SKILLED HUMAN CAPITAL AND HIGH-GROWTH ENTREPRENEURSHIP: EVIDENCE FROM INVENTOR INFLOWS

Benjamin Balsmeier
Lee Fleming
Matt Marx
Seungryul Ryan Shin

WORKING PAPER 27605

NBER WORKING PAPER SERIES

SKILLED HUMAN CAPITAL AND HIGH-GROWTH ENTREPRENEURSHIP: EVIDENCE FROM INVENTOR INFLOWS

Benjamin Balsmeier
Lee Fleming
Matt Marx
Seungryul Ryan Shin

Working Paper 27605
http://www.nber.org/papers/w27605

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2020

The authors thank Guan Cheng Li for invaluable research assistance. We also thank participants in the NBER Productivity Seminar for comments and suggestions. We gratefully acknowledge financial support from The Coleman Fung Institute for Engineering Leadership, the National Science Foundation (1360228), and the Ewing Marion Kauffman Foundation. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Skilled Human Capital and High-Growth Entrepreneurship: Evidence from Inventor Inflows
Benjamin Balsmeier, Lee Fleming, Matt Marx, and Seungryul Ryan Shin
NBER Working Paper No. 27605
July 2020
JEL No. J24,J61,L26

ABSTRACT

To what extent does high-growth entrepreneurship depend on skilled human capital? We estimate the impact of the inflow of inventors into a region on the founding of high-growth firms, instrumenting mobility with the county-level share of millions of inventor surnames in the 1940 U.S. Census. Inventor immigration increases county-level high-growth entrepreneurship; estimates range from 29-55 immigrating inventors for each new high-growth firm, depending on the region and model. We also find a smaller but significant negative effect of inventor arrival on entrepreneurship in nearby counties.

Benjamin Balsmeier
Department of Economics and Management
Université du Luxembourg
6, rue Richard Coudenhove-Kalergi
L-1359 Luxembourg
benjamin.balsmeier@uni.lu

Lee Fleming
Coleman Fung Institute for
Engineering Leadership
University of California, Berkeley
lfleming@ieor.berkeley.edu

Matt Marx
Boston University
and NBER
mattmarx@bu.edu

Seungryul Ryan Shin
Seoul National University 1,
Gwanak-ro, Gwanak-gu
Seoul, Republic of Korea
s.ryan.shin@gmail.com

Introduction

Traffic and high rents aside, policymakers worldwide have long sought to replicate Silicon Valley's entrepreneurial success—but typically with middling results (Lerner, 2012). Accordingly, scholars have sought to identify the determinants of regional entrepreneurship. Prior studies have suggested a role for local culture (Hofstede, 2001; Saxenian, 1994; Florida, 2005), technological spillovers from universities (Rosenthal and Strange, 2003; Jaffe et al, 1993), industrial diversity including "anchor" tenants (Duranton and Puga, 2001, Agrawal & Cockburn, 2003), and the munificence of financial capital (Samila & Sorenson, 2011; Grilli & Murtinu, 2014).

Glaeser & Kerr (2009) review these and other theories in an effort to understand the spatial distribution of entrepreneurial activity in U.S. manufacturing. They find that factors such as demographics, culture, and industry mix explain only a small share of the geographical variance. Much more consequential is the availability of relevant talent, accounting for which enables them to explain 60-80% of the variance and conclude that "the broad stability of this finding suggests that people and their human capital are probably the crucial ingredient for most new entrepreneurs" (p. 659). They are of course not the first to suggest that human capital is crucial to the entrepreneurial process, particularly when considering high-potential ventures. By applying scientific principles and research experience to practical problem-solving in the economy, technically-trained workers with backgrounds in science, technology and engineering play a crucial role in innovation and economic growth (Arrow and Capron, 1959; Romer, 1990; Rosenberg and Nelson, 1994; Waldinger, 2016). Jensen and Thursby (2001), for example, argue that scientific inventors need to be fully engaged and motivated for technologies to be successfully commercialized in new firms.[1] Startup activity also appears critically dependent upon inventor knowledge from industry (Klepper 2009) and universities (Zucker et.al 1999; Stuart and Ding 2006).

Evidence of the role for skilled human capital in high-growth entrepreneurship comes primarily from work illustrating correlations between the supply of technical workers levels of patenting, entrepreneurial firm founding, and employment (e.g. Kerr, 2013; Maloney and Caicedo, 2016). However, causality might run the other direction. It could be that the correlations established by Glaeser & Kerr (2009) reflect not only the importance of talent to entrepreneurship but rather the flocking of skilled workers to opportunity. Alternatively, it might be that hiring technical talent is not that important: new ventures might focus on fundraising and outsource product development, as did Slack, Skype, Whatsapp, Alibaba, and BaseCamp.

[1] Not all high-growth firms in the U.S. are high-tech firms, and vice-versa. However, Hathaway (2018) reports that high-tech firms are overrepresented by a factor of four among high-growth firms (21% vs. 5% of all firms) as defined by *Inc*. Magazine's annual list of the 5,000 fastest-growing privately held firms in the U.S.

Either way, the role of local, skilled talent in fueling high-growth entrepreneurship may not be as great as typically thought, and correlations may reflect other, unobserved factors.

To our knowledge, a *causal* link has not been established between the supply of key technical talent—including the sort of scientists and engineers who come up with original inventions—and the founding of high-growth ventures.[2] This is not to say that the entrepreneurial literature has not wrestled with the role of talent, but it has done so primarily in the realm of executive leadership. For example, Kaplan et al (2009) suggest that most successful startups remain true to their original business plan, upgrading executive as they move towards IPO. Ewens and Marx (2017) add causal evidence that startups depend on human capital in the executive suite, instrumenting investors' ability to replace the founder with a seasoned CEO via state-level shifts in policies regarding employee non-compete agreements. Again, these articles focus on the "upper echelons" of organizations, ignoring the vast majority of (potential) employees at new ventures.

Our aim is to test whether the local supply of inventors is a critical determinant of high-growth entrepreneurship. We establish that the arrival of inventors in a region drives the founding of startups that grow and become successful, addressing reverse-causality concerns by instrumenting inventor inflows with the share of inventors' surnames in that region based on the nationwide distribution of surnames from the 1940 U.S. Census. Our shift-share instrument represents an advance over prior efforts in two ways. First, because the "shares" stem from more than three million unique surnames across a large number of counties, it is less vulnerable to critiques that apply to such instruments with low variation or a few highly-determinative shares (see Borusayak et al, 2019 for a fuller discussion of the issue). Second, focusing on the U.S. lessens concerns regarding endogenous origin-destination combinations (e.g., Indian engineers migrating to Silicon Valley) and addreses the issue of potential endogenous choice of regions and selection of incoming inventors at the nation level (Moser, Voena, Waldinger 2014; Parey et al., 2017).

We first establish that the local share of surnames matching a given inventor strongly predicts immigration into a county. We then aggregate moves to create a county-wide average likelihood of arrival for inventors with a given surname. Application of the instrument indicates that the arrival of 55 inventors results in one new high-growth startup on average. In California and Massachusetts, only 29 inventors are required to achieve a similar effect. A back-of-the-envelope calculation suggests that incoming inventors

[2] Related to this paper, several studies have addressed the role of local inventors in regional productivity. For example, Agrawal, et al. (2011) show that inventor emigration decreases local knowledge flow in the source region but also drives knowledge back into the departed region. A growing and influential literature on foreign immigration suggests positive impacts on the U.S. of an influx of inventors from outside its borders, including greater patenting and innovation (Bernstein et. al., 2018; Hunt and Gauthier-Loiselle 2010; Moretti et al. 2018; Burchardi et. al. 2020; Kerr and Lincoln 2010), wages (Peri, Shih, and Sparber 2015) and TFP (Capelli, Czarnitzki, Doherr, Montobbio 2019). Our study differs from these in that we study *internal* migration within the U.S.

may explain 24.9% of high-growth startups. Results remain robust to a variety of different models, instruments, and measures, and even to the inclusion of inventor immigration to surrounding counties, which has a negative effect on entrepreneurship in the focal county.

Data description

Historic Census data

The central challenge in determining the causal effect of inventor mobility upon regional outcomes is to disentangle the choice to move from the reasons to move. Ideally, we would observe inventors moving for reasons completely unrelated to career opportunities. To approximate this as best as possible, we seek to instrument an inventor's move to a county in which they had not previously patented. We build a variety of instruments based on 1940 Census Data (http://sites.mnhs.org/library/content/1940-census), first establishing the plausibility of the instrument at the individual inventor level and then aggregating to develop a county-level instrument.

We begin with the 1940 U.S. Census records for 131,940,709 citizens in 38,382,088 households. These data include 3,363,932 different surnames, of which 27% appear only once. (The median is 3, the mean is 39, and the maximum is 1,359,079 for Smith.) Figures 1 and 2 illustrate the sparse geographical distributions of "Marx" and "Fleming". After some name cleaning and standardizing procedures, described in detail in Appendix A1, there were 42,268 Flemings, 6,232 Marxes, 153 Shins, and 9 Balsmeiers in the 1940 census data. All analyses are robust to excluding prolific surnames as indicated by high (local) frequency or wealth. The 1940 U.S. Census records cover 3,086 counties.[3]

[3] The 1940 U.S. Census records consist of 3097 counties and other districts based on the county system in 1940. In order to help matching with the location information of inventors, we translate 19 counties or districts, which are old and no longer in use, to a corresponding county in the current county system based on the location, resulting in 3086 unique counties.

