness to alternative ways of measuring our key variables. Columns (6)–(8) focus on the construction of our meetings measure. Recall that when there were no admissible $i'j'$ pairs in the donuts around i and j, we used information from ij' and $i'j$ meetings to construct our IV (7 percent of cases). Instead, column (6) treats these cases as missing. Next, we pursue an alternative method of dealing with zero-valued numerators in our logged meeting probability measures.[36] We simply add 0.01 (column 7) or 0.1 (column 8) to the numerators for both ij and $i'j'$ meeting probabilities. Results change little.

Columns (9)–(14) explore sensitivity to alternative citation measures, starting with how we allocate citations across a firm's establishments. Our baseline treats citations as establishment-to-establishment flows and allocates citations across establishments using the inventor's home-town. Column (9) reports a cruder alternative that divides firm-to-firm citations equally among their Silicon Valley establishments. The coefficients are still highly significant but half the size. This attenuation provides supportive evidence for our proposed mechanism: that face-to-face meetings generate knowledge flows that lead to patent citations. As this mechanism operates primarily through meetings involving workers at the inventors' establishments—most directly through the inventors themselves meeting—, assigning citations equally rather than based on inventors' predicted workplaces should lead to attenuation as we find.

Column (10) takes a different approach and implicitly assumes that all establishments of the firm work on every patent. Thus, citations are firm-to-firm objects and so we assign every citation to all (Silicon Valley) establishment pairs formed by the two firms.[37] Reassuringly, we still find a positive and significant coefficient. Unlike for column (9), the coefficient magnitudes are not directly comparable to our baseline; rather than reapportioning the same number of cites across establishments, here all firm cites are attributed to each establishment.

Column (11) uses an alternative citation measure that restricts attention only to citations included by the applicant at the time of submission and so may serve as cleaner proxy for knowledge flows.[38] This restriction leads us to drop 22 percent of citations marked as examiner citations and 0.5 percent from third parties or unknown sources. While the magnitude of the coefficient falls by a little over 20 percent (alongside the magnitude of the dependent variable), the estimate remains highly statistically significant.

[36] Recall we added small numbers to all numerators such that previously zero-valued probabilities lay at the 10th percentile of the new distribution on average (0.056 for ij and 0.013 for $i'j'$ meetings).

[37] As this approach results in multiple observations for multi-establishment firms, we down-weight establishment pairs so that each firm pair has a total weight of one. Given these weights and our firm-pair clustering, this specification is equivalent to regressing firm-to-firm citations on firm-to-firm meetings (averaged across establishments).

[38] One caveat is that citations added by examiners may have been knowledge known by the citing firm but not cited for strategic reasons (see (Lampe, 2012)), or may have been cited by both applicant and examiner (see Kuhn et al., 2020). Consistent with these possibilities, Alcacer and Gittelman (2006) and Thompson (2006) find that, excluding self-cites, examiner and applicant citations are equally skewed towards geographically-proximate firms.

Columns (12)–(14) explore the extensive and intensive margins of citation behavior. Column (12) replaces the dependent variable with a dummy for whether i cites j (i.e. whether $PatentCitations_{ijt} > 0$). The estimated coefficient is highly significant and about 5 times larger than our baseline. Column (13) reports an intensive margin regression of $\ln PatentCitations_{ij,t}$ on meetings (i.e. dropping all the zero-citation observations). Column (14) regresses $\text{arcsinh} PatentCitations_{ij,t}$ on meetings but excludes any firm that has no Silicon Valley inventors either citing or cited by other Silicon Valley inventors in our sample period (recall all our firms patent but may not have Silicon Valley cites in our sample period). The intensive margin coefficient is large compared to our baseline but also has a large standard error and is insignificant (the sample size is also dramatically smaller). The coefficient on the never cite/never cited subsample is ten times as large as our baseline coefficient and highly significant. Taken together, these results suggest that much of the action is on the extensive margin (whether an establishment cites another one, rather than how many citations) and among firms with Silicon Valley-based inventors working on related innovations.

5 Distinguishing the Effects of Serendipitous and Planned Meetings

Recall that the IV regressions above reveal LATE estimates that potentially combine the returns to both serendipitous meetings and to planned face-to-face meetings (those induced by our IV shifting the cost of organizing a meeting). In this section, we construct a test to determine whether our IV estimates measure the first of these, the returns to serendipity.

We formally assume the following structure that we alluded to in Section 3.1:

$$\ln PatentCitations_{ijt} = \gamma_1 f(ChanceMP_{ij,t-1}) + \gamma_2 g(PlannedMP_{ij,t-1}) + \Gamma X_{ij} + \epsilon_{ijt}, \tag{4}$$

$$f(ChanceMP_{ij,t-1}) = k(TotalMP_{i'j',t-1}, X_{ij}, \varsigma_{ij,t-1}), \tag{5}$$

$$g(PlannedMP_{ij,t-1}) = l(MeetingCost(TotalMP_{i'j',t-1}, \cdot), X_{ij}, \epsilon_{ijt}, u_{ij,t-1}), \tag{6}$$

$$\ln TotalMP_{ij,t-1} = f(PlannedMP_{ij,t-1}) + g(ChanceMP_{ij,t-1}). \tag{7}$$

Both the chance and planned meeting probabilities, $ChanceMP_{ij}$ and $PlannedMP_{ij}$, affect citation behavior. Chance meetings are a function of our instrument $TotalMP_{i'j'}$ that captures the meetings geography separating i and j. Planned meetings may also be a function of the instrument as the meetings geography reduces the cost of arranging meetings.[39] Finally, the error term ϵ_{ijt} in equation (4) may affect planned meetings, capturing the endogeneity concern that a worker at i finding a patent from j may schedule a meeting with the inventor at j to learn more.

[39] For example, if a popular coffee shop is conveniently located between two establishments, they may be more likely to arrange a face-to-face meeting rather than communicating via other modes or not communicating at all.

With this structure in hand, we define the two types of knowledge spillover as follows:

$$\frac{\partial \ln PatentCitations_{ijt}}{\partial f(ChanceMP_{ij,t-1})} = \gamma_1 \equiv \text{the return to serendipitous face-to-face meetings,}$$
$$\frac{\partial \ln PatentCitations_{ijt}}{\partial g(PlannedMP_{ij,t-1})} = \gamma_2 \equiv \text{the return to planned face-to-face meetings.}$$

Our regression of $\ln PatentCitations_{ijt}$ on $\ln TotalMP_{ij,t-1}$ instrumented by $\ln TotalMP_{i'j',t-1}$ reveals a weighted average of γ_1 and γ_2. If we could regress citations on measures of both chance and planned meetings, we could to disentangle the two returns. Sadly, our geolocation data offer no way to partition meetings into those that are planned and those that are not. Instead, we show that if we have suitable instruments for chance and planned meetings, we can test whether the previous LATE estimate uncovers the returns to serendipity γ_1.

Suppose we have two instruments Z^1_{ij} and Z^2_{ij}, both exogenous to ϵ_{ij}, with $f(ChanceMP_{ij}) = \kappa_1 Z^1_{ij} + \kappa_2 Z^2_{ij} + v_{ij}$ and $g(PlannedMP_{ij}) = \pi_1 Z^1_{ij} + \pi_2 Z^2_{ij} + \omega_{ij}$. Regressing citations on total meeting probabilities (i.e. specification 2), the two-stage least squares formula shows that we recover $\hat{\beta}^1$ when only using the instrument Z^1_{ij} and $\hat{\beta}^2$ when only using Z^2_{ij}:

$$\hat{\beta}^1 = \frac{(\gamma_1\kappa_1+\gamma_2\pi_1)+(\gamma_1\kappa_2+\gamma_2\pi_2)\frac{cov(Z^1_{ij},Z^2_{ij})}{var(Z^1_{ij})}}{(\kappa_1+\pi_1)+(\kappa_2+\pi_2)\frac{cov(Z^1_{ij},Z^2_{ij})}{var(Z^1_{ij})}}, \quad \hat{\beta}^2 = \frac{(\gamma_1\kappa_2+\gamma_2\pi_2)+(\gamma_1\kappa_1+\gamma_2\pi_1)\frac{cov(Z^1_{ij},Z^2_{ij})}{var(Z^2_{ij})}}{(\kappa_2+\pi_2)+(\kappa_1+\pi_1)\frac{cov(Z^1_{ij},Z^2_{ij})}{var(Z^2_{ij})}}.$$

The βs are different weighted average of γs. E.g., if Z^1_{ij} loads relatively heavily on planned meetings, $\frac{\pi_1}{\pi_2} > \frac{\kappa_1}{\kappa_2}$, and the returns to the planned meetings exceed the returns to chance, $\hat{\beta}^1 > \hat{\beta}^2$.

Proposition 1. *Under assumptions A.1 and A.2, $\hat{\beta}^1 = \hat{\beta}^2$ implies that both IV estimates recover the return to serendipitous meetings, $\hat{\beta}^1 = \hat{\beta}^2 = \gamma_1$.*

A1. The instruments differentially affect chance and planned meetings, $\frac{\pi_1}{\pi_2} \neq \frac{\kappa_1}{\kappa_2}$.

A2. At least one of our instruments affects chance meetings: κ_1 or $\kappa_2 \neq 0$.

Proof. $\hat{\beta}^1 = \hat{\beta}^2$ implies $\gamma_1[(\kappa_1\pi_2 - \kappa_2\pi_1)] = \gamma_2[(\kappa_1\pi_2 - \kappa_2\pi_1)]$. If $\gamma_1 = \gamma_2$, $\hat{\beta}^1 = \hat{\beta}^2 = \gamma_1$. If $\gamma_1 \neq \gamma_2$, $\kappa_1\pi_2 \neq \kappa_2\pi_1$ (A.1) and either $\kappa_1 \neq 0$ or $\kappa_2 \neq 0$ (A.2) $\implies \pi_1 = \pi_2 = 0 \implies \hat{\beta}^1 = \hat{\beta}^2 = \gamma_1$. □

Intuitively, if there is no treatment heterogeneity, our IV estimates recover both the return to serendipitous and planned meetings. If there is treatment heterogeneity and $\hat{\beta}^1 = \hat{\beta}^2$, the IVs must not be relevant for planned meetings or else estimates would differ since $\frac{\pi_1}{\pi_2} \neq \frac{\kappa_1}{\kappa_2}$. In other words, if we could regress planned meetings on both instruments, we would have no first stage.

