

2022-11

Reprint of “The startup cartography project: measuring and mapping entrepreneurial ecosystems”

R.J. Andrews, C. Fazio, J. Guzman, Y. Liu, S. Stern. 2022. "Reprint of “The Startup Cartography Project: Measuring and mapping entrepreneurial ecosystems”" Research Policy, Volume 51, Issue 9, pp.104581-104581. <https://doi.org/10.1016/j.respol.2022.104581>
<https://hdl.handle.net/2144/46152>

"Downloaded from OpenBU. Boston University's institutional repository."

THE STARTUP CARTOGRAPHY PROJECT:
Measuring and Mapping Entrepreneurial Ecosystems

RJ Andrews

Catherine Fazio

Jorge Guzman

Yupeng Liu

Scott Stern

RJ Andrews: Info We Trust. rj@infowetrust.com. Cathy Fazio: Boston University, Questrom School of Business, cfazio@bu.edu. Jorge Guzman: Columbia University, Columbia Business School, jag2367@gsb.columbia.edu. Yupeng Liu: Columbia University, Columbia Business School. y4180@gsb.columbia.edu. Scott Stern: MIT, Sloan School of Management. sstern@mit.edu

© 2020 by RJ Andrews, Catherine Fazio, Jorge Guzman, Yupeng Liu and Scott Stern. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

ABSTRACT

This paper presents the Startup Cartography Project, which offers a new set of entrepreneurial ecosystem statistics for the United States from 1988-2016. The SCP combines state-level business registration records with a predictive analytics approach to estimate the probability of “extreme” growth (IPO or high-value acquisition) at or near the time of founding for all newly-registered firms in a given year. The results indicate the ability of predictive analytics to identify high-potential start-ups at founding (using a variety of different approaches and measures). The SCP then leverages estimates of entrepreneurial quality to develop four entrepreneurial ecosystem statistics, including the rate of start-up formation, average entrepreneurial quality, the quality-adjusted quantity of entrepreneurship, and entrepreneurial ecosystem performance over time. These statistics offer sharp insight into patterns of regional entrepreneurship, their correlation with regional economic growth and the evolution of entrepreneurial ecosystems over time. The SCP includes both a public-access dataset at the state, MSA, county, and zip code level, as well as an interactive map, the U.S. Startup Map, that allows academic and policy users to assess entrepreneurial ecosystems at an arbitrary level of granularity (from the level of states down to individual street addresses).

I. Introduction

Over the past two decades, there has been a dramatic increase in interest by both academic researchers and policymakers in the role of startup companies in regional economic performance (Saxenian, 1994; Feldman, 2001; Lerner, 2009). This interest reflects both increasing appreciation for the empirical linkage between the two (Feldman et al, 2005; Glaeser et al, 2015), and also the outsized success of particular regions such as Silicon Valley that have hosted waves of start-up firms and experienced a high and sustained level of innovation-driven entrepreneurial dynamism as a result. Relative to traditional economic development incentives (e.g., such as tax breaks for large employers), the promotion of regional entrepreneurship aims to nurture the establishment and expansion of new firms at a relatively low cost in order to benefit from the growth of (some of) those firms over time. For example, in the United States, a host of programs have been initiated over the past decade to foster entrepreneurial ecosystems, ranging from the US Economic Development Association Regional Innovation Strategies (i6) (EDA, 2010), to the Kauffman Foundation ESHIP Communities initiative (Kauffman Foundation, 2019a), to private sector efforts such as “Rise of the Rest.” (Revolution, 2019). Perhaps not surprisingly, active debate exists around the design and structure of policies intended to promote regional “entrepreneurial ecosystems” (Feldman and Francis, 2004; Lerner, 2009; Audrestch and Lehmann, 2005; Stam, 2015).

Beyond important conceptual challenges in defining the nature of entrepreneurial ecosystems (Kauffman Foundation, 2019b; Feld, 2012; Murray and Stern, 2015), the evaluation of entrepreneurial ecosystems involves an important empirical challenge: how can one measure the state of an entrepreneurial ecosystem at a point in time, track changes in that system over time,

or make comparisons across regions (or within regions at different levels of geographic granularity)?