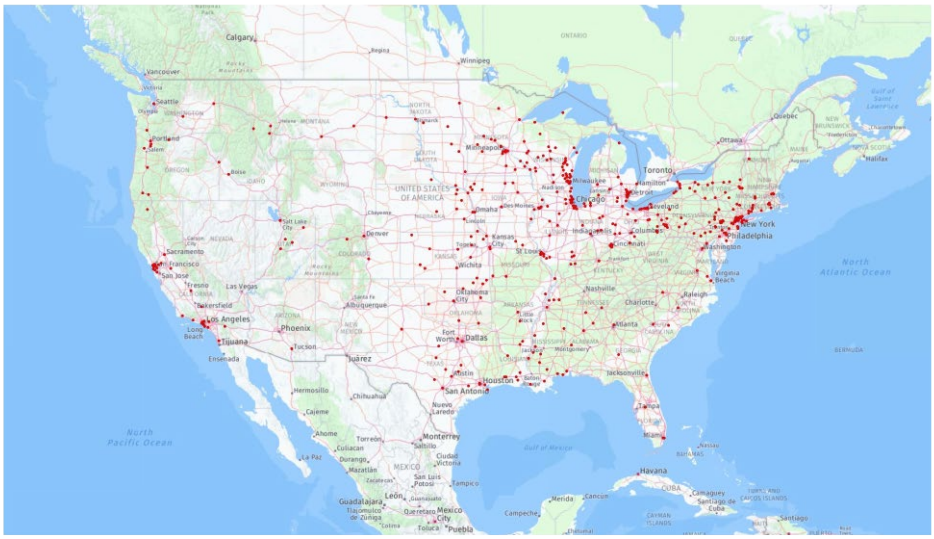

Figure 1: Spatial distribution of the surname "Marx" in 1940 (each red dot corresponds to 50 individuals, 4,762 in total).

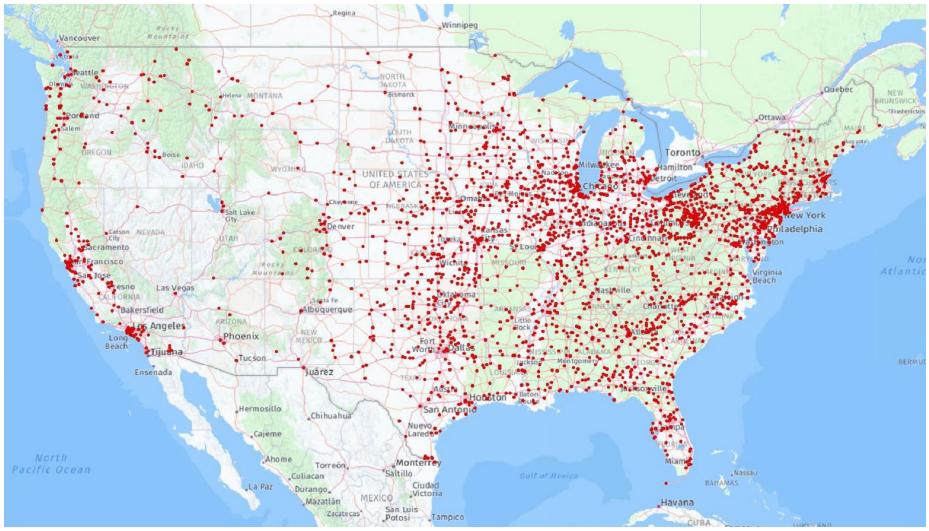

Figure 2: Spatial distribution of the surname "Fleming" in 1940 (each red dot corresponds to 50 individuals, 42,268 in total).

Inventor data

We begin with raw data from the United States Patent and Trademark Office (USPTO) from 1976-2018. Although the USPTO lists inventors for every patent, it does not provide unique identifiers for them. For example, even the relatively rare name of Matthew Marx is listed as inventing many patents, including 5,995,928, "Method and apparatus for continuous spelling speech recognition with early identification,

6,173,266, "System and method for developing interactive speech applications," and 7,271,262, "Pyrrolopyrimidine derivatives." In this simple example, it would seem reasonable based on the titles alone that the same inventor authored the first two but not the last patent, and that is indeed the case. Inventor names can be disambiguated with a variety of algorithms, here we use Balsmeier et al. (2017). After applying the name cleaning and standardizing procedures and the matching algorithm, described in detail in Appendix A1, we match 91.1% of inventors' surname to a surname from the 1940 Census. Note that the name cleaning exercise has no significant effect on the size of the estimated coefficients but decreases matching errors and improves precision of the instrument.

We used the inventor ID and location to identify inventor moves across US counties. We drop all inventors with a single patent. Then, using patent application year as a timestamp, we count an inward move in the first year we first observe an inventor in a county. As noted by Cheyre, Klepper, and Veloso (2015), patent application dates do not necessarily correspond with dates of employment and in particular may lag actual moves. Hence, the inventor may have moved into a county earlier than we detect, leading to a fuzzy lower bound of the actual lag between our variable of interest and the actual inward moves. In 96% of cases, we observe an incoming inventor patenting elsewhere up to 5 years earlier (mean = 2.6). Results are robust to excluding inventor moves with longer gaps between two patenting events, or temporary stops at a third county. If an inventor appears on two or more patents within a given year, we follow Moretti and Wilson (2017) and take the most frequent location.

Entrepreneurship data

To measure high-growth entrepreneurship, we rely primarily upon business registration data from Guzman and Stern (2019). Instead of examining selected samples such as Crunchbase or NETS, they collect annual state registers of *all* businesses founded. They link newly-founded businesses from these registers to SDC Platinum and report the number that experience a liquidity event (i.e., and Initial Public Offering or acquisition) within six years of founding. This forms our primary dependent variable. As a robustness check, we also use their measure of the number of local startups that are *expected* to experience high growth. Finally, as an alternative measure of high-growth potential we count the number of venture-financed startups in the region using VenturExpert.

Table 1 shows descriptive statistics of our key variables at the county level and Figure 3 provides a scatterplot of the relationship between incoming inventors and high-growth startups in the raw data.

Table 1: descriptive statistics at county level

Variable	N	mean	median	stdev	min	max
Incoming Inventors	82,259	2.47	0.00	14.71	0.00	939.00
Instrument	82,259	2.27	0.80	10.22	0.00	443.14
Number of new high-growth startups	82,259	0.18	0.00	1.58	0.00	90.81

Notes: This table reports summary statistics of the key variables used in our regression analyses at the county level, covering 3130 counties 1988-2014. The number of high growth startups is not an integer value in some cases because they are measured by Guzman and Stern (2020) at the zip code level and split by an algorithm in case zip code areas overlap county borders.

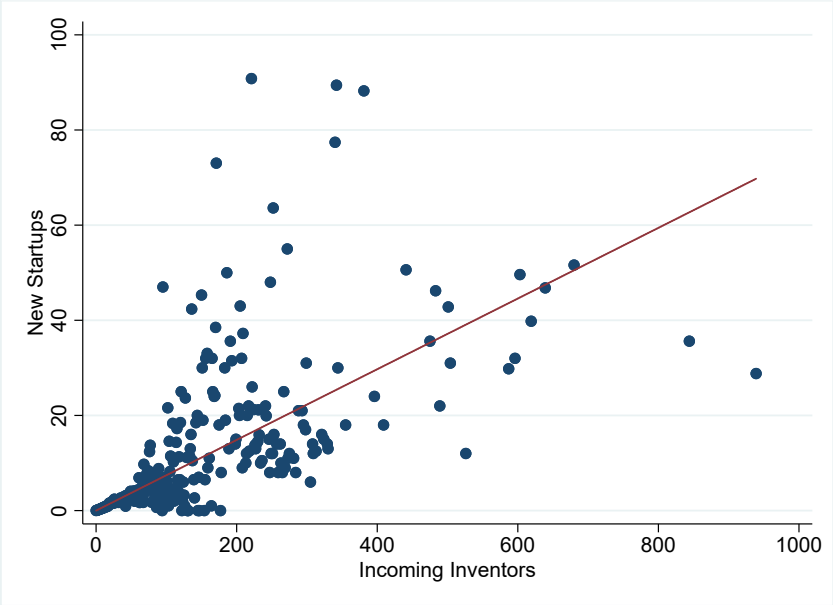

Figure 3 – Graphical representation of raw data

Methodology and instrument construction

Our goal is to estimate the causal impact of inventor inflows on high-growth entrepreneurship. We can write this in an equation that we can estimate with OLS:

$$Y_{d,t+1} = \alpha_0 + \beta \cdot Inv_{d,t} + \delta_t + \gamma_d + \varepsilon_{dt} \quad (1)$$

where $Y_{d,t+1}$ stands for a dependent variable observed for county d at time $t+1$. $Inv_{d,t}$ is the number of inventors that moved to county d in t. δ_t denotes a full set of year fixed effects to control for varying macroeconomic conditions. γ_d controls for time-invariant unobserved county characteristics that may confound our identification of β. ε_{dt} is the error term.