Thus, if A.1 and A.2 hold, we can compare the magnitudes of $\hat{\beta}^1$ and $\hat{\beta}^2$, or more formally perform a standard over-identification test used in settings where you have more instruments than endogenous variables. If magnitudes are similar, i.e. the over-identification test is rejected, $\hat{\beta}^1 = \hat{\beta}^2 = \gamma_1$ and we have recovered the returns to serendipity.[40]

[40] Note that chance meetings may lead to planned ones. E.g., a worker bumps into an acquaintance and arranges

To implement this test, we require instruments that satisfy A.1 and A.2. Our approach rests on the insight that the subset of $i'j'$ meetings (i.e. those of the neighbors of i and j) that occur during the workday are likely stronger shifters of the costs of arranging ij planned meetings than the subset of $i'j'$ meetings outside these hours. Specifically, workday $i'j'$ proximity naturally occurs in locations that workers at i and j might visit during work hours and so make sensible work meeting locations (e.g. restaurants, coffee shops, or convention centers). In contrast, $i'j'$ proximity during the night and weekend mostly occurs at locations associated with leisure or family activities—e.g. in neighborhood parks, apartment complexes, schools or supermarkets—that are less appropriate for work meetings and not visited as part of a regular work schedule. Thus, A.1 is very likely to hold with $i'j'$ workday meetings loading relatively heavily on planned meetings.

In terms of A.2, both workday and night/weekend $i'j'$ meetings capture the meetings geography of the city that directly generates serendipitous encounters, with any effect on planned meetings more indirect and due to the meetings geography changing the cost of organizing work meetings. Thus, if either instrument is relevant (i.e. we have a first stage for total meetings, which is testable), it is almost certainly a relevant instrument for ij chance meetings.

Following this logic, we construct workday $i'j'$ total meeting probabilities as described in Section 3.3 except now restricting attention only to geolocation pings between 9am and 6pm local time, and construct $i'j'$ night/weekend meeting probabilities from pings at all other times. Figures 6a and 6b display Apple employees workday and night/weekend meeting probabilities with workers at all other establishments. In support of A.1, despite considerable overlap, workday meetings are relatively more common with workers at establishments nearby Apple.

5.1 Results using Workday and Night/Weekend Instruments

Table 5 presents regressions using the workday and night/weekend IVs that allow us to test whether we are recovering the returns to serendipity. Column (1) repeats our baseline specification (column 7 of Table 3) that instruments $\ln TotalMP_{ij,t-1}$ with $\ln TotalMP^{>5km}_{i'j',t-1}$. Column (2) reruns this specification but constructs the instrument using only the subset of meetings occurring during the workday and column (3) using only the subset occurring during the night and weekend (in both cases excluding $i'j'$ meetings less than 5km from i and j).

The coefficients are similar using either instrument and an over-identification test, Sargan's J test, cannot reject that they are the same (p=0.282). Proposition 1 implies that either the returns to serendipitous and planned meetings are the same ($\gamma_1 = \gamma_2$), or that there is no first stage for planned meetings ($\pi_1 = \pi_2 = 0$). In either case, we have recovered the returns to serendipity.

We provide some support for the second possibility—that our instrument does not induce

to meet again. Since the planned meeting is induced by the same variation driving chance meetings, the return to meetings induced by chance encounters would be subsumed into the LATE returns we attribute to serendipity.

planned meetings—by performing the same exercise with workday and night/weekend instruments but no longer excluding meetings occurring less than 5km from either establishment. If variation in the meetings geography of the city does affect the probability of arranging a planned meeting, the geography pertaining to locations close to either establishment (e.g. offices, coffee shops, lunch spots) is likely to be particularly important. Thus, there is more likely to be a first stage for planned meetings (with the caveat that the $i'j'$ variation within 5km is more worrisome from an endogeneity perspective).

Columns (4)–(6) of Table 5 present these results. While far from definitive, Sargan's J test is now closer to rejection (p=0.158) as would be the case if we had a first stage for planned meetings (i.e. $\pi_1 \neq 0$ and/or $\pi_2 \neq 0$) and their returns differed (i.e. $\gamma_1 \neq \gamma_2$). Additionally, we now find that the coefficient on total meetings is higher when using the workday rather than the night/weekend IV. Assuming we do have a first stage from workday meetings ($\pi_1 \neq 0$ where Z_{ij}^1 is the workday IV), this ordering implies that $\gamma_1 < \gamma_2$—consistent with our strong prior that the returns to planned meetings exceed the returns to serendipity. In contrast, that ordering that was reversed for the >5km IVs in columns (2) and (3) which implies the counterintuitive ordering $\gamma_1 > \gamma_2$ if there was a first stage.[41] Thus, this set of results suggest that our baseline estimates recover the return to serendipity because, once we exclude $i'j'$ meetings variation close to either workplace, there is no first stage for planned meetings.

In summary, we find that our estimates of the returns to face-to-face interactions in Section 4 reflect the returns to serendipity at the heart of Jane Jacobs' work, memorably summarized by the Glaeser (2009) quote reproduced in the introduction.

6 Conclusions

The goal of this paper is to make progress unpacking the black box that is knowledge spillovers. Leveraging rich smartphone geolocation data to capture meetings between workers at different establishments, we document that face-to-face interactions—instrumented by the meetings of workers in adjacent establishments in unconnected industries—substantially increase knowledge flows as measured by citation activity.

The importance of serendipitous meetings in generating knowledge flows is central to much of Jane Jacobs' work on the role cities play in economic development. However, our IV estimates potentially combine both the returns to serendipity and the returns to planned meetings induced by a favorable meetings geography. The second contribution of the paper is constructing and implementing a test for whether our estimates reflect the returns to serendipity by exploiting the differential loading of the workday and non-workday meetings geography on planned

[41] More precisely, here we are comparing the LATE returns to serendipity to the LATE returns to the types of planned meetings induced by changing the meetings geography of the city. The latter may be smaller or larger than the average returns to planned meetings. For example, shifting a conversation from online to face-to-face due to low meeting costs may have much lower returns than having two R&D teams collaborate in person.

and chance meetings. We find support for Jacobs' ideas with our estimates providing evidence for sizeable knowledge flows generated through serendipitous interactions.

Our analysis sheds light on several important policy questions regarding the forces of agglomeration in cities; the gains from encouraging more serendipitous interactions through better urban planning; and the costs of larger fractions of the workforce working from home. We leave questions of what types of interactions are the most fruitful and how to better design cities to maximize these to future work. More generally, highly granular geolocation data combined with existing establishment-level or firm-to-firm datasets can provide new insights into the agglomeration forces that generate our modern urban geography and much-studied patterns of industry clustering.

References

ALCACER, J. AND M. GITTELMAN (2006): "Patent Citations as a Measure of Knowledge Flows: The Influence of Examiner Citations," *Review of Economics and Statistics*, 88, 774–779.

ALTONJI, J. G., T. E. ELDER, AND C. R. TABER (2005): "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy*, 113, 151–184.

ALVAREZ, F. E., F. J. BUERA, AND J. LUCAS, ROBERT E (2013): "Idea Flows, Economic Growth, and Trade," Working Paper 19667, National Bureau of Economic Research.

ANDREWS, M. (2020): "Bar Talk: Informal Social Interactions, Alcohol Prohibition, and Invention," Working paper, University of Maryland Baltimore County.

ARZAGHI, M. AND J. V. HENDERSON (2008): "Networking off Madison Avenue," *The Review of Economic Studies*, 75, 1011–1038.

ATHEY, S., B. A. FERGUSON, M. GENTZKOW, AND T. SCHMIDT (2020): "Experienced Segregation," Working Paper 27572, National Bureau of Economic Research.

BARWICK, P. J., Y. LIU, E. PATACCHINI, AND Q. WU (2019): "Information, Mobile Communication, and Referral Effects," Working Paper 25873, National Bureau of Economic Research.

BÜCHEL, K. AND M. V. EHRLICH (2020): "Cities and the Structure of Social Interactions: Evidence from Mobile Phone Data," *Journal of Urban Economics*, 119, 103276.

CATALINI, C. (2018): "Microgeography and the Direction of Inventive Activity," *Management Science*, 64, 4348–4364.

CATALINI, C., C. FONS-ROSEN, AND P. GAULE (2020): "How Do Travel Costs Shape Collaboration?" *Management Science*, 66, 3340–3360.

CHARLOT, S. AND G. DURANTON (2006): "Cities and Workplace Communication: Some Quantitative French Evidence," *Urban Studies*, 43, 1365–1394.

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.

DURANTON, G. AND D. PUGA (2001): "Nursery Cities: Urban Diversity, Process Innovation, and the Life Cycle of Products," *American Economic Review*, 91, 1454–1477.

ELLISON, G., E. L. GLAESER, AND W. R. KERR (2010): "What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns," *American Economic Review*, 100, 1195–1213.

GLAESER, E. L. (2009): "What A City Needs," *The New Republic*, 240, 42–45.

GLAESER, E. L., H. D. KALLAL, J. A. SCHEINKMAN, AND A. SHLEIFER (1992): "Growth in Cities," *Journal of Political Economy*, 100, 1126–1152.

JACOBS, J. (1961): *The Death and Life of Great American Cities*, Vintage.

——— (1969): *The Economy of Cities*, Vintage.

——— (1984): *Cities and the Wealth of Nations: Principles of Economic Life*, Vintage.

JAFFE, A. B., M. TRAJTENBERG, AND M. S. FOGARTY (2000): "Knowledge Spillovers and Patent Citations: Evidence from a Survey of Inventors," *American Economic Review*, 90, 215–218.

JAFFE, A. B., M. TRAJTENBERG, AND R. HENDERSON (1993): "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations," *The Quarterly Journal of Economics*, 108, 577–598.

JOVANOVIC, B. AND R. ROB (1989): "The Growth and Diffusion of Knowledge," *The Review of Economic Studies*, 56, 569–582.

KRUGMAN, P. R. (1991): *Geography and Trade*, MIT Press.

KUHN, J., K. YOUNGE, AND A. MARCO (2020): "Patent Citations Reexamined," *The RAND Journal of Economics*, 51, 109–132.

LAMPE, R. (2012): "Strategic Citation," *Review of Economics and Statistics*, 94, 320–333.

LEAMER, E. E. (2007): "A Flat World, a Level Playing Field, a Small World After All, or None of the Above? A Review of Thomas L Friedman's The World is Flat," *Journal of Economic Literature*, 45, 83–126.

MARSHALL, A. (1920): *Principles of Economics*, MacMillan.

MORETTI, E. (2012): *The New Geography of Jobs*, Houghton Mifflin Harcourt.

——— (2021): "The Effect of High-Tech Clusters on the Productivity of Top Inventors," *American Economic Review*, 111, 3328–75.

PAULY, S. AND F. STIPANICIC (2022): "The Creation and Diffusion of Knowledge: Evidence from the Jet Age," Working paper, Toulouse School of Economics.