Confronting this challenge requires addressing three interrelated issues: skewness, lagged performance and multiple levels of geographic analysis. First, while the bulk of regional economic growth is linked to the scaling of young firms, startup growth, in and of itself, is heavily skewed. A relatively small number of “successes” from any given cohort of startups has a disproportionate impact on the overall cohort’s economic performance. Measurement of an entrepreneurial ecosystem needs to somehow link the measurement of entrepreneurship to these potential skewed outcomes. Put another way, in evaluating the potential of an entrepreneurial ecosystem, it is important to measure not only the *quantity* of new startups being formed there, but also their potential for growth (i.e., their “entrepreneurial *quality*”). Second, the impact of an entrepreneurial ecosystem on a regional economy occurs only after a considerable lag in time between the founding of new firms there and the emergence of skewed outcomes (at least five but maybe as many as 10-20 years!). The combination of skewed outcomes and long time lags makes the assessment of an entrepreneurial ecosystem, and the timely evaluation of policies and programs intended to promote it, challenging even when the ecosystem is strong and programs are working as intended. Third, ecosystems occur at multiple levels of geographic analysis, ranging from small clusters of firms in a few individual locations (e.g., the ecosystem surrounding a university campus) to broader regions at the level of cities, counties, states, or even countries. Any empirical assessment of entrepreneurial ecosystems thus also needs to account for the fact that ecosystems can be evaluated (and will need to be measured) at multiple levels of analysis.

By and large, prior efforts to overcome these measurement challenges have relied on one of two broad approaches. One central approach focuses on measuring the quantity of new firms

founded at a given point in time (within a fixed geographic domain). Most notably, the Longitudinal Business Database (LBD) (Jarmin and Miranda, 2002) provides enormously valuable insight into the number of new firms (with at least one employee) founded by year and state. However, despite its considerable strengths, the LBD (and related datasets such as the Business Dynamics Statistics) tend to abstract away from differences among firms at the time of founding, and the statistics that are made available may only be produced (at least for public-use) at a relatively high rate of aggregation (such as a state). Alternatively, it is possible to simply condition the study of an entrepreneurial ecosystem on those firms resident there that satisfy a pre-determined performance criteria, such as the receipt of venture capital (less than 0.1% of all newly founded firms receive venture capital in any given year), or achievement of a certain level of employment growth within a limited time-frame (see Stanger and Bell-Materson, 2015). However, for many purposes, conditioning the analysis of an ecosystem on such milestones conflates the measurement of the rate of entrepreneurship found there with the assessment of overall entrepreneurial performance. If the rate of venture financing in a given region is low or declining, for example, does that imply that there is “too little” venture capital or too few firms found there with the potential to attract it?

The purpose of this paper is to introduce a new database and mapping platform, the Startup Cartography Project (“SCP”), that aims to address these challenges in an integrated manner for both academic and policy users.¹ Specifically, the core objective of the SCP is to provide a

¹ The Startup Cartography Project builds upon but also substantially extends our own prior work on measuring the quantity and quality of entrepreneurship (Guzman and Stern, 2015, 2017; Fazio, et al, 2017). Specifically, as described below, the SCP has now been extended to 49 states (and Washington, D.C.) through 2014 (and 46 through 2016), and also includes a more comprehensive underlying dataset in terms of equity growth events. As well, building on Guzman and Stern (2020), we now develop a more comprehensive set of entrepreneurial ecosystem metrics, including REAI (which measures the performance of entrepreneurial ecosystems over time). Finally, this paper provides the foundation for the underlying now-released SCP data and interactive U.S. Startup Map mapping tool.

consistent, transparent and accessible data resource that allows for granular (as well as aggregated) and timely (as well as retrospective) measurements of entrepreneurial ecosystems. The SCP incorporates three broad elements. First, building on Guzman and Stern (2015, 2017, 2020), the SCP uses a predictive analytics approach to estimate, for any given startup, the probability of growth of that firm at or near the time of founding (a measure of its quality). Second, leveraging this measure of entrepreneurial “quality” for all firms, the SCP builds a set of novel entrepreneurship statistics that capture the quantity, quality and performance of any given set of firms, allowing for consistent measures of entrepreneurship across time and place. Finally, we then translate the core SCP statistics into an interactive mapping tool, the U.S. Startup Map, that allows for dynamic and interactive visualization of entrepreneurship at an arbitrary level of aggregation (from an individual street address up to the level of the United States).