The key econometric challenge with equation (1) is that unobserved factors influence both the rate of incoming inventors and local economic conditions. Counties with a high innovation rate are attractive to

inventors. Although county fixed effects will effectively control for any persistent differences in innovation levels across counties, it is difficult to control for temporary local trends that might attract inventors. To address this threat to identification, we construct a shift-share instrument for inventor inflows that builds on the work of Bartik (1991) and its application to the case of international immigration to the US (Card, 2001). Prior immigration studies noted that immigrants from a certain country of origin tend to locate near previous immigrants from the same origin country (Bartel, 1989, and Lalonde and Topel, 1991). Card (2001) and others (see Jaeger et al. 2018 for an overview) have exploited this to predict immigrant inflows into certain regions by interacting past shares of immigrants from an origin country in a given region with the contemporaneous inflow of migrants from the same country at the national level.

We leverage this idea further to create an instrument for the contemporaneous inflow of US inventors to a certain county based on the spatial distribution of US surnames across counties in 1940. Analogous to the prior immigration literature, we utilize the observation that people with a certain family name can be found more frequently at places where there were other people with same name in the past (see Darlu, Brunet, and Barbero, 2011, for the example of Savoy, France and Clark & Cummins, 2014, for England). We illustrate below that these patterns hold for individual US inventors.

Specifically, we define our instrument as:

$$\widetilde{Inv_{dt}} = \sum_n \frac{P_{dn}^{1940}}{P_n^{1940}} \cdot Inv_{nt} \qquad (2)$$

where P_{dn}^{1940} is the population of people in county *d* with surname *n* in 1940, P_n^{1940} is the number of people with surname *n* in the entire US in 1940 and Inv_{nt} is the number of inventors with surname *n* who move from any county in the US to any other county in the US in year *t*. The expected inflow of inventors $\widetilde{Inv_{dt}}$ in county *d* at time *t* is thus the weighted sum of inventors that move across the US with surname *n* (the "shift") with the historical distribution of the same family names (the "shares") serving as weights. The intuitive appeal behind this instrument (as in prior immigration studies) is that it generates variation at the local level by exploiting variation at the national level, which is arguably not influenced by local conditions. (That is, the total number of inventors with the name Fleming who move from within entire US is unlikely to be driven by the local economic conditions of one out of the more than 3,000 U.S. counties.)

One advantage of this instrument over prior shift-share instruments generally, and settlement instruments in particular, is the extensive amount of variation in the distribution of names (i.e., the "shares") that stem from more than 3 million unique surnames in 1940 across varying destination and origin areas. (By contrast, immigration studies typically analyze 192 different countries, often with particularly influential origin-destination relationships.) Our estimation should therefore be less vulnerable to problems that arise from

low variation in shares or overly strong influences of a single or few shares (see critiques in the recent literature, Borusyak et al., 2019, Goldsmith-Pinkham et al., 2020, or Adao et al. 2019).

A second advantage of our U.S.-focused shift-share instrument is that a given surname is typically not bound to a specific county of origin (as is more common with country-level analysis, see Moser, Voena, Waldinger 2014; Parey et al., 2017). Thus the spatial distribution of origin of mobile inventors with a surname varies substantially over time (and is the only variation we exploit in our IV). This makes an endogenous origin-destination combination (such as Indian engineers coming into Silicon Valley for long periods of time) highly unlikely to drive our results. Put differently, that mobile inventors with certain names come from various origin counties means that it is less likely that our "shift" is correlated with unobserved endogenous characteristics of origin areas. Our instrument thus minimizes serial correlation between specific origin and destination regions, as criticized in studies of international migration.

Figures 4-6 illustrate the variation over time and space with the example of all inventors that moved across the US between 1976 to 2014 and have the last name Fleming (75 moves in total). Figure 2 shows how the number of mobile inventors with surname Fleming varies over time yet does not exhibit a trend. The heatmaps in Figure 3 show to which counties the Flemings moved to in the 1980s, 1990s, and 2000s, and the heatmaps in Figure 4 show the origin counties the Flemings moved away from in the 1980s, 1990s, and 2000s, illustrating significant variation in origin and destination counties over time.