ROACH, M. AND W. M. COHEN (2013): "Lens or Prism? Patent Citations as a Measure of Knowledge Flows from Public Research," *Management Science*, 59, 504–525.

ROSENTHAL, S. S. AND W. C. STRANGE (2001): "The Determinants of Agglomeration," *Journal of Urban Economics*, 50, 191–229.

SAXENIAN, A. (1996): *Regional Advantage*, Harvard University Press.

STARTZ, M. (2021): "The Value of Face-to-Face: Search and Contracting Problems in Nigerian Trade," Working paper, Dartmouth.

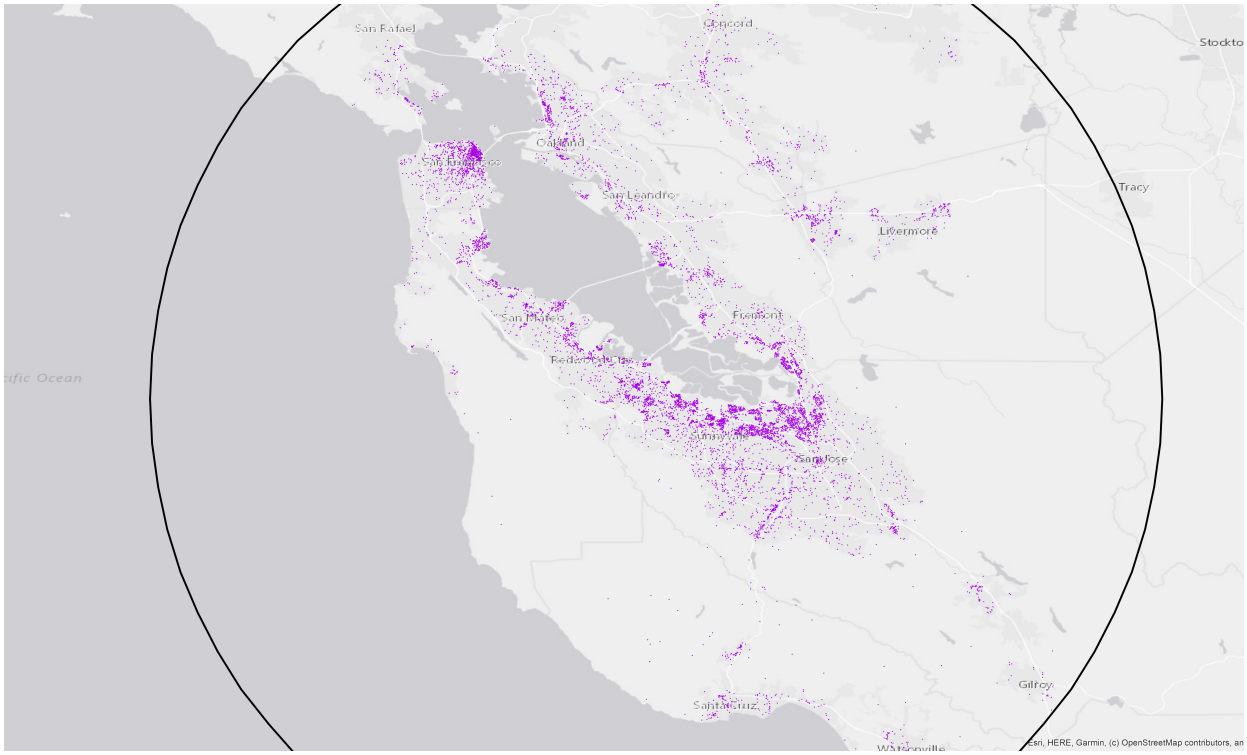
STORPER, M. AND A. J. VENABLES (2004): "Buzz: Face-to-Face Contact and the Urban Economy," *Journal of Economic Geography*, 4, 351–370.

THOMPSON, P. (2006): "Patent Citations and the Geography of Knowledge Spillovers: Evidence from Inventor-and Examiner-Added Citations," *Review of Economics and Statistics*, 88, 383–388.

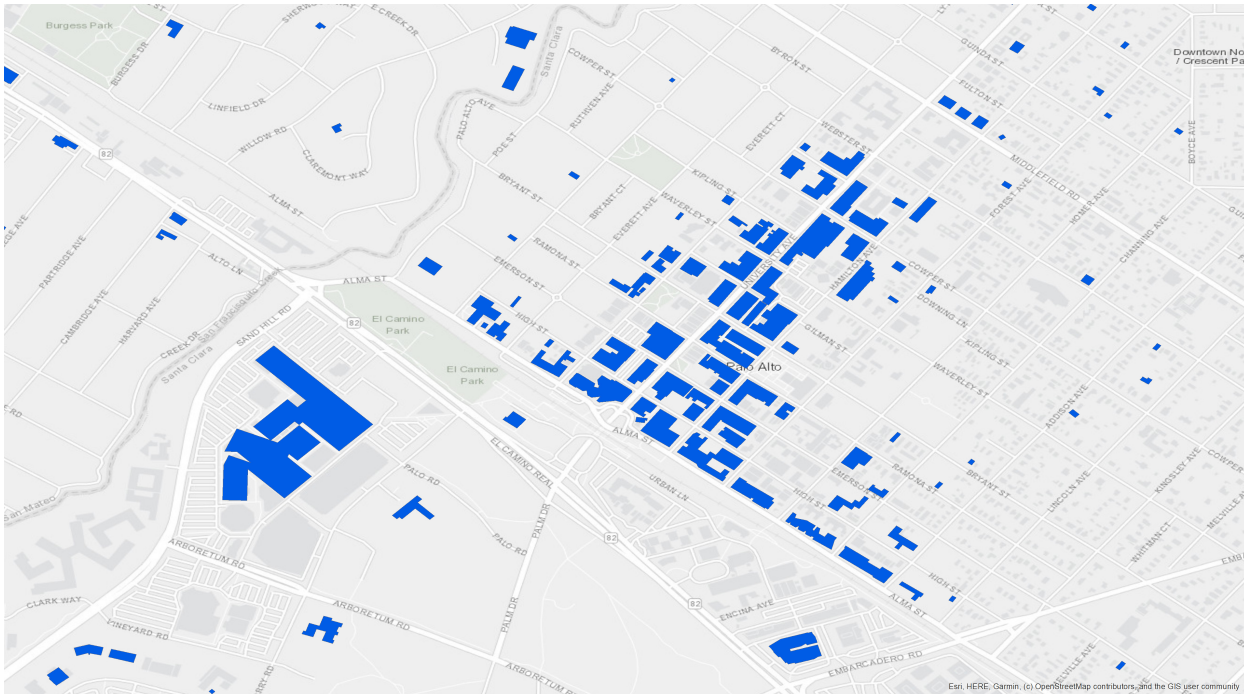
THOMPSON, P. AND M. FOX-KEAN (2005): "Patent Citations and the Geography of Knowledge Spillovers: A Reassessment," *American Economic Review*, 95, 450–460.

Figure 1: Patenting establishments

(a) All patenting establishments within 50 miles of Stanford University

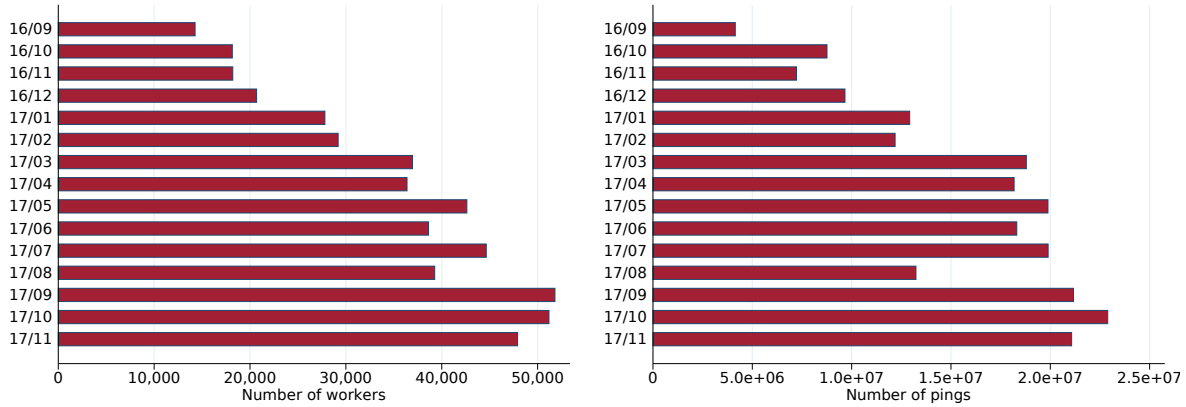


(b) Buildings in Palo Alto containing patenting establishments



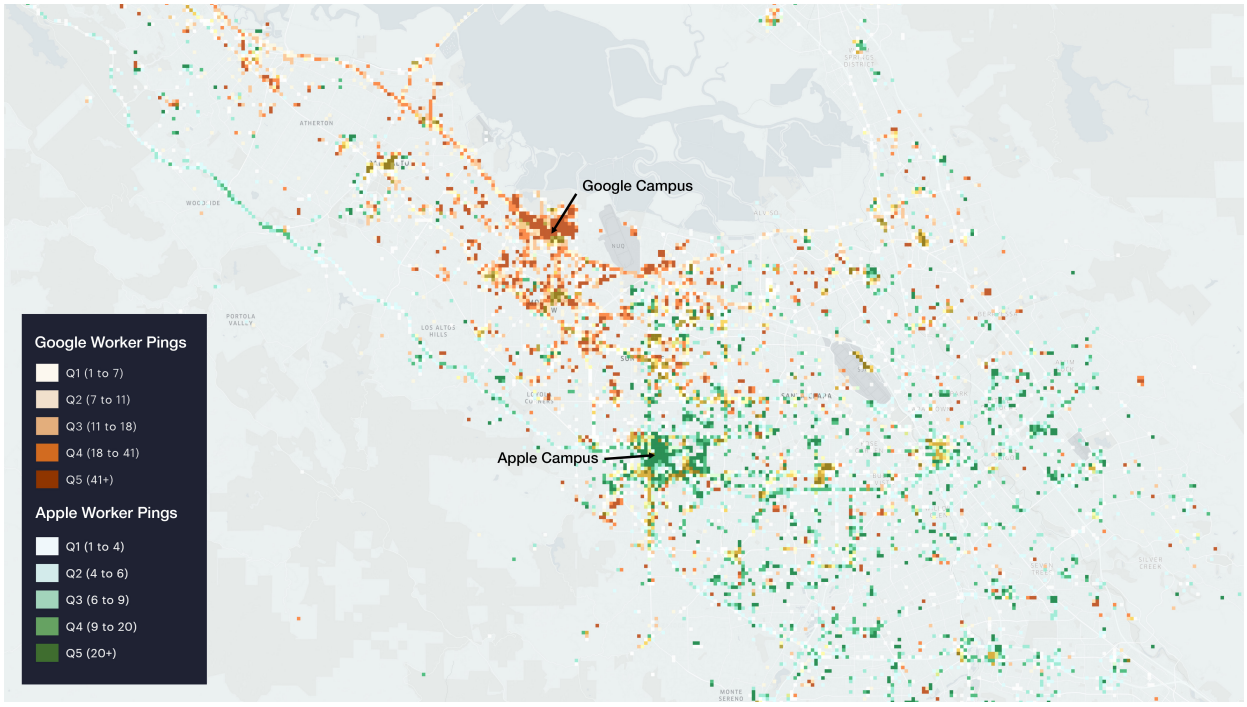
Notes: Figure 1a plots patenting establishments from Orbis located within 50 miles of Stanford University. Figure 1b marks rooftops of buildings in Palo Alto containing patenting establishments, using Microsoft's rooftop shapefile.