Our approach builds upon and extends our own prior work (including Guzman and Stern (2015, 2017, 2020)) leveraging three core insights. First, because the challenges to growth as a sole proprietorship are formidable, a practical requirement for any growth-oriented entrepreneur that would contribute to an entrepreneurial ecosystem is business registration (as a corporation, partnership, or limited liability company). We take advantage of the public nature of business registration records to define a population sample of entrepreneurs observed at a similar (and foundational) stage of the entrepreneurial process. Second, moving beyond simple counts of business registrants (Klapper, Amit, and Guillen, 2010), we are able to measure characteristics related to entrepreneurial quality *at or close to the time of registration*. For example, we can measure start-up characteristics such as whether the firm is organized in order to facilitate equity financing (e.g., registering as a corporation or in Delaware), how the firm is named (e.g., whether it signals a high-tech sector versus a local focus) or whether the firm acquires or develops

measurable innovations (e.g., a patent or trademark). Third, we leverage the fact that, though rare, we observe meaningful growth outcomes for some firms (e.g., those that achieve an IPO or high-value acquisition within six years of founding), and are therefore able to estimate the relationship between these growth outcomes and start-up characteristics. In other words, our approach implements a predictive analytics approach to entrepreneurship which allows us to estimate, for any given firm, its underlying level of quality (as linked to particular observables) at or near the time of founding.

We apply this predictive analytics approach in the context of 49 U.S. states and Washington D.C. from 1988-2014, and 46 U.S. states within the year 2014-2016 (a significant extension beyond our earlier work). Consistent with Guzman and Stern (2015, 2017), we find that a small number of characteristics allow us to develop a robust predictive model that distinguishes firm quality. In an out-of-sample test, we find that 56% of realized growth outcomes occur in the top 5% of our estimated quality distribution (and nearly 37% in the top 1% of the estimated quality distribution). Moreover, we find that a small number of governance and intellectual property characteristics – Delaware registration, registering for a trademark or patent application – are the single largest factors predicting subsequent start-up performance. However, our work with policymakers and other analysts suggested that the ability to actually utilize a predictive analytics approach was more persuasive if we focused on a modified version of the model where we focus only on those start-up characteristics that are closely linked with the legal and intellectual property environment surrounding the firm (i.e., Delaware registration, and the registration of patents or trademarks). For concision, we refer to the richer model incorporating such features as eponymy as the “academic” model, and we refer to the model that relies exclusively on institutional features as the “policy” model.

We then use these estimates to generate four aggregate economic statistics for the measurement of entrepreneurship: the Startup Formation Rate (SFR), the Entrepreneurship Quality Index (EQI), the Regional Entrepreneurship Cohort Potential Index (RECPI) and the Regional Ecosystem Acceleration Index (REAI). SFR is simply a measure of new firm formation (within a cohort of firms defined by a given time period and geographic scope). EQI is a measure of *average quality* within any cohort, allowing for the calculation of the probability of a growth outcome within a specified population of start-ups. RECPI multiplies SFR and EQI within a given geographic domain (e.g., a town or even the state of Massachusetts), yielding a measure of the quality-adjusted quantity of entrepreneurship within that ecosystem. Whereas EQI compares entrepreneurial quality across different groups (and so facilitates apples-to-apples comparisons across groups of different sizes), RECPI allows the direct calculation of the expected number of growth outcomes from a given start-up cohort within a given regional boundary. Finally, REAI, measured as the ratio of realized to expected growth events, is a measure of entrepreneurial ecosystem performance in accelerating startups after founding. While RECPI estimates the expected number of growth events for a given group of firms, over time we can observe the realized number of growth events from that cohort. This difference (reflected in REAI) can be interpreted as the relative ability of firms within a given region to grow, conditional on their initial entrepreneurial quality. Variation in ecosystem performance could result from differences across regional ecosystems in their ability to nurture the growth of start-up firms, or changes over time or location in financing availability, economic conditions, or economic policies or programs.