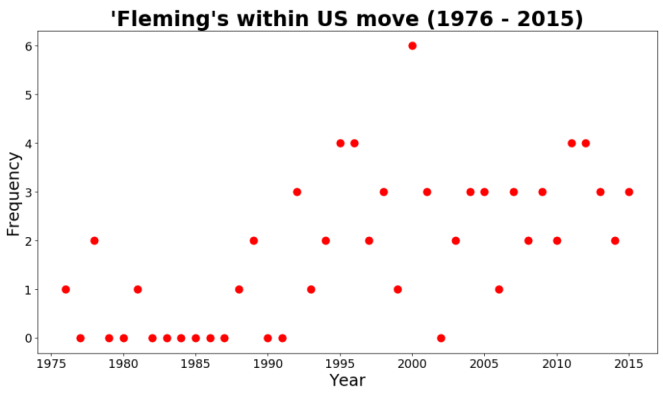

Figure 4: Frequency of moving inventors within the US named Fleming over time

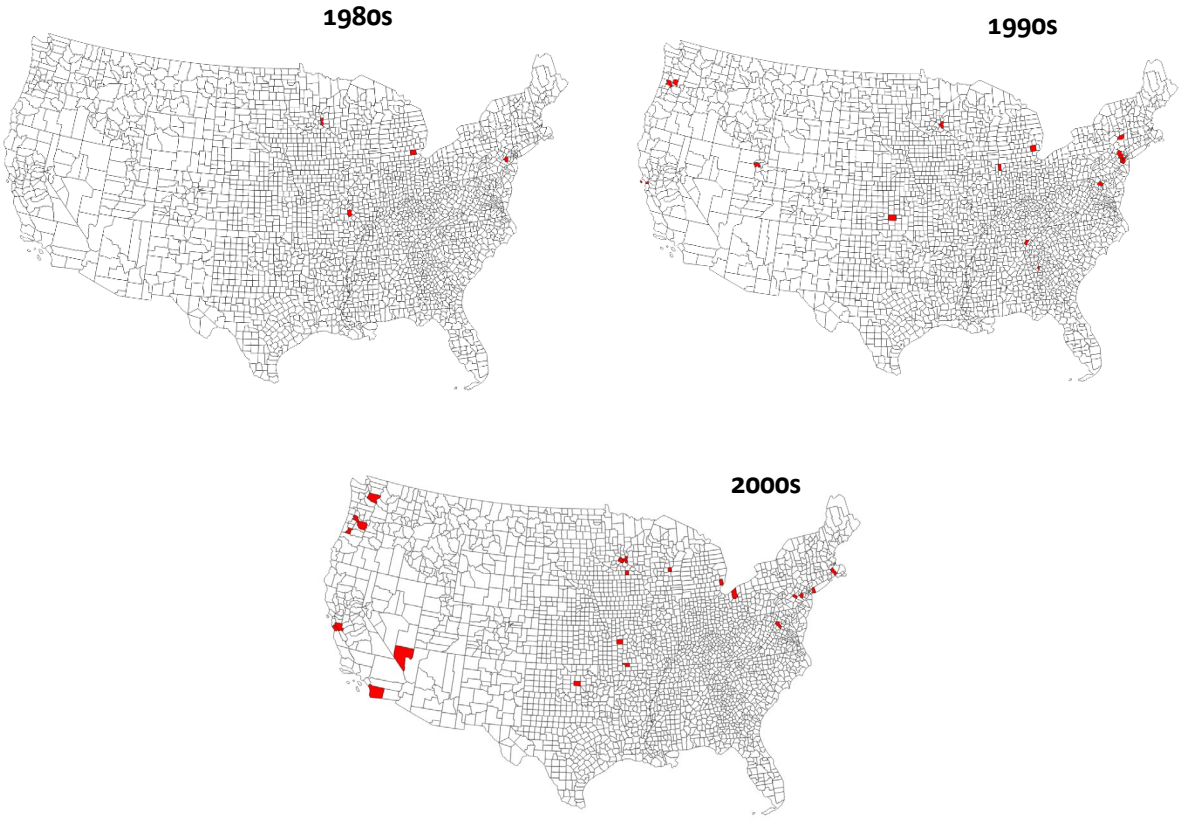

Figure 5: Destination counties of moving inventors within the US named Fleming

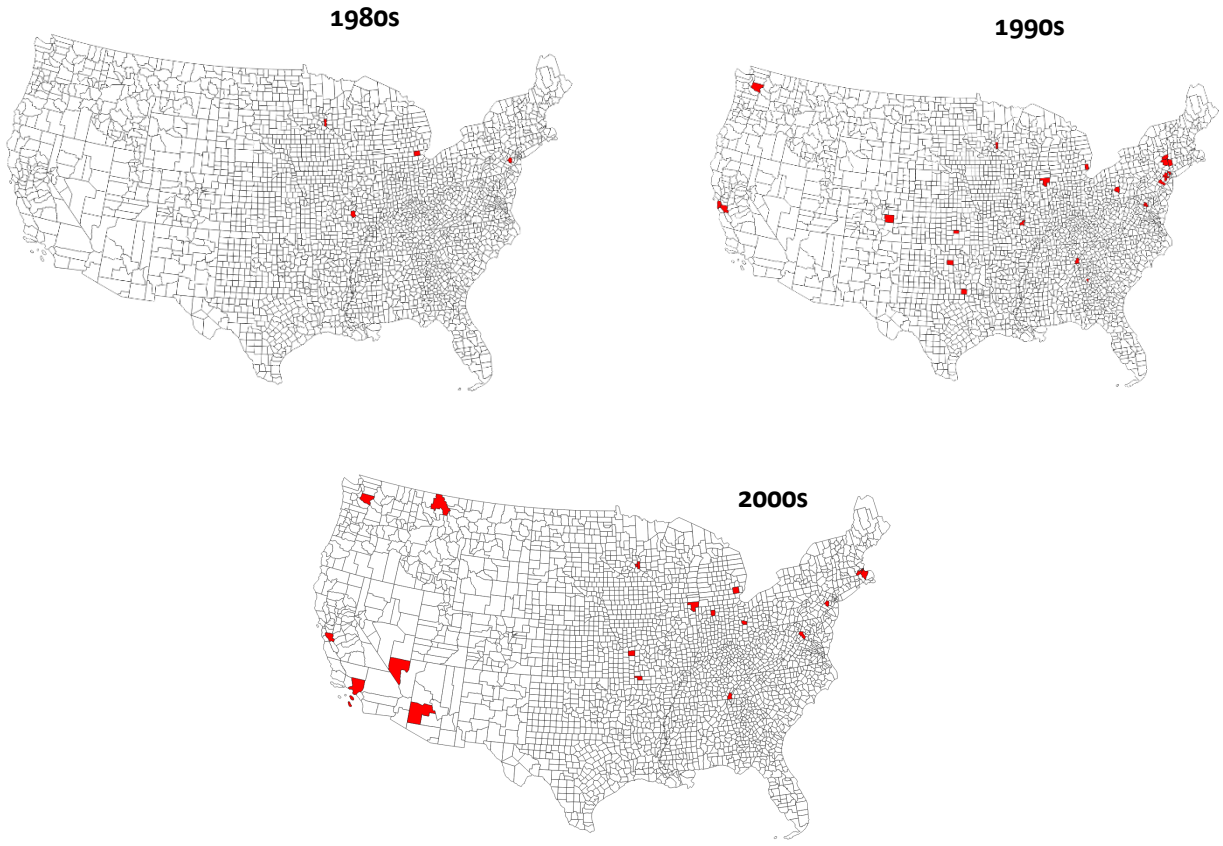

Figure 6 - Origin counties of moving inventors within the US named Fleming

A remaining concern could be that at least some national movements of inventors are still driven by local economic conditions, and that these might be correlated with past shocks. It could be, for instance, that inventors and families with the name Fleming were always interested in mechanical engineering and thus would have settled in areas where mechanical engineering was in high demand as of 1940. If the same area experiences a high demand in mechanical engineering today, then inventors with the name Fleming might more likely move to that region for endogenous reasons. To reduce endogeneity concerns in this respect, we augment our instrument by leaving out county *d*'s own inflows from the national flow of inventors with same surname. (Buchardi et al. 2020, Wozniak & Murray, 2012 and Hunt, 2017 use similar strategies in the immigration literature). Our preferred instrument is thus:

$$\widetilde{Inv_{dt,leave-out}} = \sum_n \frac{P_{dn}^{1940}}{P_n^{1940}} \cdot Inv_{nt,leave-out(n,d)} \quad (3)$$

where $Inv_{nt,leave-out(n,d)}$ is the total number of inventors with name *n* who move to counties outside of *d*.

The leave-out strategy ensures that the potentially-endogenous choice of Flemings to move to county d does not drive changes in our instrument.

It should further be noted that both stages of our IV regression include county fixed effects. Identification thus derives from weighted time-varying changes in the number of moving inventors with a given surname at the national level, excluding those moving to county d, combined with representation of the same surname in county d in 1940. Since it is highly unlikely that the Flemings who move in year t are equally distributed across the country as the Flemings that move in year t+1 (as illustrated above), our instrument is also unlikely to suffer from any unobserved persistent endogenous relationship between any pair of counties (e.g. trade relationships or the oft-cited, persistent link between Indian software engineers and immigration from India to Silicon Valley). The considerable variation in the distribution of surnames over time also addresses the "persistence problem" with shift-share instruments in the immigration literature (Jaeger et al., 2018). Arguably, $Inv_{dt,leave-out}$ is truly exogenous and can be used to estimate the causal impact of inventor inflows on using equation (1), instrumenting $Inv_{d,t}$ with $Inv_{dt,leave-out}$ as in (3).

First stage plausibility check - individual inventor level regressions

Before applying our instrument at the county level, we establish its plausibility by investigating the linkage between the historical surname distribution and geographical mobility of individual inventors. Given widely accepted transmission rules (Piazza et al., 1987; Rossi, 2013) and diversity of surnames (3,363,932 unique surnames appear in 1940 US census data), demographics on surnames are adopted in a variety of studies, such as research on migration of people, social networks and mobility. Piazza et al. (1987) tracks migration rates using surname distribution in Italy. Degioanni and Darlu (2001) infer geographical origin of migrants in a given area using surnames. Darlu, Brunet, and Barbero (2011) show that surname distribution can be used to estimate mobility using the example of Savoy, France. Studies also use surnames to investigate social mobility, e.g., whether social status changes over centuries (Clark & Cummins, 2014) and whether wealth moves over generations (Clark & Cummins, 2015). In a recent study, Grilli and Allesina (2017) perform a surname analysis on academic professors to compare academic systems in the US, France, and Italy. Our first stage plausibility check contributes to this literature by exploiting the complete US 1940 Census surname and location information, linking to the complete set of US patenting inventors.

Our IV approach rests on the assumption that historic surname shares can discriminate between

destination counties of moving inventors with a given last name, conditional on moving.[4] We empirically test this assumption by estimating a dyadic model that reflects the complete choice set a moving inventor is confronted with. To this end, we construct a dataset at the inventor-origin-destination county level that contains each potential destination county combined with the actual county a given inventor is emigrating from. We mark the county the inventor actually moved to with a dummy and for the actual and each potential destination county, include the share of people in the 1940 Census with the same surname. Armed with this dyadic dataset covering 258,657 moves from 1988-2014, we estimate the following model with OLS:

$$Pr(d.cty\#o.cty_{i,o,d,t} = 1|Move\ out_{o,t}) = \alpha_0 + \beta \cdot (\frac{P_{dn}^{1940}}{P_n^{1940}}) + \delta_t + \gamma_d + \gamma_o + \varepsilon_{i,d,o,t} \quad (4)$$

where $Pr(d.cty\#o.cty_{n,o,d,t} = 1|Move\ out_{o,t})$ is a dummy indicating the destination county ($d.cty$) a given inventor *i* with name *n* moved to from origin county $o.cty$ in year *t*. P_{dn}^{1940} is the population in county *d* with surname *n* in 1940; P_n^{1940} is the population with surname *n* in the entire U.S. in 1940; δ_t denotes a full set of year fixed effects to control for varying macroeconomic conditions; γ_d controls for time-invariant unobserved destination county characteristics; and γ_o controls for time-invariant unobserved origin county characteristics that may confound our identification of β, and $\varepsilon_{i,d,o,t}$ is the error term. We estimate four versions of equation (4): (a) only with year fixed effects; (b) year and destination-county fixed effects; (c) year and origin-county fixed effects; (d) year and destination-origin county combination fixed effects. Variant (d) absorbs time-invariant county-pair relationship characteristics including, for instance, the geographic distance between two counties. Table 2 presents the results.

[4] Adding to the plausibility of our instrument, we also find that the historical share of the same surname in a given location is negatively associated with the inventor's emigration from the location. This supports the argument that inventors are not only more likely to move to regions with a higher historic share of the same surname but also more likely to stay in a region in which more of their families and relatives have resided. Several additional analyses verify the robustness of the results and suggest conditions in which the historical surname effect is moderated. The surname effect is amplified as the average value of houses owned by individuals with the same surname in the county increases, as the foreign-born ratio of individuals with the same surname in the county decreases, or when the inventor resides in a state that enforces non-compete agreements. We find no evidence that the surname effect is susceptible to invention-related inventor characteristics, such as invention productivity, quality, or years of experience as an inventor.

Table 2 – Destination county choice

	origin-destination county move			
	a	b	c	d
Destination county Historic surname fraction	0.044*** (0.006)	0.021*** (0.002)	0.044*** (0.006)	0.013*** (0.001)
N	524,583,139	524,583,139	524,583,139	523,553,217
Year FEs	Yes	Yes	Yes	Yes
Destination county FEs	No	Yes	No	No
Origin county FEs	No	No	Yes	No
Origin-destination county FEs	No	No	No	Yes
R^2	0.000	0.008	0.000	0.061

Notes: This table presents OLS regressions of a dummy indicating a origin-destination county move of an inventor within the period 1980 to 2015 on destination counties' historic surname shares in 1940. Unit of observation is the origin-destination county dyad. Standard errors clustered at the destination county appear in parentheses. ***, ** and * indicate a significance level of 1%, 5%, and 10%, respectively.

Although we cannot interpret our LPM specification as a probability model, all specifications consistently show that an increase in the historic surname share in a potential destination county leads to a significantly higher probability of observing a given inventor moving to that specific destination county as compared to all other potential destination choices. The results in Table 2 support the plausibility of our instrument. (Since the dependent variable vector is sparse a low R^2 is to be expected.) The increase in explained variation when destination and destination-origin county fixed effects are included reinforces that unobserved time invariant factors explain mobility decisions.