Figure 2: Counts of workers in Silicon Valley patenting establishments and their pings, September 2016–November 2017



Notes: Figure plots worker counts and pings by month, with workers identified from smartphones that leave pings in an establishment of a patenting firm in at least 20 different hours in that month.

Figure 3: Overlay of pings for workers at Apple and Google



Notes: Figure shows overlays of pings for workers at Apple's and Google's headquarters. Apple worker pings are shades of green, Google worker pings are shades of orange, and browns denote overlapping pings. Figure displays quintiles of Apple's and Google's respective ping counts, with darker shades denoting more pings.

Figure 4: Apple and Google meetings measures

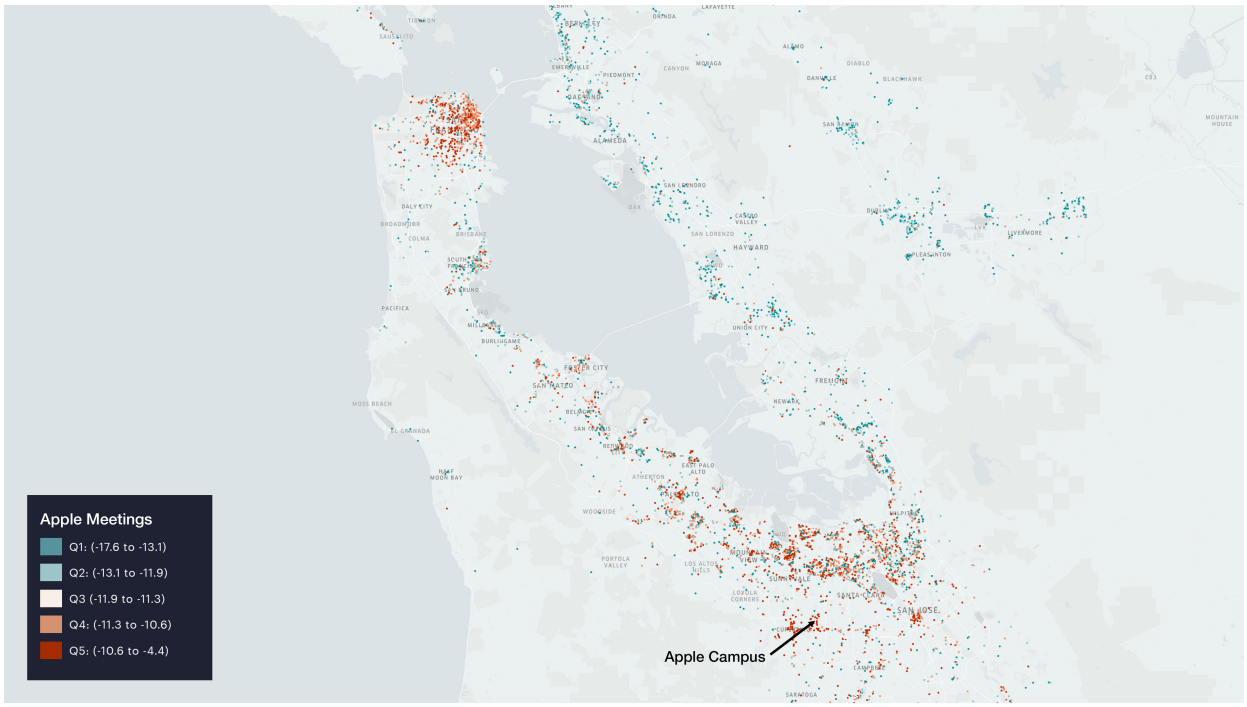
(a) Coincidences of workers from Apple and Google ($TotalMeetings_{ij}$)

(b) Coincidences of workers from firms adjacent to Apple and Google ($TotalMeetings_{i'j'}$)

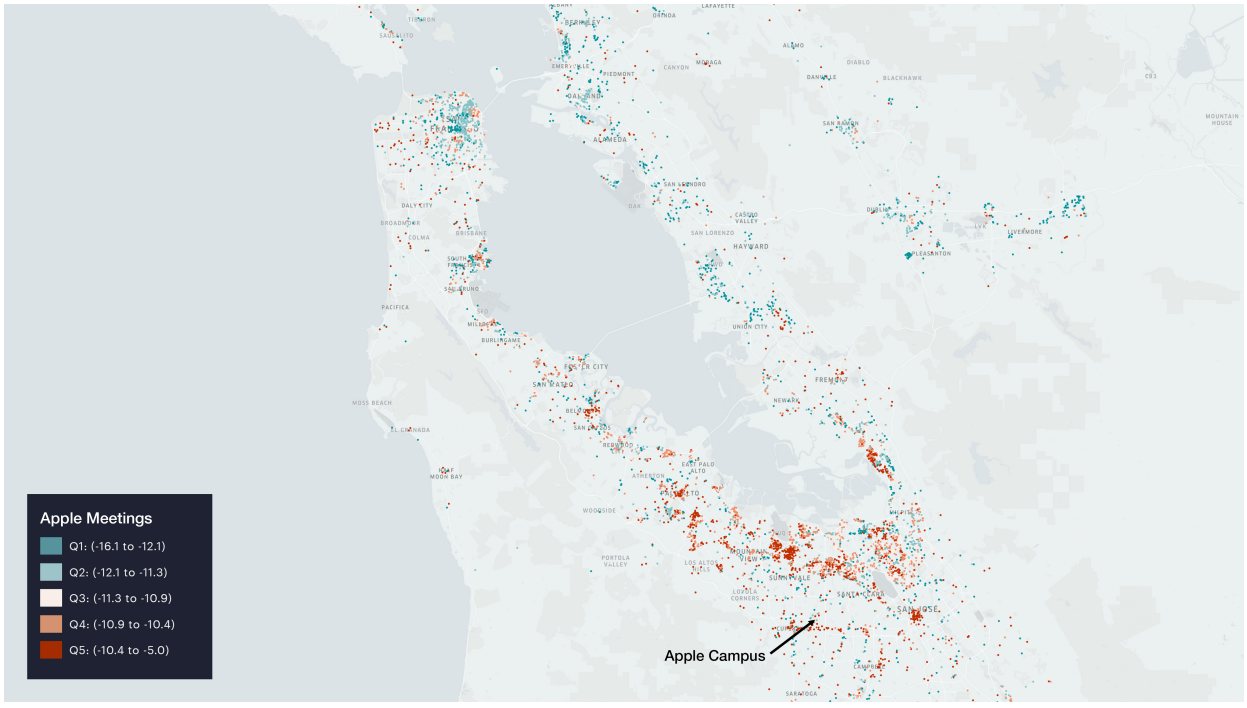
Notes: Figure 4a shows $TotalMeetings_{ij}$, i.e. 'coincidences' where workers from Apple and Google are in the same location at the same time. Figure 4b shows $TotalMeetings_{i'j'}$, coincidences between workers at establishments i' within 1km of Apple and j' within 1km of Google, restricting attention to $i'j'$ pairs whose industries neither cite nor supply each other. Meetings within 5km radius circles around either Apple (marked with a green circle) or Google (an orange circle) are excluded in the $TotalMeetings_{i'j'}^{>5km}$ measure. For both figures, reds denote more coincidences, blues fewer, and labels indicate the locations of the coincidences.

Figure 5: $\ln TotalMP_{ij}$ and $\ln TotalMP_{i'j'}^{>5km}$ variation between Apple and other establishments

(a) $\ln TotalMP_{ij}$ variation between Apple and all other establishments



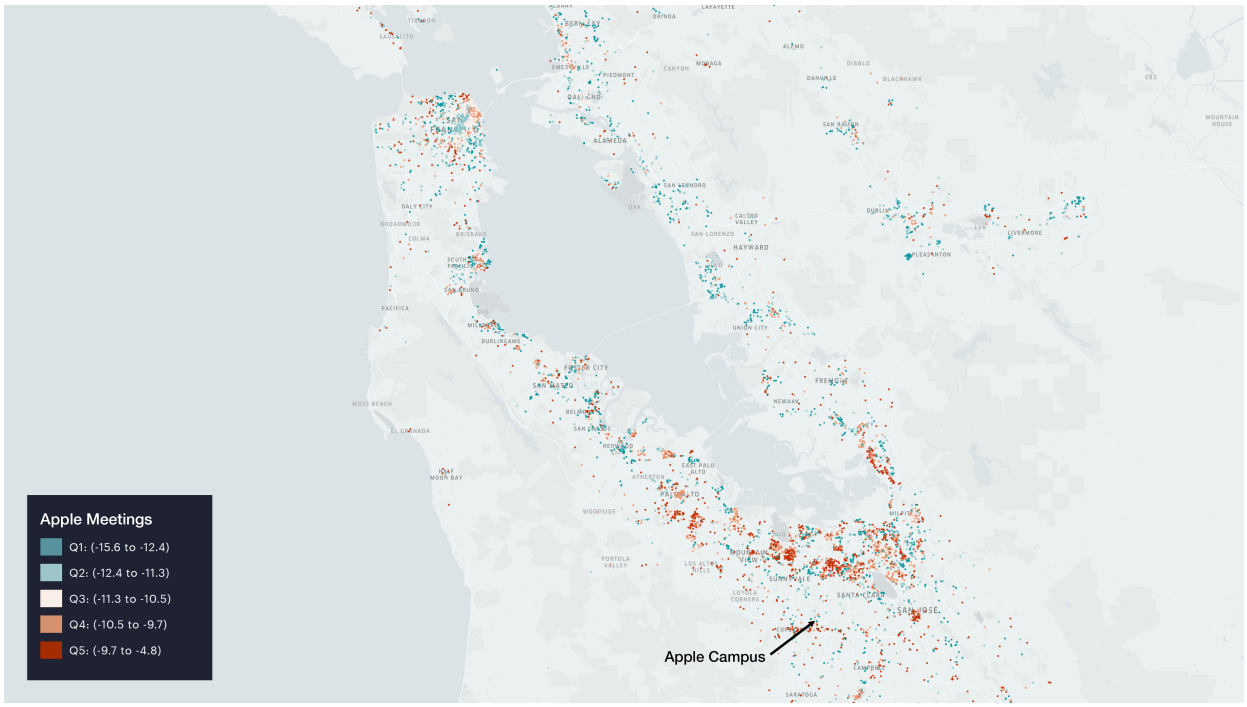
(b) $\ln TotalMP_{i'j'}^{>5km}$ variation between Apple and all other establishments