We construct these statistics at the state, MSA, county and zip code level, and illustrate the potential of these data for regional analysis by undertaking a descriptive examination of the 100 largest MSAs in the United States. A few key findings stand out. On the one hand, both the

average level of EQI and RECPI/Population are much higher for key regions that have been traditionally associated with growth entrepreneurship, such as the Bay Area (San Francisco and San Jose, CA), as well as Boston, MA and Austin, TX. At the same time, the SCP captures (in a timely way) the recent growth in entrepreneurial ecosystems such as Provo, UT, and Denver, CO. In addition, our descriptive analysis offers a novel lens through which to view the linkage between entrepreneurship and regional economic growth. Whereas the quantity of entrepreneurship is essentially uncorrelated with subsequent regional economic growth, the quality of entrepreneurship in a given ecosystem is strongly correlated with subsequent regional economic performance.

Finally, we use these statistics to build an interactive map, the U.S. Startup Map, that visualizes entrepreneurial ecosystems across time and place. Specifically, the U.S. Startup Map allows individual users to choose both the timeframe for analysis (i.e., a given year) as well as the level of geographic granularity (ranging from the United States down to the level of individual street addresses), and provides a visualization of both the SFR and EQI for that chosen geography. As suggested earlier, feedback from policy users suggested that we adopt the “policy” model for our visualization, since users of the interactive map are more likely to be interested in identifying start-up populations linked to institutional factors such as Delaware registration, or trademark or patent applications. By helping to stakeholders to see the results of quantitative academic empirical research, our work with the U.S. Startup Map also holds broader implications for policy and practice.

The rest of this paper proceeds as follows. Section II presents our approach for constructing entrepreneurial ecosystem statistics. Section III reviews data and estimation. Section IV calculates the key entrepreneurial quality statistics, and overviews some broad descriptive patterns we

observe from the SCP metrics. Section V describes the application of the SCP for policy, and the development of an interactive and dynamic visualization tool, the U.S. Startup Map, that allows users to assess entrepreneurial ecosystems in any time period of their choosing and at an arbitrary level of granularity. Section VI concludes.

II. Measuring Entrepreneurial Ecosystem Quantity, Quality and Performance²

A central challenge in assessing entrepreneurial ecosystems (for a wide range of both academic and policy questions) is the development of measures of entrepreneurial ecosystems that are (at least potentially) comparable across different ecosystems and over time. A central contribution of the Startup Cartography Project is to introduce a set of consistent measures that account not only for the quantity of entrepreneurs but also for the quality of those entrepreneurs at the time of founding. Our approach leverages that fact that while there are a very large number of new businesses established at any point in time (and so attempting to categorize them through an external assessment would be burdensome), entrepreneurs themselves have information about both their underlying idea and ambition, and make choices at the time of founding consistent with their objectives and potential for growth. Specifically, by starting with the entire population of business registrants (a relatively low-stakes requirement for any business in the United States that have any ambition to move beyond self-employment), it is possible to use a predictive analytics approach that relates the ultimate performance of start-up firms to initial early-stage choices by the entrepreneur that are also observable at or around the time of founding as a “digital signature” for

² This section draws upon but also extends the discussion of the estimation of entrepreneurial quality, as well as the development of statistics for a “cohort” of start-ups from Guzman and Stern (2015, 2017, 2020). However, relative to those earlier works, this paper both extends the range of statistics reported, the coverage of data that are now incorporated into the SCP, and links the statistics to the U.S. Startup Map, a novel interactive mapping tool.

each firm. We measure entrepreneurial quality by estimating the relationship between observed growth outcomes and start-up characteristics using the population of at-risk firms. For a firm i born in region r at time t , with at-birth start-up characteristics $H_{i,r,t}$, we observe growth outcome $g_{i,r,t+s}$ s years after founding and estimate:

$$\theta_{i,r,t} = P(g_{i,r,t+s}|H_{i,r,t}) = f(\alpha + \beta H_{i,r,t}) \quad (1)$$

This model allows us to *predict* quality as the probability of achieving a growth outcome given start-up characteristics at founding, and so estimate entrepreneurial quality as $\hat{\theta}_{i,r,t}$. As long as the process by which start-up characteristics map to growth remain stable over time (an assumption which is itself testable), this mapping allows us to form an estimate of entrepreneurial quality for any business registrant within our sample (even those in recent cohorts where a growth outcome (or not) has not yet had time to be observed).