The impact of incoming inventors on regional entrepreneurship

We now move to our IV regression analysis to the county level, where the dependent variable is a (logged) measure of the number of high-growth startups founded in county *d* during year *t*. "High-growth" is determined retrospectively as the number of firms founded in year *t*+1 that achieved an IPO or successful acquisition within 6 years after founding. Results are in Table 3. Model (a) estimates equation (1) via OLS and without the instrument. Model (b) re-estimates model (a) using the instrument from (3). Models (c) and (d) and state-year and then county fixed effects. Finally, in model (e) we re-estimate (d) but exclude California and Massachusetts.

Table 3 – Impact of incoming inventors on local high growth startup foundation

	High-growth startups founded				
	a	b	c	d	e
	OLS	IV	IV	IV	IV (w/o CA, MA)
Incoming Inventors$_{t-1}$	0.188***	0.377***	0.380***	0.248***	0.232***
	(0.015)	(0.027)	(0.027)	(0.043)	(0.043)
N	82,259	82,259	82,259	82,259	80,330
First Stage F		400.2	387.8	115.5	107.2
Year FE	Yes	Yes	No	No	No
State FE	Yes	Yes	No	No	No
State-Year FE	No	No	Yes	Yes	Yes
County FE	No	No	No	Yes	Yes
R^2	0.341				

Notes: This table presents OLS regression of log(number of high growth startup foundation + 1), where high growth startups are defined following Guzman and Stern (2020) as newly registered companies that complete either an IPO or successful acquisition within 6 years. Incoming inventors as well as the instrument are log-transformed. Specifications (b)-(e) show results of our IV regression as described above, where incoming inventors are instrumented with $Inv_{dt,leave-out}$ in the first stage. First stage F is the Kleibergen-Paap Wald F statistic of the first stage regression. Model e excludes all counties of California and Massachusetts. Standard errors clustered at the county level appear in parentheses. ***, ** and * indicate a significance level of 1%, 5%, and 10%, respectively.

We begin in model (a) of Table 3 by correlating of the number of incoming inventors in a county with the count of high-growth startups founded the following year. Consistent with Glaeser & Kerr (2009), we observe a strong relationship between the supply of relevant talent and entrepreneurial activity, as shown by the positive and statistically significant estimated coefficient on Incoming Inventors. The remaining models (b-e) employ our IV approach, all showing a significant positive impact of incoming inventors in a given county on the local rate of high growth startup formations. The strength of the instrument and the coefficient size drops after the inclusion of county fixed effects, however, the first stage *F* value is still far above conventional levels, suggesting that our IV regression does not suffer from weak instrument biases.

Under the assumption that a log specification can be interpreted as an elasticity, model (d) suggests that a 10% increase in the rate of incoming inventors increases the rate of high growth startup foundations by 2.48% against the mean. Translating the relative increases into absolute numbers suggests that 10 more inventors lead to 0.181 more startups. Put differently, a county can expect one additional high growth startup per 55 incoming inventors. It should be noted though that the distribution of incoming inventors, as well as the number of high-growth startups founded in a county, is highly skewed. Hence, mean values are not representative. The log transformation addresses skewness, but caution is warranted when calculating elasticities in absolute terms. Taking, for instance, the higher averages of incoming inventors (27.79) and growth events (2.96) of California and Massachusetts counties as the benchmark, we approach an elasticity of 1 additional high growth startup per 29 incoming inventors. The figure is even higher if we restrict the sample to California and Massachusetts (see Appendix Table A1), as we find not only a higher mean value

but also a higher estimated coefficient (β=0.325).

Taking our full sample estimates and summary statistics as a baseline, a back-of-the-envelope calculation suggests that incoming inventors may explain 24.9% of all high growth startups. We took all observed high growth startups in the sample (14,783) and the total number of incoming inventor events (202,911). Given the 55:1 elasticity we would expect a total of 202,911/55 = 3,689 new startups from all incoming inventors, which represents 24.9% of high-growth startups in the sample.

Robustness checks

Alternative instrument constructions

Although the validity of shift-share instruments does not require exogeneity of the shares, and although concerns should be lessened by the inclusion of county fixed effects, we nonetheless estimate robustness checks that should further alleviate concerns of potentially-endogenous share characteristics. To this end, we re-estimate model (d) of Table 3, replacing the instrument with alternative calculations of the historic name shares (while still applying our leave out strategy).

For the first alternative instrument, we consider only people in US 1940 Census that lived in a given county before 1935. We thus effectively enlarge the gap between the shares and the actual moves of inventors and reduce potential correlation between historic and current inventor migration shocks. Second, we exclude the 50 surnames that appear most frequently in the historic data, which should reduce concerns that correlated shares of two counties may lead to an over-rejection problem (as shown by Adao et al., 2019). In our third construction, we exclude wealthy families of each county as inventors may benefit even generations later from their ancestors' wealth. Using the historic house value in the 1940 Census, we excluded families holding more than 1% of the total house value of a given county.

Our fourth construction departs from the shift-share approach, instead calculating the inventor's separation from his or her surname's historic geographic centroid. We use the inverse squared geographic distance between each county centroid and the geographic centroid for an inventor's surname as weights when constructing the instrument. The distance between a county's centroid and a surname's historic geographic centroid has the advantage of a very low correlation with any future county or inventor specific characteristics.[5] Table 4 shows the results for these alternative instruments.

[5] A limitation of this fourth instrument construction is that some surname are clustered in multiple geographic

Table 4 – Alternative instruments

	High-growth startups founded			
	a	b	c	d
	IV	IV	IV	IV
Incoming Inventors$_{t-1}$	0.255***	0.248***	0.253***	0.261**
	(0.047)	(0.043)	(0.044)	(0.113)
N	82,259	82,259	82,259	82,259
First Stage F	101.1	115.5	116.7	25.56
State-Year FE	Yes	Yes	Yes	Yes
County FE	Yes	Yes	Yes	Yes

Notes: This table presents OLS regression of log (number of high-growth startups founded + 1), where high-growth startups are defined following Guzman and Stern (2020) as newly registered companies that complete either an IPO or successful acquisition within 6 years. Incoming inventors as well as the instrument are log-transformed. Model (a) restricts the instrument to those who settled in the county of the 1940 Census by 1935; (b) excludes excludes the 50 most frequent surnames; (c) excludes the wealthiest 1% of surnames per 1940 Census house value; (d) replaces the shift-share approach with the inverse squared geographic distance between the county and the centroid for the inventor's surname. First stage F is the Kleibergen-Paap Wald F statistic of the first stage regression. Standard errors clustered at the county level appear in parentheses. ***, ** and * indicate a significance level of 1%, 5%, and 10%, respectively.

The coefficient sizes remain robust across different specifications, although the strength of the instrument declines in models (a) and (d) compared to our original instrument. Especially with respect to our centroid-distance instrument, this is not surprising. That the instrument strength and coefficient size does not decline greatly when excluding particularly influential families supports the assumption that either 1) there is no direct link between the historic name shares and the second stage regression, or 2) the county fixed effects effectively absorb such potentially worrying relationships.

Alternative dependent variables

We further verify the robustness of our results by testing two alternative dependent variables. First, we replicate the results of Table 4 by replacing the count of startups that achieved a liquidity even with Guzman & Stern's (2020) projection of the number of startups in a county that are *expected* to do so (whether or not they actually did). Their Regional Entrepreneurship Cohort Potential Index is an aggregation of several startup characteristics including high tech industry classification (Biotech, IT, E-Commerce, Medical and

regions. Thus, even if there is one largest centroid, we will calculate distance from it even if a somewhat smaller but much-closer aggregation exists. The share-of-surnames instrument does not suffer from this limitation.

Semiconductors), trademark registration, incorporation in Delaware, having applied for a patent, and eponymy (Belenzon et al., 2017). The RECPI has the advantage of more broadly reflecting a county's startup quality. It will thus not only capture startups that actually achieve high growth within 6 years but also those that did not achieve a liquidity event, or that took longer to do so.[6] Table 5 shows similar findings when using this dependent variable.

Table 5 – Impact of incoming inventors on regional entrepreneurship potential

	Regional Entrepreneurship Cohort Potential Index				
	a	b	c	d	e
	OLS	IV	IV	IV	IV (w/o CA, MA)
Incoming Inventors$_{t-1}$	0.203***	0.410***	0.412***	0.475***	0.435***
	(0.013)	(0.027)	(0.027)	(0.039)	(0.042)
N	82,259	82,259	82,259	82,259	80,330
First stage F		400.2	387.8	115.5	107.2
Year FE	Yes	Yes	No	No	No
State FE	Yes	Yes	No	No	No
State-Year FE	No	No	Yes	Yes	Yes
County FE	No	No	No	Yes	Yes
R^2	0.517				

Notes: This table presents OLS regression of log(RECPI + 1), where RECPI is the Regional Entrepreneurship Cohort Potential Index as defined and calculated by Guzman and Stern (2020). Incoming inventors as well as the instrument are log-transformed. Specifications b to e represent results of our IV regression as described above, where Incoming Inventors are instrumented with $Inv_{dt,leave-out}$ in the first stage. First stage F is the Kleibergen-Paap Wald F statistic of the first stage regression. Model e excludes all counties of California and Massachusetts. Standard errors clustered at the county level appear in parentheses. ***, ** and * indicate a significance level of 1%, 5%, and 10%, respectively.