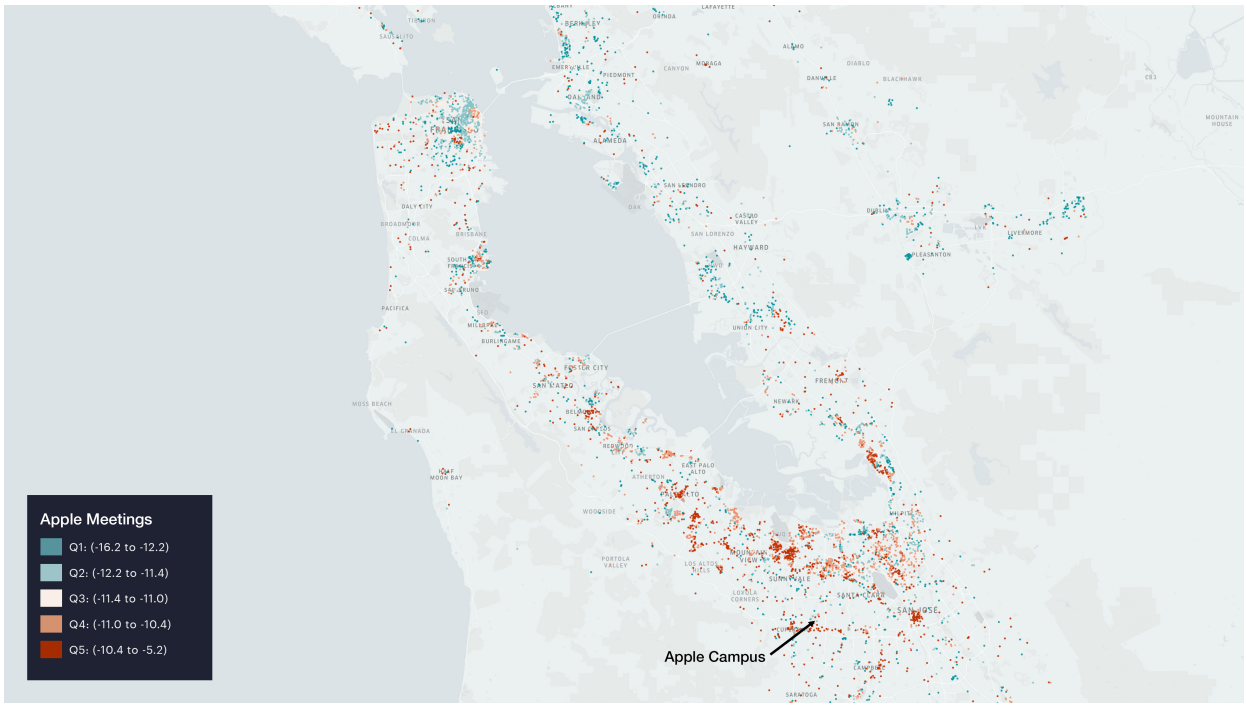
Notes: Figure 5a shows the log meeting probabilities between the headquarters of Apple and all other establishments in our sample (i.e. $\ln TotalMP_{ij}$ for i=Apple HQ and all j). Figure 5b shows our instrument, the log meeting probabilities calculated using meetings between workers at establishments i' within 1km of Apple HQ and establishments j' within 1km of establishment j, restricting attention to $i'j'$ pairs whose industries neither cite nor supply each other and removing meetings occurring within 5km of either i or j (i.e. $\ln TotalMP_{i'j'}^{>5km}$). Both figures display quintiles of their respective meeting probabilities, with reds denoting higher probabilities and blues denoting lower probabilities.

Figure 6: Workday and night/weekend $\ln TotalMP_{i'j'}^{>5km}$ variation between Apple and others

(a) Workday $\ln TotalMP_{i'j'}^{>5km}$ variation between Apple and all other establishments



(b) Night and weekend $\ln TotalMP_{i'j'}^{>5km}$ variation between Apple and all other establishments



Notes: Figures show the log meeting probabilities calculated using meetings between workers at establishments i' within 1km of Apple HQ and establishments j' within 1km of establishment j, restricting attention to $i'j'$ pairs whose industries neither cite nor supply each other and removing meetings occurring within 5km of either i or j. Figure 6a shows these meeting probabilities calculated solely from meetings occurring Monday–Friday 9am–6pm and Figure 6b shows these meeting probabilities calculated from meetings occurring at all other times. Both figures display quintiles of their respective meeting probabilities, with reds denoting higher probabilities and blues denoting lower probabilities.

Table 1: Representativeness of the smartphone data

	Med. Age		Med. Income		% College		% Black		% Hispanic	
Bay-Area County	Cen.	Imp.	Cen.	Imp.	Cen.	Imp.	Cen.	Imp.	Cen.	Imp.
Alameda	37.3	37.5	85.7k	84.2k	44.7	44.4	11.1	11.7	22.5	22.5
San Francisco	38.3	38.1	96.3k	99.3k	55.8	57.0	5.28	5.49	15.3	17.2
San Mateo	39.6	40.0	106k	103k	48.5	47.8	2.43	2.36	24.9	24.8
Santa Clara	37.0	36.9	107k	107k	50.0	49.0	2.53	2.59	26.1	26.7
All four counties together ("Silicon Valley")					49.3	49.1	5.93	5.67	23.0	23.0

Notes: Table compares census demographics ("Cen." columns) from the 2013–2017 ACS for the four Bay Area counties that make up Silicon Valley with imputed demographics ("Imp." columns) based on the distribution of smartphone pings across census block groups within these counties. We consider five demographic variables; median age, median household income, percent with some college, percent Black and percent Hispanic. For each smartphone owner in our sample who likely lives in Silicon Valley (N = 426,955), their imputed demographics were constructed using their likely home census block group based on their regular nighttime location. These demographics are then aggregated to the county level, and compared with actual county-level census statistics to asses the representativeness of our smartphone sample. Using county populations, our smartphone sample can also be compared to all four counties at once (which we call Silicon Valley) for our three mean-based statistics.

Table 2: First stage regressions: $\ln TotalMP_{ij}$ on $\ln TotalMP_{i'j'}$

	$\ln TotalMP_{ij}$		
	(1)	(2)	(3)
$\ln TotalMP_{i'j'}$	0.378***		
	(0.00013)		
$\ln TotalMP^{>5km}_{i'j'}$		0.221***	0.0594***
		(0.000114)	(0.00009)
$SameIndustry_{ij}=1$	0.115***	0.138***	0.0699***
	(0.00085)	(0.000960)	(0.00066)
3 distance cubics			✓
$\Delta_{ij} Demographics$ cubics			✓
Establishment i and j FEs	✓	✓	✓
Observations	220,001,468	219,103,680	218,084,204
R-squared	0.333	0.280	0.407
First-stage F	8,454,675	3,758,156	429,850

Notes: Table shows regressions of the log meeting probability $\ln TotalMP_{ij}$ (i.e. the probability that a worker from i meets a worker from j) on our instrument $\ln TotalMP_{i'j'}$—the same object calculated from the meetings between workers at establishments i' adjacent to i and j' adjacent to j, excluding any $i'j'$ pairs where their industries cite or supply each other. As we only have one period of meetings observations collected between September 2016 and November 2017, we omit period $t-1$ subscripts. All columns include establishment i and j fixed effects and a same-industry dummy. Columns (2)–(3) replace $\ln TotalMP_{i'j'}$ with $\ln TotalMP^{>5km}_{i'j'}$ which further restricts the meetings used to calculate the adjacent-establishment meeting probabilities to those occurring more than 5km away from either establishment i or j. Column (3) additionally includes cubics in distance, road distance, and travel time between i and j, as well as cubic controls for each of 11 demographic differences between workers at establishments i and j. Standard errors clustered at the level of i and j's firms shown in parentheses.

Table 3: Second stage regressions: Citations on face-to-face meetings

	arcsinh$PatentCitations_{ij,t}$							
	OLS	OLS	OLS	IV $TotalMP_{i'j',t-1}$	IV Split Sample	IV $TotalMP^{>5km}_{i'j',t-1}$	IV $TotalMP^{>5km}_{i'j',t-1}$	IV $TotalMP^{>5km}_{i'j',t-1}$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ln$TotalMP_{ij,t-1}$	4.65e-05***	4.26e-05***	4.18e-05***	7.17e-05***	8.79e-05***	7.56e-05***	0.000102***	0.000116***
	(1.75e-06)	(1.67e-06)	(1.65e-06)	(4.20e-06)	(3.28e-06)	(5.42e-06)	(1.83e-05)	(1.97e-05)
ln$TotalMP_{ij,t-1}\times$								2.76e-05***
$\ln\sqrt{Workers_i}\sqrt{Workers_j}$								(5.27e-06)
$SameIndustry_{ij}=1$	0.00111***	0.00110***	0.00111***	0.00109***	0.00110***	0.00110***	0.00110***	0.00110***
	(5.92e-05)	(5.92e-05)	(5.94e-05)	(5.92e-05)	(5.92e-05)	(5.92e-05)	(5.95e-05)	(5.93e-05)
3 distance cubics		✓					✓	✓
$\Delta_{ij}Demographics$ cubics			✓				✓	✓
Establishment i and j FEs	✓	✓	✓	✓	✓	✓	✓	✓
Observations	222,401,518	222,401,016	221,341,416	220,001,484	222,113,164	219,103,680	218,084,204	218,084,204
R-squared	0.012	0.012	0.012	0.000	0.000	0.000	0.000	0.000
First-stage F				2.040e+07	6.750e+07	6.034e+06	478,335	250,238