The predictive analytics approach implemented in (1) allows us to recover an estimate for the entrepreneurial quality of any given firm at (or near) the time of founding. However, we then need to undertake a second step in which we form consistent and meaningful entrepreneurial ecosystem metrics that allow for comparisons across different ecosystems and across time. Specifically, the Startup Cartography Project provides users with four key entrepreneurship statistics capturing the rate of formation of registered firms, the level of entrepreneurial quality for a given population of start-ups, the potential for growth entrepreneurship within a given region and start-up cohort, and the performance over time of a regional entrepreneurial ecosystem in realizing the potential performance of firms founded within a given location and time period.

The Startup Formation Rate (SFR) represents the quantity of for-profit, new business registrants within a given population. It mimics other quantity based measures available such as the Business Dynamics Statistics (BDS) or the Global Entrepreneurship Monitor (GEM).

The Entrepreneurial Quality Index (EQI). To create an index of entrepreneurial quality for any group of firms (e.g., all the firms within a particular cohort or a group of firms satisfying a particular condition), we simply take the *average* quality within that group. Specifically, in our regional analysis, we define the *Entrepreneurial Quality Index (EQI)* as an aggregate of quality at the region-year level by simply estimating the average of $\theta_{i,r,t}$ over that region:

$$EQI_{r,t} = \frac{1}{N_{r,t}} \sum_{i \in \{I_{r,t}\}} \theta_{i,r,t} \quad (2)$$

where $\{I_{r,t}\}$ represents the set of all firms in region r and year t , and $N_{r,t}$ represents the number of firms in that region-year. To ensure that our estimate of entrepreneurial quality for region r reflects the quality of start-ups in that location rather than simply assuming that start-ups from a given location are associated with a given level of quality, we exclude any location-specific measures $H_{r,t}$ from the vector of observable start-up characteristics.³

The Regional Entrepreneurship Cohort Potential Index (RECPI). From the perspective of a given region, the overall inherent potential for a cohort of start-ups combines both the quality of

³ Three particular features of EQI are notable. First, while the general form of $EQI_{r,t}$ is a panel format, it is possible to construct a cross-sectional distribution of quality at a moment in time (i.e., EQI_{r,t_0}) to facilitate analyses such as spatial mapping. Second, the level of geographical aggregation is arbitrary: while the discussion of a “region” may connote a large geographic area, it is possible to calculate EQI at the level of a city, ZIP code, or even individual addresses. Finally, we can extend EQI in order to study an arbitrary grouping of firms (i.e., we do not need to select exclusively on geographic boundaries). For example, we can examine start-ups whose founder share a common demographic characteristic (e.g., sex), or firms that undertake a specific strategic action (e.g., engage in crowdfunding).

entrepreneurship in a region and the number of firms in such region (a measure of quantity). To do so, we define *RECPI* as simply $EQI_{r,t}$ multiplied by the number of firms in that region-year:

$$RECPI_{r,t} = EQI_{r,t} \times N_{r,t} \quad (3)$$

Since our index multiplies the *average* probability of a firm in a region-year to achieve growth (quality) by the number of firms, it is, by definition, the expected number of growth events from a region-year given the start-up characteristics of a cohort at birth. This measure of course abstracts away from the ability of a region to realize the performance of start-ups founded within a given cohort (i.e., its ecosystem performance), and instead can be interpreted as a measure of the “potential” of a region given the “intrinsic” quality of firms at birth, which can then be affected by the impact of the entrepreneurial ecosystem, or shocks to the economy and the cohort between the time of founding and a growth outcome.