One concern with using RECPI in this context is that one of the variables used in its construction is whether the startup had a patent upon being incorporated. This might be problematic, as we use patent data to observe mobility. Thus as a second robustness check we re-estimate our baseline models using the count of startups in county *d* during year *t* that eventually received venture capital financing. This variable is calculated using VenturExpert, a dataset of venture-backed companies, covering 1980 -2010. The results of Table 6 are consistent with Tables 3-5.

[6] Guzman and Stern (2020) and Andrews et al. (2019) show that the RECPI accurately reflects a region's startup potential and that it is not dependent on modelling choice (results below are based on their so called 'academic' approach which takes twelve startup characteristics into account and are robust to using their less sophisticated but easier to calculate 'policy' approach). The correlation of the log transformed variables is r=0.839.

Table 6 – Impact of incoming inventors on local venture backed startups

	Venture-backed companies founded				
	a	b	c	d	e
	OLS	IV	IV	IV	IV (No CA, MA)
Incoming Inventors$_{t-1}$	0.299***	0.592***	0.607***	0.154***	0.123***
	(0.020)	(0.036)	(0.037)	(0.034)	(0.031)
N	97,247	97,247	97,247	97,247	95,015
First stage F		388.9	366.4	240.9	221.1
Year FE	Yes	Yes	No	No	No
State FE	Yes	Yes	No	No	No
State-Year FE	No	No	Yes	Yes	Yes
County FE	No	No	No	Yes	Yes
R^2	0.341				

Notes: This table presents OLS regression of log (number of venture-backed startups + 1) from VentureExpert. Incoming inventors as well as the instrument are log-transformed. Specifications (b)-(e) show results of our IV regression as described above, where incoming inventors are instrumented with $Inv_{dt,leave-out}$ in the first stage. First stage F is the Kleibergen-Paap Wald F statistic of the first stage regression. Model e excludes all counties of California and Massachusetts. Standard errors clustered at the county level appear in parentheses. ***, ** and * indicate a significance level of 1%, 5%, and 10%, respectively.

Spillover vs. displacement effects

Locally increased entrepreneurial activity spurred by incoming inventors might imply a positive impact of inventor mobility on aggregated entrepreneurship for the country as a whole. Our regressions, however, do not account for potential spillovers to neighboring counties, and internal migration is a zero-sum game. Regional spillovers could be positive if incoming inventors' entrepreneurial activities are not bound to county borders—i.e., at least some inventors are involved with new ventures in counties other than where they live. Alternatively, incoming inventors may displace entrepreneurial activity in nearby counties if startups in proximate counties compete for resources. Given that venture capitalists prefer to invest locally (Stuart & Sorenson, 2003; Bernstein, Giroud, & Townsend, 2016), a surge of high-growth new ventures following an influx of inventors may crowd out funding possibilities for companies in neighboring counties.

We address the possibility of negative spillovers by adding the spatially surrounding counties' incoming inventors to our baseline regressions. Motivated by Moretti and Wilson (2014) and Agrawal et al., (2017) we weigh the number of incoming inventors in surrounding counties by the inverse distance between the focal county's *i* geographic centroid and each surrounding county's *j* geographic centroid:

$$SpatialInventors_{i,t} = \sum_{j \neq i}^{J} w_{ij} \cdot IncomingInventors_{j,t-leave\ out\ inventors_{i,t}} \quad (5)$$

where w_{ij} represents the inverse distance between each county pair i and j and $IncomingInventors_{j,t-leave\ out\ inventors_{i,t}}$ is the number of incoming inventors into county j excluding those that come from i.

We generate three spatial lags—incoming inventors from counties within a distance of 50 miles, 70 miles and 100 miles of the focal county—and add this variable to our baseline model estimated above. We construct the corresponding instrument for equation (5) using the same weights w_{ij} combined with our instrument defined in (2). To avoid the confounding influences of inventors moving from the surrounding regions, we deviate from our baseline model by excluding inventors moving in from surrounding counties j when measuring the number of incoming inventors into i (within 50, 70, and 100 miles respectively). Our spillover regressions thus also address the question to which degree our baseline estimates reflect effects of short distance moves within broader regions.

All models of Table 7 indicate that inventors moving into surrounding counties lead to a displacement effect, (i.e. fewer high-growth startups in the focal county), as shown by the negative and statistically-significant estimated coefficients for the SpatialInventors covariates at various distances. Accounting for spatial correlations and disentangling the mechanisms of this effect is beyond the scope of this paper but unsurprising given localized knowledge spillovers (Jaffe, Tratenberg, & Henderson, 1993, Singh & Marx, 2013; Balsmeier et al. 2020). Also consistent with localized knowledge flows and agglomerative forces, we find that the displacement effect is decreasing in the distance between the focal and neighboring counties.

It is important to note that accounting for the displacement effect of inventors arriving in neighboring counties does little to disturb the main findings of Tables 3-6. Excluding inventors coming in from surrounding counties still leads to declining effect sizes with increasing distance, though the effects remain economically and statistically significant, suggesting that the main impact comes from inventors moving in from counties farther away than 100 miles. The results also cannot be explained by short distance moves within broader regions. Although it is impossible to derive an estimate of the impact of inventor mobility on an aggregated and national level entrepreneurship without a structural model, our regressions suggest that displacement effects may play a crucial role in such calculations.

Table 7 – Impact of local and spatially proximate incoming inventors on local high growth startup foundation

	High-growth startups founded		
	a	b	c
	IV	IV	IV
Incoming Inventors$_{t-1,\ ex\ 50\ miles}$	0.330***		
	(0.076)		
SpatialInventors$_{t-1,\ 50\ miles}$	-0.162**		
	(0.069)		
Incoming Inventors$_{t-1,\ ex\ 70\ miles}$		0.312***	
		(0.066)	
SpatialInventors$_{t-1,\ 70\ miles}$		-0.118**	
		(0.048)	
Incoming Inventors$_{t-1,\ ex\ 100\ miles}$			0.297***
			(0.062)
SpatialInventors$_{t-1,\ 100\ miles}$			-0.088**
			(0.038)
N	82,259	82,259	82,259
First Stage F Incoming Inventors$_{t-1,\ ex\ X\ miles}$	52.22	61.31	66.60
First Stage F SpatialInventors$_{t-1,\ X\ miles}$	51.80	63.39	72.30
State-Year FE	Yes	Yes	Yes
County FE	Yes	Yes	Yes

Notes: This table presents OLS regression of log(number of high growth startup foundation + 1), where high growth startups are defined following Guzman and Stern (2020) as newly registered companies that complete either an IPO or successful acquisition within 6 years. Incoming Inventors, Spatial Inventors as well as the corresponding instruments are log-transformed. First stage Fs are Sanderson-Windmeijer multivariate F test statistics of the first stage regressions where the linear projection of the other endogenous regressor is partialled out. Standard errors clustered at the county level appear in parentheses. ***, ** and * indicate a significance level of 1%, 5%, and 10%, respectively.

Discussion

The importance of scientists and inventors to positive economic and societal outcomes has been recognized since at least Vannevar Bush in 1945. We explicitly linked one mechanism for that positive relationship, tying scientists and inventors to entrepreneurship, which in turn has been shown to lead to other desirable outcomes, such as future employment growth (Glaeser, Kerr, and Kerr 2015). Bringing evidence to the argument, this work estimates the benefits of inventors to high-growth regional entrepreneurship. Geography in this respect is just an instrument to get at that estimate, however, it allowed us to back out an estimate that the arrival of 55 inventors to a county increased entrepreneurship in that county by one firm, on average.

While the work remains preliminary and premature for policy recommendations, it would seem that regions should seek to bolster their STEM workforces. Given that many children demonstrate the capability to become inventors at an early age, and yet many do not, would appear to offer a straightforward (though certainly not inexpensive) path to increasing entrepreneurship, through early STEM education (Bell et. al. 2019). While this work used immigration to back out the value of an inventor, it would seem reasonable that a home grown inventor might be just as useful to local entrepreneurship. Indeed, if a home grown inventor had easier access to existing networks of friends, family, investors, and fellow entrepreneurs, they might be even more effective at starting high tech firms. It would also be interesting to understand if immigration crowds out – or complements -- locally grown inventors.

Taking our full sample estimates and summary statistics as the baseline for a back-of-the-envelope calculation of the relative importance of inventors suggests that incoming inventors can explain 24.9% of all high growth startups. This estimate is preliminary and does not include home grown inventor contributions, scientists who do not patent, or the negative effects on surrounding counties. If this estimate holds, however, it would indicate that inventors and technical professionals create a very substantial part of high growth entrepreneurship.

Caveats notwithstanding, we connect the current results with prior work on research on competition covenants. Strengthening of non-compete policies has been shown to precede inventor emigration (Marx, Singh, and Fleming, 2015). While policy makers might hope to encourage regional investment in research with stronger non-competes, because such policies make it more difficult for an engineer to leave for a local competitor, such policies might also dampen local entrepreneurship (startups are also typically covered by non-competes). Emigration to entrepreneurial opportunities might be one cause of the historically observed shift of inventors, especially the most highly cited and most collaborative, from states which enforce non-competes to those that do not (Marx and Fleming 2012).