Notes: Table shows cross-sectional regressions of arcsinh$PatentCitations_{ij,t}$, the inverse hyperbolic sine of patent citations between establishments i and j allocated from firm-level patents using inventor hometowns, on the log meeting probabilities ln$TotalMP_{ij,t-1}$, the probability a worker from i meets a worker from j face-to-face. Citations come from patent applications filed between March 2017 and May 2018 (period t), meetings measures calculated using smartphone geolocation data collected between September 2016 and November 2017 (period $t-1$). All columns include establishment i and j fixed effects and a same-industry dummy. Column (1) reports the OLS. Column (2) includes cubics in distance, road distance and travel time between i and j, and column (3) includes cubics for each of 11 demographic differences between workers at establishments i and j. Column (4) instruments ln$TotalMP_{ij,t-1}$ with ln$TotalMP_{i'j',t-1}$, the log meeting probability calculated from the meetings between workers at establishments i' adjacent to i and j' adjacent to j, excluding any $i'j'$ pairs whose industries cite or supply each other. Column (5) uses a split sample IV, calculating ln$TotalMP_{ij,t-1}$ from meetings occurring in odd-numbered geo7s and instrumenting with ln$TotalMP_{ij,t-1}$ calculated from meetings in even-numbered geo7s. Column (6) instruments ln$TotalMP_{ij,t-1}$ with ln$TotalMP^{>5km}_{i'j',t-1}$ which further restricts the meetings used to calculate the adjacent-establishment meeting probabilities to those more than 5km away from either establishment i or j. Column (7) uses this same instrument but additionally includes the distance and demographic controls introduced in columns (2) and (3). Column (8) repeats this specification but adds an interaction between ln$TotalMP_{ij,t-1}$ and the product of the number of workers at each establishment, $\ln\sqrt{Workers_i}\sqrt{Workers_j}$ (instrumented by the the interaction between ln$TotalMP^{>5km}_{i'j',t-1}$ and $\ln\sqrt{Workers_i}\sqrt{Workers_j}$). Main effect of $\ln\sqrt{Workers_i}\sqrt{Workers_j}$ swept out by the i and j fixed effects. Standard errors clustered at the level of i and j's firms shown in parentheses.

Table 4: Second stage regressions: Robustness

	arcsinh$PatentCitations_{ij,t}$						
	IV (>5km) Baseline	Transition controls	$Dist_{ij}>5$km	One year lag	Pure cross-section	Only ($i'j'$) used	Add 0.01 to numerators
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ln$TotalMP_{ij,t-1}$	0.000102***	0.000102***	0.000108***	8.00e-05***	9.82e-05***	9.62e-05***	8.46e-05***
	(1.83e-05)	(1.84e-05)	(1.82e-05)	(1.60e-05)	(1.69e-05)	(2.61e-05)	(1.58e-05)
Controls and FEs	✓	✓	✓	✓	✓	✓	✓
Observations	218,084,204	218,084,204	200,693,626	218,084,204	218,084,204	202,371,810	218,084,204
R-squared	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Cragg-Donald Wald F-stat	478,335	478,510	470,621	478,335	478,335	238,425	352,457

	arcsinh$PatentCitations_{ij,t}$				$PatentCitations_{ij,t}$		
	Add 0.1 to numerators	Equal alloc. $Citations_{ij}$	Firm-to-firm $Citations_{ij}$	Only applicant $Citations_{ij}$	Extensive margin	Intensive margin	No never cite/cited
	(8)	(9)	(10)	(11)	(12)	(13)	(14)
ln$TotalMP_{ij,t-1}$	8.31e-05***	5.45e-05***	0.000892***	7.70e-05***	0.000531***	0.0186	0.000996***
	(1.24e-05)	(1.75e-05)	(0.000298)	(1.68e-05)	(8.13e-05)	(0.0189)	(0.000291)
Controls and FEs	✓	✓	✓	✓	✓	✓	✓
Observations	218,084,204	218,084,204	218,084,204	218,084,204	218,084,204	367,879	19,625,110
R-squared	0.000	0.000	0.005	0.000	0.003	-0.000	0.001
Cragg-Donald Wald F-stat	$1.34{\cdot}10^6$	478,335	478,335	478,335	478,335	141	18,468

Notes: Table explores the robustness of our baseline specification that regresses the inverse hyperbolic sine of patent citations between establishments i and j, arcsinh$PatentCitations_{ij,t}$, on the log meeting probability ln$TotalMP_{ij,t-1}$. Column (1) repeats our baseline specification (column 7 of Table 3). Subsequent columns modify the baseline specification with modification noted in the column header and described in more detail in Section 4.3. All columns use the IV specification instrumenting ln$TotalMP_{ij,t-1}$ with ln$TotalMP^{>5km}_{i'j',t-1}$, the log meeting probability calculated from the meetings between workers at establishments i' adjacent to i and j' adjacent to j, excluding any $i'j'$ pairs whose industries cite or supply each other and removing meetings occurring more than 5km away from either establishment i or j. All columns also include cubics in distance, road distance and travel time between i and j, cubics for each of 11 demographic differences between workers at establishments i and j, a same industry dummy, and establishment i and j fixed effects. Standard errors clustered at the level of i and j's firms shown in parentheses.

Table 5: Testing whether the LATE coefficient equals the returns to serendipity

	arcsinh$PatentCitations_{ij,t}$					
	Meetings $>$ 5km from i or j (ln$TotalMP^{>5km}_{i'j',t-1}$)			Meetings $\geq$ 0km from i or j (ln$TotalMP_{i'j',t-1}$)		
	Day & Night IV	Z^1_{ij} Day IV	Z^2_{ij} Night IV	Day & Night IV	Z^1_{ij} Day IV	Z^2_{ij} Night IV
	(1)	(2)	(3)	(4)	(5)	(6)
ln$TotalMP_{ij,t-1}$	0.000102***	8.48e-05***	0.000105***	9.47e-05***	0.000102***	8.16e-05***
	(1.83e-05)	(1.68e-05)	(1.84e-05)	(1.35e-05)	(1.25e-05)	(1.38e-05)
Controls and FEs	✓	✓	✓	✓	✓	✓
Observations	218,084,204	216,523,520	217,822,128	218,970,004	218,897,870	218,961,576
R-squared	0.000	0.000	0.000	0.000	0.000	0.000
Sargan's J Test		p=0.282			p=0.158	
Cragg-Donald Wald F-stat	478,335	796,203	479,350	1.034e+06	1.648e+06	975,567

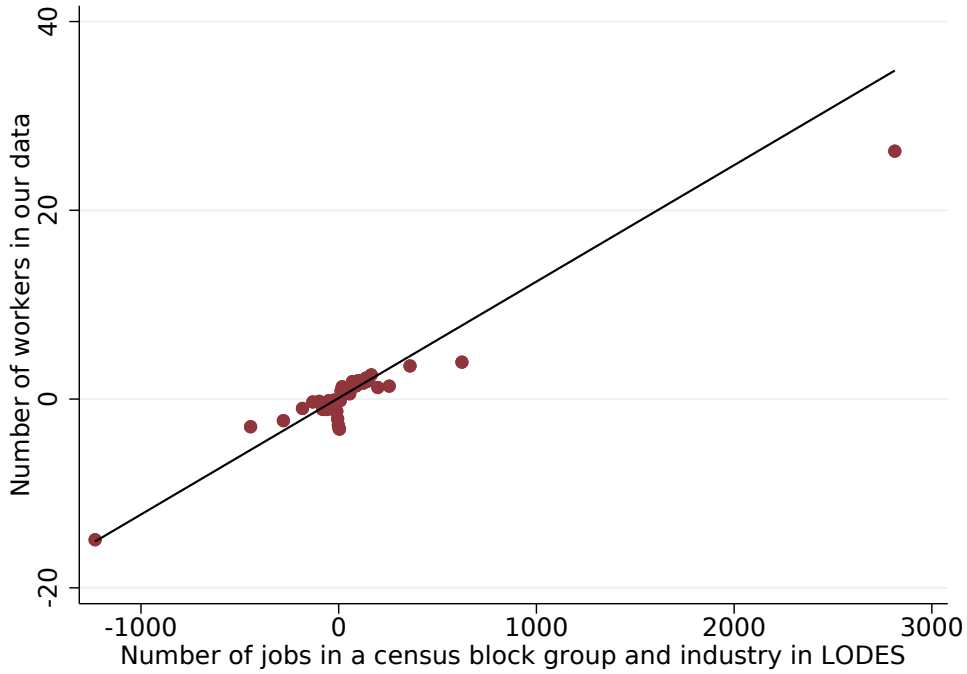
Notes: Table shows regressions of the inverse hyperbolic sine of patent citations between establishments i and j, arcsinh$PatentCitations_{ij,t}$, on the log meeting probability ln$TotalMP_{ij,t-1}$ instrumented either by night/weekend meeting probabilities or workday meeting probabilities of workers at adjacent establishments i' and j' (as previously, using meetings of $i'j'$ pairs whose industries neither cite nor supply each other and excluding meetings occurring less than 5km away from either establishment i or j). Column (1) repeats our baseline IV specification using ln$TotalMP^{>5km}_{i'j',t-1}$ that combines both workday and night/weekend meetings (column 7 of Table 3). Column (2) reports the IV regression that uses only workday meeting probabilities of workers at adjacent establishments. Column (3) reports the IV regression using only night/weekend meeting probabilities of workers at adjacent establishments. Columns (4)–(6) repeat the previous three columns but also including meetings occurring between 0km and 5km from either establishment i or j. All columns include cubics in distance, road distance and travel time between i and j, cubics for each of 11 demographic differences between workers at establishments i and j, a same industry dummy, and establishment i and j fixed effects. Standard errors clustered at the level of i and j's firms shown in parentheses.