The Regional Ecosystem Acceleration Index (REAI). While *RECPI* estimates the *expected* number of growth events for a given group of firms, over time we can observe the *realized* number of growth events from that cohort. This difference can be interpreted as the relative ability of firms within a given region to grow, conditional on their initial entrepreneurial quality. Variation in ecosystem performance could result from differences across regional ecosystems in their ability to nurture the growth of start-up firms, or changes over time due to financing cycles or economic conditions. We define *REAI* as the ratio of realized growth events to expected growth events:

$$REAI_{r,t} = \frac{\sum g_{i,r,t}}{RECPI_{r,t}} \quad (4)$$

A value of *REAI* above one indicates a region-cohort that realizes a greater than expected number of growth events (and a value below one indicates under-performance relative to expectations). *REAI* is a measure of a regional performance premium: the rate at which the regional business ecosystem supports high potential firms in the process of becoming growth firms.

Together, SFR, EQI, RECPI, and REAI offer researchers and regional stakeholders the ability to undertake detailed evaluations (over time, and at different levels of geographic and sectorial granularity) of entrepreneurial ecosystem performance.

III. Data and Estimation The foundational data source for the SCP are state-level business registration records, a potentially rich and systematic data set for the study of entrepreneurship. Business registration records are public records created endogenously when an individual registers a new business as a corporation, LLC or partnership. Our data covers 49 states and Washington, D.C. from 1988-2014, and 46 states and Washington, D.C. from 2014-2016.⁴ While it is possible to found a new business without business registration (e.g., a sole proprietorship), the benefits of registration are substantial, and include limited liability, various tax benefits, the ability to issue and trade ownership shares, and credibility with potential customers. Furthermore, all corporations, partnerships, and limited liability companies must register with a Secretary of State (or Secretary of the Commonwealth) in order to take advantage of these benefits: the act of *registering* the firm triggers the legal creation of the company. As such, these records reflect the population of businesses that take a form that is a practical prerequisite for growth.⁵ Concretely, our analysis draws on the complete population of firms satisfying one of the following conditions: (a) a for-profit firm in the local jurisdiction or (b) a for-profit firm whose jurisdiction is in Delaware but whose principal office address is in the local state. In other words, our analysis excludes non-profit organizations as well as companies whose primary location is not in the state. The resulting

⁴ These are all US states except for Delaware from 1988-2014 and all US States except for Delaware, Illinois, South Carolina and Michigan from 2014-2016 (these three states significantly increased the fees and/or administrative burden with using state-level registration data for the most recent years)

⁵ This section draws on Guzman and Stern (2015, 2017, 2020), where we introduce the use of business registration records in the context of entrepreneurial quality estimation. Please also see data appendices in those earlier papers.

dataset contains 38,506,776 observations.⁶ For each observation we construct variables related to: (a) a growth outcome for each start-up; (b) start-up characteristics based on business registration observables; and (c) start-up characteristics based on external observables that can be linked directly to the start-up. We briefly review each one in turn. We provide a more detailed summary relating to each observable in our data appendix. *Growth*. The growth outcome utilized in the SCP, *Growth*, is a dummy variable equal to 1 if the start-up achieves an initial public offering (IPO) or is acquired at a meaningful positive valuation within 6 years of registration⁷, as reported in Thomson Reuters SDC database.⁸ During the period of 1988 to 2010, we identify 17,453 firms that achieve growth, representing 0.07% of the total sample of firms in that period. *Start-Up Characteristics*. At the center of our analysis is an empirical approach to map growth outcomes to observable characteristics of start-ups at or near the time of business registration. We develop two types of measures of start-up characteristics: (a) measures based on business registration data observable in the registration record itself, and (b) measures based on external indicators of start-up quality that are observable at or near the time of business registration. *Measures Based on Business Registration Observables*. We construct twelve measures based on information observable in business registration records. We first create two binary measures that relate to how

⁶ The number of firms founded in our sample is substantially higher than the US Census Longitudinal Business Database (LBD), done from tax records. For example, for Massachusetts in the period 2003-2012, the LBD records an average of 9,450 new firms per year and we record an average of 24,066 firm registrations. We have yet to explore the reasons for this difference. However, we expect that it may be explained, in part by: (i) partnerships and LLCs that do not have income during the year do not file a tax returns and are thus not included in the LBD, and (ii) firms that have zero employees and thus are not included in the LBD.