There are some obviously immediate directions for the current work. While this paper established a baseline for all inventors, it would be straightforward to investigate the impact of a variety of characteristics on entrepreneurship, such as highly cited or collaborative inventors, inventors in different fields, inventors who rely upon or publish in the scientific literature, those whose professional history is in large firms, small firms, startups, or academia, or those who have been supported by Federal research. Individual level analyses could also be productive, for example, if one can instrument away the technology motivation for moving towards or away from similar inventors, then what is the productivity and career impact of moving towards or away from the center of technology in an inventor's field? Furthermore, what is the impact of more diverse inventors on a region's innovation and productivity? Does the region become more creative with increasing diversity, or does it become less productive? Are there time lags as inventors learn new

fields and develop new collaborative relationships? The methods here might also be applied to scientists, though the lack of widely available and robust disambiguation of the science bibliometric databases presents a sobering barrier to easy progress. Inventors and scientists may also have other positive impacts on a regional economy, for example, by increasing productivity and decreasing unemployment.

Conclusion

Has Silicon Valley become the world's hub of entrepreneurship due in part to its supply of inventive human capital, or has the region simply acted as a "magnet" to attract scientists and inventors to its former orchards? Should would-be entrepreneurs locate in a region with a vibrant supply of creatives and inventors (Florida, 1995), or should founders focus on finding sufficient financial capital to outsource and/or in-license innovative resources? Ought policymakers—especially those in areas without a strong record of entrepreneurship—prioritize science parks and accelerators, or might their efforts be productively redirected toward attracting and retaining skilled human capital locally?

Our results suggest that the local availability of skilled human capital is a critical determinant of high-growth entrepreneurship. Our shift-share instrument, based on the county-level distribution of surnames in the 1940 census, overcomes limitations of prior approaches, enabling us to provide estimates of the impact of a marginal inventor on entrepreneurial activity. We find that in some regions, an increase of only 29 inventors predicts one additional high-growth startup in a county. We also find that immigration of inventors to surrounding counties has a negative and significant effect on entrepreneurship in the focal county.

The mobility instrument developed here might also provide purchase into other questions on regional success, for example, how important is the agglomeration of inventors – is Silicon Valley exciting because it draws inventors, or because inventors moved there? We know that scientists and engineers tend to be in areas where there is more going on, but is that demand or supply? Surely they flock to opportunities, so it is hard to know which way the arrows point, or assuming they point in both directions, their net effect. The work underscores prior findings regarding the importance of labor mobility for entrepreneurship, for example by limiting the ability of firms to use employee non-compete agreements (Samila & Sorenson, 2011). In short, both stocks and flows matter.

References

Acemoglu, D., Akcigit, U., Alp, H., Bloom, N., Kerr, W. (2018). Innovation, Reallocation and Growth. American Economic Review 108(11), 3450-91

Ackerberg, D. A., Caves, K., & Frazer, G. (2015). Identification properties of recent production function estimators. Econometrica, 83(6), 2411-2451.

Adao, R., Kolesár, M. and Morales, E., 2019. Shift-share designs: Theory and inference. The Quarterly Journal of Economics, 134(4), pp.1949-2010.

Agrawal, A. and Cockburn, I., 2003. The anchor tenant hypothesis: exploring the role of large, local, R&D-intensive firms in regional innovation systems. International journal of industrial organization, 21(9), pp.1227-1253.

Agrawal, A., Galassso, A. and A. Oettl (2017). Roads to Innovation. Review of Economics and Statistics 99 (3), 417-434.

Agrawal, A., Kapur, D., McHale, J. and Oettl, A., 2011. Brain drain or brain bank? The impact of skilled emigration on poor-country innovation. Journal of Urban Economics, 69(1), pp.43-55.

Arrow, K.J. and Capron, W.M., 1959. Dynamic shortages and price rises: the engineer-scientist case. The Quarterly Journal of Economics, 73(2), pp.292-308.

Balsmeier, B., Fierro, G., Li, G., Johnson, K., Kaulagi, A., O'Reagan, D., Yeh, W., Lueck, S., and L. Fleming (2017). Machine learning and natural language processing applied to the patent corpus. Forthcoming, Journal of Economics and Management Strategy.

Bartel, A.P., 1989. Where do the new US immigrants live? Journal of Labor Economics, 7(4), pp.371-391.

Bartik, T.J., 1991. Who benefits from state and local economic development policies?

Bell, A. and R. Chetty, X. Jaravel, N. Petkova, J. Van Reenen 2019. "Who Becomes an Inventor in America? The Importance of Exposure to Innovation." The Quarterly Journal of Economics, Volume 134, Issue 2, May 2019, Pages 647–713.

Bernstein, S. and R. Diamond, T. McQuade, B. Pousada (2018). "The Contribution of High-Skilled Immigrants to Innovation in the United States, " Working paper, Stanford University.

Bernstein, S., Giroud, X., & Townsend, R. R. (2016). The impact of venture capital monitoring. The Journal of Finance, 71(4), 1591-1622.

Borusyak, K., Hull, P. and Jaravel, X., 2018. Quasi-experimental shift-share research designs (No. w24997). National Bureau of Economic Research.

Burchardi, K.B., Chaney, T., Hassan, T.A., Tarquinio, L. and Terry, S.J., 2020. Immigration, Innovation, and Growth (No. w27075). National Bureau of Economic Research.

Cappelli, R., Czarnitzki, D., Doherr, T. and Montobbio, F., 2019. Inventor mobility and productivity in Italian regions. Regional Studies, 53(1), pp.43-54.

Card, D., 2001. Immigrant inflows, native outflows, and the local labor market impacts of higher immigration. Journal of Labor Economics, 19(1), pp.22-64.

Clark, G. and Cummins, N., 2015. Intergenerational wealth mobility in England, 1858–2012: surnames and social mobility. The Economic Journal, 125(582), pp.61-85.

Cheyre, C., Klepper, S., & Veloso, F. (2015). Spinoffs and the mobility of US merchant semiconductor inventors. Management Science, 61(3), 487-506.

Darlu, P., Brunet, G. and Barbero, D., 2011. Spatial and temporal analyses of surname distributions to estimate mobility and changes in historical demography: the example of Savoy (France) from the eighteenth to the twentieth century. In Navigating time and space in population studies (pp. 99-113). Springer, Dordrecht.

Duranton, G. and Puga, D., 2001. Nursery cities: Urban diversity, process innovation, and the life

cycle of products. American Economic Review, 91(5), pp.1454-1477.

Ewens, M. and Marx, M., 2018. Founder replacement and startup performance. The Review of Financial Studies, 31(4), pp.1532-1565.

Florida, R., 2005. Cities and the creative class. Routledge.

Glaeser, E.L. and Kerr, W.R., 2009. Local industrial conditions and entrepreneurship: how much of the spatial distribution can we explain?. Journal of Economics & Management Strategy, 18(3), pp.623-663.

Glaeser, E.L. and Kerr, S., Kerr, W. 2015. Entrepreneurship and Urban Growth: An Empirical Assessment with Historical Mines. Review of Economics and Statistics. 97:2: 498-520.

Grilli, L. and Murtinu, S., 2014. Government, venture capital and the growth of European high-tech entrepreneurial firms. Research Policy, 43(9), pp.1523-1543.

Guzman, J. and S. Stern, (2019). "The State of American Entrepreneurship: New Estimates of the Quality and Quantity Of Entrepreneurship for 32 US States, 1988-2014," NBER Working paper 22095.

Hathaway, I. (2018). "High-growth firms and cities in the US: an Analysis of the Inc. 5000." Brookings Institution report.

Hofstede, G., 2001. Culture's consequences: Comparing values, behaviors, institutions and organizations across nations. Sage publications.

Hunt, J., 2017. The impact of immigration on the educational attainment of natives. Journal of Human Resources, 52(4), pp.1060-1118.

Hunt, J. and Gauthier-Loiselle, M., 2010. How much does immigration boost innovation?. American Economic Journal: Macroeconomics, 2(2), pp.31-56.

Jaffe, A.B., Trajtenberg, M. and Henderson, R., 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. the Quarterly journal of Economics, 108(3), pp.577-598.

Jaeger, D.A., Ruist, J. and Stuhler, J., 2018. Shift-share instruments and the impact of immigration (No. w24285). National Bureau of Economic Research.

Jensen, Richard, and Marie Thursby. 2001. "Proofs and Prototypes for Sale: The Licensing of University Inventions." American Economic Review, 91 (1): 240-259.

Kerr, W.R., 2013. US high-skilled immigration, innovation, and entrepreneurship: Empirical approaches and evidence (No. w19377). National Bureau of Economic Research.

Kerr, W.R. and Lincoln, W.F., 2010. The supply side of innovation: H-1B visa reforms and US ethnic invention. Journal of Labor Economics, 28(3), pp.473-508.

King, G., and Nielsen, R. (2017). Why Propensity Scores Should Not Be Used for Matching. Harvard Working Paper.

Klepper, S., 2009. Spinoffs: A review and synthesis. European Management Review, 6(3), pp.159-171.

Kogan L., D. Papanikolaou, A. Seru and N. Stoffman (2017). Technological Innovation, Resource Allocation and Growth. Quarterly Journal of Economics, 132(2), 665–712.

LaLonde, R.J. and Topel, R.H., 1991. Labor market adjustments to increased immigration. In Immigration, trade, and the labor market (pp. 167-199). University of Chicago Press.