Online Appendix to: The Returns to Face-to-Face Interactions: Knowledge Spillovers in Silicon Valley

David Atkin, Keith Chen and Anton Popov

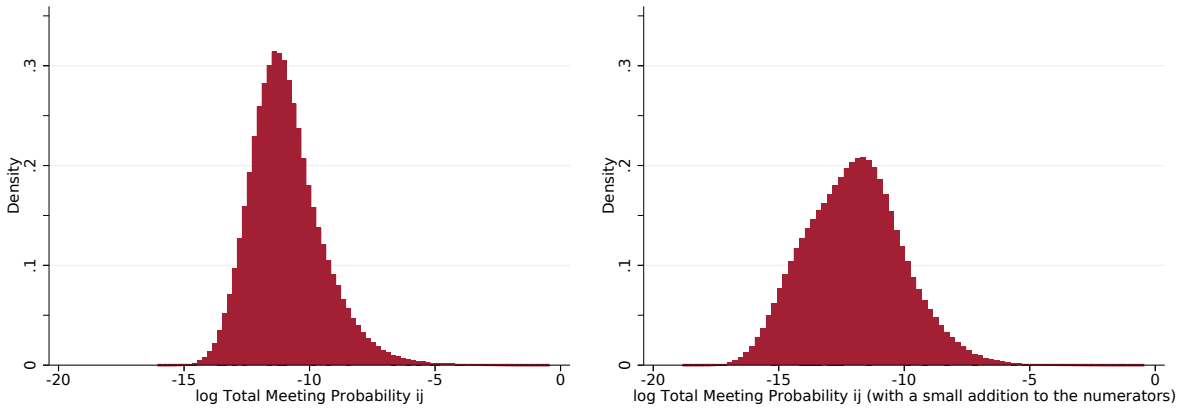
A Additional Figures and Tables

Figure A.1: Comparing worker counts in smartphone geolocation data and LODES



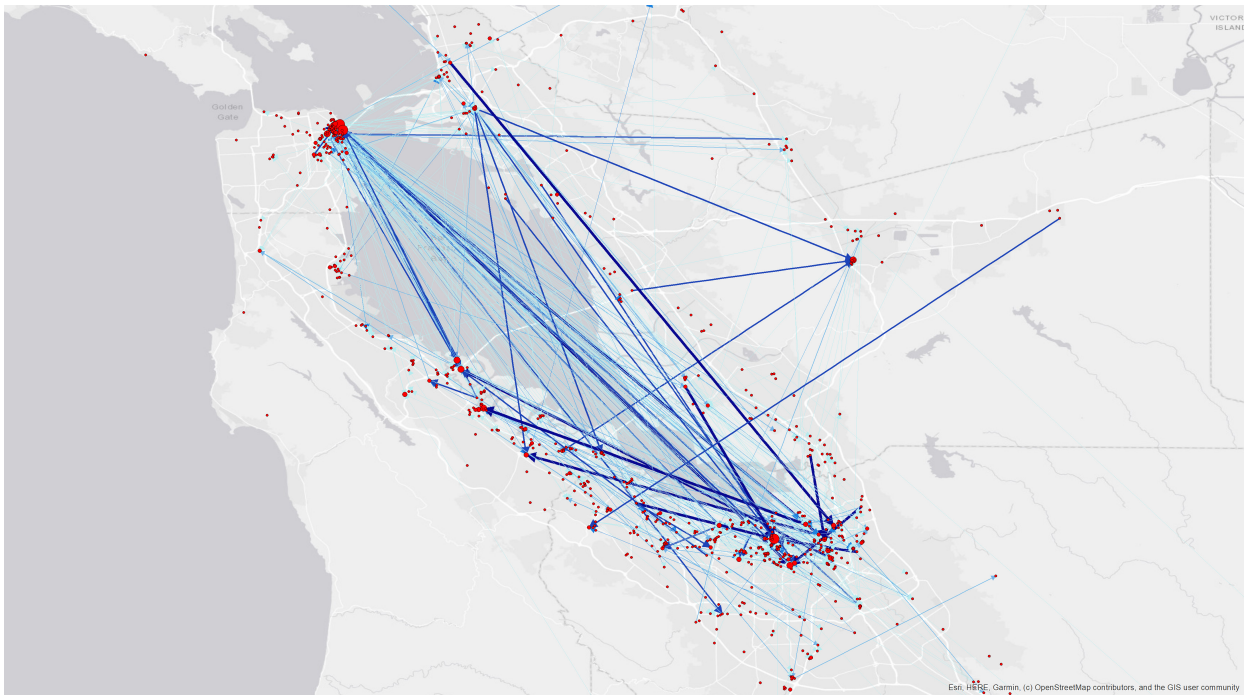
Notes: Figure plots a binscatter of worker counts estimated from our smartphone geolocation data matched to patenting establishment rooftops for each census block group c and industry i on workers in the LEHD Origin-Destination Employment Statistics in that same cell, after residualizing on census block group and industry fixed effects. Slope=0.0123, s.e.=0.0001, within-R^2=0.227.

Figure A.2: $\ln TotalMP_{ij}$ distribution pre and post adding a small number to all numerators



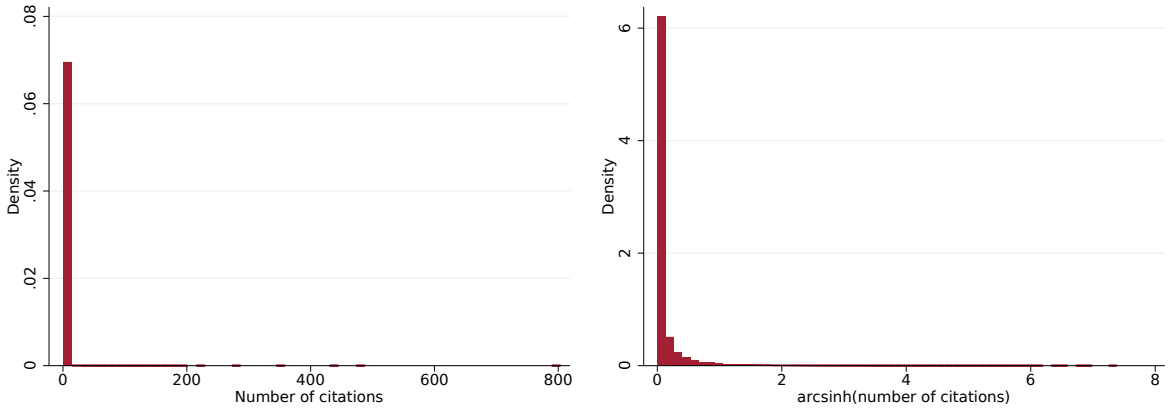
Notes: Left panel: distribution of $\ln TotalMP_{ij}$ when the numerator is non-zero. Right panel: distribution of $\ln TotalMP_{ij}$ where a small number (0.0561) is added to all $TotalMP_{ij}$ numerators with the number chosen such that the mean of the previously zero-numerator observations lies at the 10th percentile of the full post-addition distribution.

Figure A.3: A sample of citation links between Silicon Valley establishments



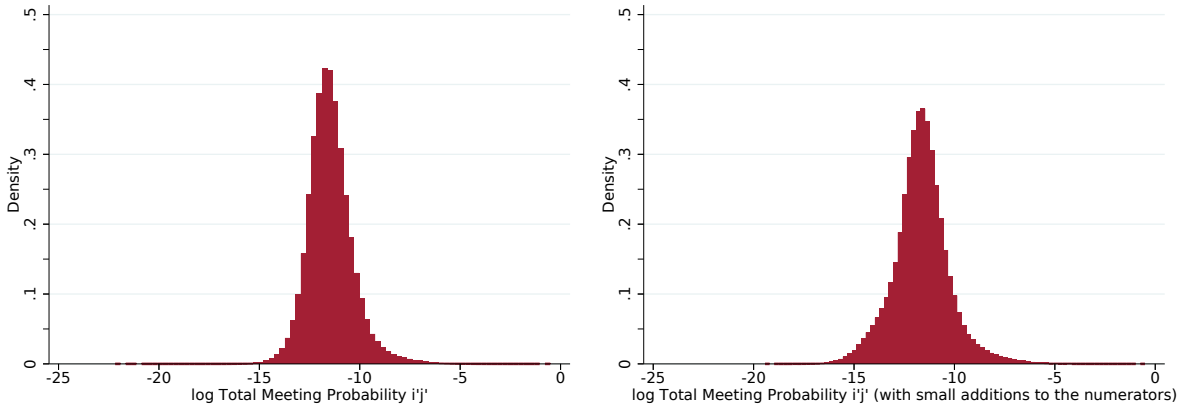
Notes: Figure shows citations between a random sample of 1000 firms in Silicon Valley using USPTO patent application data from March 2017 to May 2018. Arrow originates from citing firm and points towards cited firm. Thicker arrows represent more cites

Figure A.4: Citation distributions for non-zero ij citations



Notes: Left panel: distribution of non-zero ij citations from patent applications filed between March 2017 and May 2018. Right panel: inverse hyperbolic sine of non-zero ij citations (arcsinh$x = \ln(x+\sqrt{x^2+1})$).

Figure A.5: $\ln TotalMP_{i'j'}^{>5km}$ distribution pre and post adding a small number to all numerators



Notes: Left panel: distribution of $\ln TotalMP_{i'j'}^{>5km}$ based on non-missing observations from ($D(i)$,$D(j)$) measure, and the average of (i,$D(j)$) and (j,$D(i)$) measures if ($D(i)$,$D(j)$) is missing. Right panel: distribution of $\ln TotalMP_{i'j'}^{>5km}$ where small numbers (0.013, 0.014 and 0.014, respectively) are added to the numerators of the three measures with numbers chosen to ensure that the mean of the previously zero-numerator observations lies at the 10th percentile of the full post-addition distribution.

Table A.1: Regressing worker counts in smartphone geolocation data on LODES jobs

	$NumWorkers_{cm}$			
	(1)	(2)	(3)	(4)
$NumJobsLODES_{cm}$	0.0123***	0.0305***	0.00664**	0.0168***
	(0.00285)	(0.00757)	(0.00273)	(0.00161)
2-digit NAICS Industry		Manufacturing	Information	Professional, scientific and technical services
Industry m FE	✓			
Census-block-group c FE	✓			
Observations	38,840	1,942	1,942	1,942
R-squared	0.355	0.504	0.121	0.722
Within R-squared	0.227			

Notes: Table shows regressions of the number of workers in a particular industry m census-block group c cell (identified from our smartphone geolocation data matched to patenting establishment rooftops) on the total number of workers in the same cell in the LEHD Origin-Destination Employment Statistics. Column (1) includes m and c fixed effects, while columns (2) to (4) focus on specific industries with high shares of patenting firms. Robust standard errors shown in parentheses.

Table A.2: Citations on distance and face-to-face meetings

	arcsinh$PatentCitations_{ij,t}$	
	OLS	OLS
	(1)	(2)
$\ln Distance_{ij}$	-5.12e-05***	-1.93e-05***
	(7.21e-06)	(7.15e-06)
$\ln TotalMP_{ij,t-1}$		4.22e-05***
		(8.40e-06)
$SameIndustry_{ij}=1$	0.00111***	0.00110***
	(1.70e-04)	(1.69e-04)
Establishment i and j FEs	Yes	Yes
Observations	222,401,518	222,401,518
R-squared	0.012	0.012

Notes: Table shows cross-sectional OLS regressions of arcsinh$PatentCitations_{ij,t}$, the inverse hyperbolic sine of patent citations between establishments i and j (allocated from firm-level patents using inventor hometowns) on log as-the-crow flies distance (in km) between establishments i and j. Column (2) further includes the log meeting probabilities $\ln TotalMP_{ij,t-1}$, the probability a worker from i meets a worker from j face-to-face. Citations come from patent applications filed between March 2017 and May 2018 (period t), meetings measures calculated using smartphone geolocation data collected between September 2016 and November 2017 (period $t-1$). All columns include establishment i and j fixed effects and a same-industry dummy. Standard errors clustered at the level of i and j's firms shown in parentheses.