⁷ In the Data Appendix to Guzman, Stern, 2020 (Section III, Table A4), we investigate changes in this measure both in the threshold of growth (e.g. only IPOs) as well as the time to grow, all results are robust to these variations

⁸ Although the coverage of IPOs is likely to be nearly comprehensive, the SDC data set excludes some acquisitions. SDC captures their list of acquisitions by using over 200 news sources, SEC filings, trade publications, wires, and proprietary sources of investment banks, law firms, and other advisors (Churchwell, 2016). Barnes, Harp, and Oler (2014) compare the quality of the SDC data to acquisitions by public firms and find a 95% accuracy; Nette, Stegemoller, and Wintoki (2011), perform a similar review. While we know this data not to be perfect, we believe it to have relatively good coverage of 'high value' acquisitions. Further, none of the cited studies found significant false positives, suggesting that the only effect of the acquisitions we do not track will be simply an attenuation of our estimated coefficients.

the firm is registered, *Corporation*, whether the firm is a corporation rather than an LLC or partnership, and *Delaware Jurisdiction*, whether the firm is registered in Delaware. We then create two additional measures based directly on the name of the firm. *Eponymy* is equal to 1 if the first, middle, or last name of the top managers is part of the name of the firm itself.⁹ We hypothesize that eponymous firms are likely to be associated with lower entrepreneurial quality. Our second measure relates to the structure of the firm name. Based on our review of naming patterns of growth-oriented start-ups versus the full business registration database, a striking feature of growth-oriented firms is that the vast majority of their names are at most two words (plus perhaps one additional word to capture organizational form (e.g., “Inc.”)). We define *Short Name* to be equal to one if the entire firm name has three or less words, and zero otherwise.¹⁰

We then create several measures based on how the firm name reflects the industry or sector within which the firm is operating, taking advantage of the industry categorization of the US Cluster Mapping Project (“US CMP”) (Delgado, Porter, and Stern, 2016) and a text analysis approach. We develop eight such measures. The first three are associated with broad industry sectors and include whether a firm can be identified as local (*Local*), or traded (*Traded*), or traded within resource intensive industries (*Traded Resource Intensive*). The other five industry groups are narrowly defined high technology industries that could be expected to have high growth, including whether the firm is associated with biotechnology (*Biotech Sector*), e-commerce (*E-Commerce*), other information technology (*IT Sector*), medical devices (*Medical Dev. Sector*) or semiconductors (*Semiconductor Sector*).

⁹ Belenzon et al (2014; 2017), perform a more detailed analysis of the interaction between eponymy and firm performance, highlighting name as a signal chosen by entrepreneurs given differences in growth intention.

¹⁰ Companies such as Akamai or Biogen have sharp and distinctive names, whereas more traditional businesses often have long and descriptive names (e.g., “New England Commercial Realty Advisors, Inc.”).

Measures based on External Observables. We construct two measures related to start-up quality based on intellectual property data sources from the U.S. Patent and Trademark Office. *Patent* is equal to 1 if a firm holds a patent application within the first year and 0 otherwise. We include patents that are filed by the firm within the first year of registration and patents that are assigned to the firm within the first year from another entity (e.g., an inventor or another firm). Our second measure, *Trademark*, is equal to 1 if a firm applies for a trademark within the first year of registration.

Table 1 group these measures in five categories: outcome variables, name-based observables, intellectual property observables and industry measures (US CMP Clusters and US CMP High-Tech Clusters), and reports the summary statistics and sources for these measures. For a more detailed description of all variables as well as the specific set of US CMP clusters used to develop each industry classification, please see Table A1 and Guzman and Stern, 2019 Appendix C.

III.A Entrepreneurial Quality Models.

We use this data to estimate two alternative logit regression models that allow one to examine how the presence or absence of a startup characteristic correlates with the probability of growth: the ‘academic model’, which includes all measures, and the policy model which exclusively utilizes jurisdiction (i.e., *Delaware*), legal form (*Corporation*), and intellectual property measures (*Patent* and *Trademark*) only.