Lerner, J., 2012. Boulevard of broken dreams: why public efforts to boost entrepreneurship and venture capital have failed--and what to do about it. Princeton University Press.

Lerner, J. and Seru, A., (2017). The Use and Misuse of Patent Data: Issues for Corporate Finance and Beyond, NBER Working Paper No. 24053.

Maloney, W.F. and Valencia Caicedo, F., 2016. The persistence of (subnational) fortune. The Economic Journal, 126(598), pp.2363-2401.

M. Marx and J. Singh, L. Fleming, “Regional Disadvantage? Employee Non-compete Agreements and Brain Drain.” Research Policy 44 (2015) 941-955.

Marx, M., and L. Fleming, 2012. “Noncompetes: Barriers to Exit and Entry?” National Bureau of Economic Research Innovation Policy and the Economy, eds. Stern and Lerner, 12: 39-64. University of Chicago Press.

Marx, M. and Hsu, D.H., 2019. The Entrepreneurial Commercialization of Science: Evidence From 'Twin' Discoveries. Boston University Questrom School of Business Research Paper.

Moretti, E. (2012). The New Geography of Jobs. Houghton Mifflin Harcourt, N. Y., N.Y.

Moretti, E., Steinwender, C., and Van Reenen, J. (2016). The Intellectual Spoils of War? Defense R&D, Productivity and International Technology Spillovers, Working Paper.

Moretti, E. and D. J. Wilson (2014). State incentives for innovation, star scientists and jobs:

Evidence from biotech. Journal of Urban Economics 79, 20-38.

Moser, P., A. Voena, and F. Waldinger. 2014. "German Jewish Émigrés and US Invention." American Economic Review, 104 (10): 3222-55.

Parey, M., Ruhose, J., Netz, N., and F. Waldinger 2017. The selection of high-skilled emigrants. The Review of Economics and Statistics 99(5), pp. 776-792.

Peri, G., Shih, K. and Sparber, C., 2015. STEM workers, H-1B visas, and productivity in US cities. Journal of Labor Economics, 33(S1), pp.S225-S255.

Romer, P.M., 1990. Endogenous technological change. Journal of political Economy, 98(5, Part 2), pp.S71-S102.

Rosenberg, N. and Nelson, R.R., 1994. American universities and technical advance in industry. Research policy, 23(3), pp.323-348.

Rosenthal, S.S. and Strange, W.C., 2004. Evidence on the nature and sources of agglomeration economies. In Handbook of regional and urban economics (Vol. 4, pp. 2119-2171). Elsevier.

Samila, S. and Sorenson, O., 2011. Venture capital, entrepreneurship, and economic growth. The Review of Economics and Statistics, 93(1), pp.338-349.

Saxenian, A., 1996. Regional advantage. Harvard University Press.

J. Singh and M. Marx. “Geographic Constraints on Knowledge Diffusion: Political Borders vs. Spatial Proximity.” Management Science 59(9):2056-2078 (2013).

Stuart, T.E. and Ding, W.W., 2006. When do scientists become entrepreneurs? The social structural antecedents of commercial activity in the academic life sciences. American journal of sociology, 112(1), pp.97-144.

Stuart, T., & Sorenson, O. (2003). The geography of opportunity: spatial heterogeneity in founding rates and the performance of biotechnology firms. Research policy, 32(2), 229-253.

Toivanen, O., and Väänänen, L. (2016). Education and Invention. Review of Economics and Statistics, 98(2), 382 - 396.

Waldinger, F. 2016. Bombs, Brains, and Science: The Role of Human and Physical Capital for the Production of Scientific Knowledge. The Review of Economics and Statistics 98(5), pp. 811-831.

Wozniak, A. and Murray, T.J., 2012. Timing is everything: Short-run population impacts of immigration in US cities. Journal of Urban Economics, 72(1), pp.60-78.

Zucker, L.G., Darby, M.R. and Brewer, M.B., 1994. Intellectual capital and the birth of US biotechnology enterprises (No. w4653). National Bureau of Economic Research.

Appendix A1: Matching between surnames in patent and Census data

Matching surnames between Census and patent data required several steps. First, we cleaned the surnames in the inventor data and matched them with surnames in the census data. We converted all surnames to lower cases and deleted unnecessary punctuations and other noise in the surnames of the inventor data (e.g., ' "_ @ / & ; ? ` () # =, which were particularly important when they were the first or last character). We also removed suffixes and other extra words after commas (e.g., 'Foster', 'Sr.', 'deceased'). This process enabled 43,592 additional matches for raw inventor surnames. Then, for surnames containing a dash or apostrophe, e.g., "O'Brien", "Villa-Real", we replaced the punctuation with choices of having a space or without a space, and collected matching candidates of each surname from the census data. If there existed both surname cases with a space and without a space, we set the algorithm to select the one with a higher frequency. There were a total of 396 surnames matched in this process (221 for surnames with a dash, 175 for surnames with an apostrophe), for example: inventor surname "O'Brien" matched with census surname "Obrien"; inventor surname "Ben-Bassat" matched with census surname "Benbassat".

We tokenized the remaining surnames, i.e., split surnames with multiple words by spaces, and matched them with census surnames that contained all the tokens. To be conservative, we set the algorithm to collect surname candidates from the census data for only those surnames that contained all tokens as well as surnames with a string length shorter than the length of original inventor surname. After collecting matching candidates, we compared the text to the original inventor surname and selected the one with the highest similarity. For the similarity measurement, we used both Jaro Winkler and Damerau/Levenshtein similarity measurements. The value of a Damerau/Levenshtein similarity of 0.75 was used as a threshold to be considered as a matching candidate. Conditional on satisfying this threshold, the algorithm compared the census surname candidates using the Jaro Winkler similarity. If multiple candidates have the same Jaro Winkler similarity value, the algorithm compared Damerau/Levenshtein similarity values. If multiple candidates with the same similarity value still existed, our algorithm was set to choose one with a higher frequency in the census data. There were a total of 724 surnames matched in this procedure, for example, the surname "de la Merced" matches with census surname "Delamerced" (Jaro Winkler similarity: 0.96 / Damerau/Levenshtein: 0.83), and inventor surname "van de Vaart" matches with census surname "Vande vaart" (Jaro Winkler similarity: 0.96 / Damerau/Levenshtein: 0.92). In this case, there was also another census surname candidate "Vandervaart", but it was not selected as it has a lower Jaro Winkler similarity value compared to "Vande vaart". Inventor surname 'van der Loo' matched with census surname 'Vanderloo' (Jaro Winkler similarity: 0.96 / Damerau/Levenshtein: 0.81). There was another census surname candidate 'Vander loo', but it was not selected as it had a lower Jaro Winkler similarity than 'Vanderloo'. The 'Vanderloof', 'Vanderlool', 'Vanderloon', 'Vanderloop', 'Vanderloot' surnames were not considered as candidates as their Damerau/Levenshtein similarity values failed to exceed the threshold of 0.75.

Finally, we matched multi-word census surnames to inventor surnames that failed to match in the prior procedures. After splitting the surnames by spaces, we set the algorithm to find inventor surname candidates that contained every tokenized word of the census surname. Then, applying the same similarity indexes and processes described above, the algorithm found a final match between a census surname and an inventor surname. There were a total of 716 inventor surnames matched through this step, including, for example, "Von Doenhoff" matched with inventor surname "Vondoenhoff" (Jaro Winkler similarity: 0.98 / Damerau levenshtein: 0.92), and census surname "Mc Ellistrem" matched with inventor surname "Mcellistrem" (Jaro Winkler similarity: 0.98 / Damerau levenshtein: 0.92). As a result of the entire matching processes described above, a total of 275,849 out of 374,988 unique surnames of inventors (73.6%) found a match in the census surname. Compared to the matching without these whole processes, which found 230,421 census surname matches out of 374,988 unique inventor surname raw strings (61.4%), our algorithm added 12.2% of matches. In our data sample specifically, out of 769,625 unique inventors that applied for at least one patent in US, 701,215 inventors (91.1%) matched their surname to the 1940 Census data.

Table A1 – Impact of incoming inventors on local high growth startup foundation in California and Massachusetts

	High-growth startups founded			
	a	b	c	d
	OLS	IV	IV	IV
Incoming Inventors$_{t-1}$	0.493***	0.569***	0.569***	0.325**
	(0.045)	(0.059)	(0.059)	(0.136)
N	1,929	1,929	1,929	1,929
First Stage F		117.3	115.6	16.24
Year FE	Yes	Yes	No	No
State FE	Yes	Yes	No	No
State-Year FE	No	No	Yes	Yes
County FE	No	No	No	Yes
R^2	0.677			

Notes: This table presents OLS regression of log(number of high growth startup foundation + 1), where high growth startups are defined following Guzman and Stern (2020) as newly registered companies that complete either an IPO or successful acquisition within 6 years. Incoming inventors as well as the instrument are log-transformed. Specifications (b)-(d) show results of our IV regression as described above, where incoming inventors are instrumented with $Inv_{dt,leave-out}$ in the first stage. First stage F is the Kleibergen-Paap Wald F statistic of the first stage regression. Model e excludes all counties of California and Massachusetts. Standard errors clustered at the county level appear in parentheses. ***, ** and * indicate a significance level of 1%, 5%, and 10%, respectively.