Table A.3: Reduction in citations if most-cited firms left Silicon Valley

Firm Name	Actual Citations	Counterfactual Citations	Percent Reduction
Apple	3681	3649	-0.88
Google	3387	3305	-2.40
Cisco Systems	2176	1873	-13.91
Yahoo!	1792	1765	-1.53
International Business Machines	1788	1678	-6.18
Microsoft	1729	1551	-10.30
Pelican Imaging	1528	1522	-0.40
Oracle	1275	1034	-18.91
Regents of University of California	1002	977	-2.45
Board of Trustees at Stanford	986	689	-30.16
Genentech	968	842	-13.05
Qualcomm Inc	960	902	-6.10
Juniper Networks	887	836	-5.79
Applied Materials	804	732	-8.89
Ebay	770	715	-7.14
Agilent Technologies	750	703	-6.31
Palo Alto Research Center	718	709	-1.22
VMware	702	644	-8.15
Sony Computer Entertainment of America	660	602	-8.73
Nvidia	631	591	-6.33
Average			-7.94

Notes: Table reports absolute and percentage reduction in citations if all establishments i reduced their meeting probabilities with establishments of the top 20 most cited firms $\tilde{j}$ to the 5th percentile of their meeting probabilities with all establishments. I.e. $\text{arcsinh}^{-1}(\text{arcsinh}PatentCitations_{i\tilde{j}t} - \beta\Delta\ln TotalMP_{i\tilde{j},t-1})$ where $\Delta\ln TotalMP_{i\tilde{j},t-1}$ is the change in log meeting probabilities that brings $\ln TotalMP_{i\tilde{j},t-1}$ to the 5th percentile of establishment i's meetings with all other establishments. Citation reduction calculated using the coefficients in Column (8) of Table 3 that include interactions between meetings and the number of workers at i and $\tilde{j}$. Column (1) reports the actual number of citations to patents and patent applications of the listed firm. Column (2) lists the counterfactual number of citations if meetings with establishments of other firms were reduced to the 5th percentile. Column (3) reports the percentage reduction between the actual and counterfactual columns.

B Validating Worker Counts with LODES Data

To provide supporting evidence that our methodology for identifying workers by mapping smartphone geolocation data to building rooftops is effective, we correlate our worker counts with the LEHD Origin-Destination Employment Statistics (LODES) data. These data provide employment counts for 20 industries m across 1,942 Silicon Valley census block groups c. Of course, we should not expect the counts to be similar in magnitude. First, we only have a subset of smartphones (about one fifth) and for the smartphones we do have, we may not record pings at the office (e.g., if the office has poor reception, they turn their cellphone off, or do not open the apps our data originate from during work hours). Second, we only match workers to patenting establishments so do not include the large number of workers at non-patenting establishments.

To explore whether our worker counts correlate with LODES data across locations c and industries m, we plot the number of workers we identify, $NumWorkers_{cm}$, against the LODES employment counts, $NumJobsLODES_{cm}$, both at the cm level:

$$NumWorkers_{cm} = \alpha + \beta NumJobsLODES_{cm} + \gamma_m + \gamma_c + \varepsilon_{cm}. \tag{A.1}$$

The inclusion of both census block group and industry fixed effects ensures that we are not simply measuring the fact that more populous locations or industries have more workers in both our data and LODES.

Appendix Figure A.1 displays a binscatter of this regression and shows a strong positive relationship while column (1) of Appendix Table A.1 presents the regression coefficients. After conditioning on our fixed effects (the within R-squared), we can explain 23 percent of the spatial variation despite only having a subset of smartphones, being unable to separate workers in the same building, and missing workers at non-patenting firms. To address this last issue, columns (2)–(4) of Appendix Table A.1 report similar regressions for three industries where a particularly large share of firms patent: manufacturing, information, and professional, scientific, and technical services. As measured by their R-squared, we can explain 50, 12, and 72 percent, respectively, of the variation in these patent-intensive industries. Taken together, we believe that our smartphone data matched to establishment rooftops do a good job at identifying workers at patenting firms in Silicon Valley.

C Chance Meetings and Labor Flows

In this appendix, we explore one mechanism through which face-to-face meetings may affect knowledge flows between establishments; worker transitions. Face-to-face meetings may provide workers with information about job vacancies and job suitability. For example, a worker from j may by chance meet an old acquaintance from college who works at i. The worker from j knows the acquaintance is an excellent programmer and well suited to a current vacancy that her firm is trying to fill and suggests she applies. Face-to-face meetings facilitating worker transitions in this way is of independent interest, but is also a mechanism that generates knowledge flows between i and j, as measured by citations, as workers bring some of their previous employer's knowledge with them.

Specifically, we test whether worker transitions respond to face-to-face meetings by running a similar specification to our regression of patent citations on meeting probabilities, but replacing citations from establishment i to j with worker flows from building i to j:

$$\text{arcsinh}WorkerTransitions_{ijt} = \beta \ln TotalMP_{ij,t-1} + \Gamma X_{ij} + \delta_i + \delta_j + \varepsilon_{ijt} \tag{C.1}$$

We draw on our smartphone data to measure the worker transitions on the left hand side. If a smartphone is identified as a worker of building i in one month and then in the next month for which data about their device is available it is identified as a worker of building j,we count this as one worker transition from building i to building j. Note that, in contrast to the establishment-level regressions we ran previously, this analysis is at the building level. Since we do not need to merge in firm-level data on citations, there is no benefit to running at the establishment level when we cannot distinguish workers from different firms in the same building.

Many of the endogeneity concerns that arose for the knowledge flows analysis are relevant here. Most obviously, a firm may be interested in hiring a worker and arrange to meet him (e.g. for an interview). Once again, we instrument the log meeting probability with $\ln TotalMP^{>5km}_{i'j',t-1}$ to address such reverse causality, where $\ln TotalMP^{>5km}_{i'j',t-1}$ is the meeting probability calculated from meetings between workers in buildings i' and j' adjacent to i and j, restricting attention to $i'j'$ pairs whose industries (i.e. the industries of the establishments in those buildings) neither cite nor supply each other and excluding meetings occurring less than 5km from either building. Thus, if a worker from i interviews at j, this variation will not appear in our instrument since it uses meetings between workers in adjacent buildings i' and j'. (The need for an instrument precludes us from regressing worker-level transitions on the meetings for that specific worker.)

As with the patent citation analysis, we regress worker transitions on lagged meeting probabilities given the mechanism we have mind. Specifically, we infer worker transitions from the second half of our smartphone geolocation data, March 2017 to November 2017, and calculate meeting probabilities from the full sample collected between September 2016 and November 2017. Thus, we allow worker transitions to be affected by meetings occurring prior to, or at the time of, the transition. However, once again, serial correlation means that we are essentially running cross sectional regressions asking whether workers in buildings who bump into each other more, transition between jobs located in those buildings more.

We present the results of this regression in Appendix Table C.1. The estimates are qualita-

tively similar to the knowledge flows analysis. First, in all specifications face-to-face meetings have positive and significant effects on worker transitions. Second, the OLS coefficient in column (1) is smaller than the vanilla IV in column (2) consistent with measurement error in our meetings measures attenuating the OLS. Comparing columns (2) and (3), there is substantial attenuation of the coefficient when including cubics in three ij distance measures and cubics for each of 11 demographic differences between workers in buildings i and j. Finally, our preferred specification is column (4), which includes both these sets of controls and instruments the log meeting probability using the subset of $i'j'$ meetings that occur more than 5 km from both i and j. The coefficient changes little. Interpreting magnitudes, a 1 percent decrease in the meeting probability $TotalMP_{ij,t-1}$ reduces worker transitions by 1.15 percent.

This preferred IV specification further alleviates endogeneity concerns related to endogenous firm and worker location choices based on the same arguments discussed in the Section 3.2. However, in the case of worker transitions, we are particularly concerned about an upwards bias if workers choose job locations in order to stay near their favored amenities or transportation routes and these preferences also increase meetings between workers at buildings close to these amenities or routes (an issue we partially address in our citation analysis by controlling for these transitions). Conversely, the endogenous location choices of firms may be biasing us downwards if firms choose to locate in places where there are fewer serendipitous meetings in order to avoid their workers being poached.[1] Thus, endogeneity bias is likely to be a more substantial concern in these regressions than it was in our knowledge flows analysis.

[1]Of course, firms may also choose to locate far away from competitors to avoid their knowledge spilling over which may bias our previous citation results downwards as well. However, firms can seek legal remedies to minimize undesired knowledge flows of this sort (e.g. NDAs) while they are not legally allowed to limit worker flows.

Table C.1: Worker transitions and face-to-face meetings

	archsinh$WorkerTransitions_{ij,t}$			
	OLS	IV $\ln TotalMP_{i'j',t-1}$	IV $\ln TotalMP_{i'j',t-1}$	IV $\ln TotalMP^{>5km}_{i'j',t-1}$
	(1)	(2)	(3)	(4)
$\ln TotalMP_{ij,t-1}$	0.000351***	0.000420***	0.000271***	0.000240***
	(1.34e-05)	(2.37e-05)	(3.35e-05)	(3.70e-05)
3 distance cubics			✓	✓
$\Delta_{ij} Demographics$ cubics			✓	✓
Building i and j FEs	✓	✓	✓	✓
Observations	41,790,078	41,073,362	40,771,438	40,735,554
R-squared	0.005	0.001	0.002	0.001
First-stage F		2.582e+06	248,432	132,356

Notes: Table shows regressions of archsinh$WorkerTransitions_{ij,t}$, the inverse hyperbolic sine of worker transitions between buildings i and j measured through smartphone geolocation data, on the log meeting probability $\ln TotalMP_{ij,t-1}$, the probability a worker from i meets a worker from j. Column (1) shows the OLS. Columns (2)–(3) instrument $\ln TotalMP_{ij,t-1}$ with $\ln TotalMP_{i'j',t-1}$, the log meeting probability calculated from the meetings between workers in buildings i' adjacent to i and j' adjacent to j, excluding any $i'j'$ pairs whose industries cite or supply each other. Column (4) instruments $\ln TotalMP_{ij,t-1}$ with $\ln TotalMP^{>5km}_{i'j',t-1}$ which which further restricts the meetings used to calculate the adjacent-building meeting probabilities to those more than 5km away from either building i or j. All columns include building i and j fixed effects and a same-industry dummy. Columns (3)–(4) include cubics in distance, road distance, and travel time between i and j, as well as cubics for each of 11 demographic differences between workers in buildings i and j. Standard errors two-way clustered at the level of building i and building j shown in parentheses.