Table 2 reports our results for the academic model for all registered firms in the dataset between 1988 and 2010. The results are striking. We find an extremely strong (and robust) correlation between startup characteristics and the probability of growth. Substantial changes in the predicted likelihood of a growth outcome are associated with characteristics observable at

founding from business registration records as well as characteristics observable with a lag (e.g., patent and trademark applications). On the one hand, startups founded as corporations are 220% *more* likely to grow. Similarly, firms with short names are close to 80% *more* likely to grow. On the other, eponymous firms are roughly 70% *less* likely to achieve an equity growth outcome. Startups that apply for a patent or trademark in their first year after founding are 328% and over 1900%, respectively, more likely to achieve an equity growth outcome within 6 years of founding.¹¹ Moreover, these changes in predicted probabilities are multiplicative in nature: a startup that registers in Delaware *and* applies for a patent in its first year is over 83 times more likely to grow than a firm that only registers in its home state and does not apply for intellectual property protection.^{12,13}

Not surprisingly, findings from the policy model are comparable. As reported in Table 3, forming as a corporation, registering in Delaware, and filing for a patent or trademark within the first year are correlated with increases in the likelihood of a growth outcome of 100%, 2,449%, 1700% and 43%, respectively. And, like the academic model, the predictive power of these startup characteristics is multiplicative in nature. A startup that is registered in Delaware and files for both a patent and trademark within its first year is 950 times more likely to achieve a growth outcome than one that does not.

¹¹ Preliminary models which center on one or two startup characteristics find similar, albeit slightly higher correlations. Corporations and firms registered in Delaware are 283% and 2,261% more likely to achieve a growth outcome, respectively. Firms with short names are 93% more likely to grow, while eponymous firms are close to 80% less likely to achieve a growth outcome. Those startups that apply for a patent or trademark in their first year are 4,327% and 737% more likely to achieve an equity growth outcome, respectively.

¹² It is very important to emphasize that these startup characteristics are not the *causal* drivers of growth, but instead are “digital signatures” that allow us to distinguish firms in terms of their entrepreneurial quality. Registering in Delaware or filing for a patent will not guarantee a growth outcome for a new business, but the firms that historically have engaged in those activities have been associated with skewed growth outcomes.

¹³ The precision allowed by our definition of quality comes nonetheless at a cost. Our definition does not allow us to include all the richness of social outcomes through which companies help communities or individuals. In principle, however, a richer version of our approach that includes multiple outcomes and a larger number of observables might be able to achieve this result.

Robustness and Predictive Quality. In Figure 1, we evaluate the predictive quality of the academic model estimates by undertaking a tenfold cross-validation test (Witten and Frank, 2005),¹⁴ and report the out-of-sample share of realized growth outcomes at different portions of the entrepreneurial quality distribution. The results are striking. On average, 65% of all growth firms are included within the top 10% of our estimated growth probability distribution. 56% and 37% of all growth outcomes are included within the top 5% and 1% of the estimated entrepreneurial quality distribution, respectively. Growth, however, is still a relatively rare event even among those startups with the highest estimated entrepreneurial quality: the average firm within the top 1% of estimated entrepreneurial quality distribution has only a 2.5% chance of realizing a growth outcome. Figure A1 demonstrates that estimates generated by the Policy model are similarly robust in terms of predictive quality. The top 10% of the distribution of estimated entrepreneurial quality includes 56% of all growth firms.

Entrepreneurial Ecosystem Statistics. We then use our measures of estimated quality to develop economic indices that simultaneously account for both the quantity and the quality of entrepreneurship (and which are outlined in the empirical framework section):

- SFR—the Startup Formation Rate—the quantity of new business registrants within a given population.
- EQI—the Entrepreneurial Quality Index—the *average* growth potential (or “quality”) of any given group of new firms.

¹⁴ Specifically, we divide our sample into 10 random subsamples, using the first subsample as a testing sample and use the other 9 to train the model. For the retained test sample, we compare realized performance with entrepreneurial quality estimates from the model resulting from the 9 training samples. We then repeat this process 9 additional times, using each subsample as the test sample exactly once. This approach allows us to estimate average out of sample performance, as well as the distribution of out of sample test statistics for our model specification.

- RECPI—the Regional Entrepreneurship Cohort Potential Index—the number of startups within a particular location or region expected to later achieve a significant growth outcome.
- REAI—the Regional Entrepreneurship Acceleration Index—the ability of a region to convert entrepreneurial potential into realized growth.

Each index calculates a different quantitative measure of entrepreneurship that can be aggregated and systematically compared across entrepreneurial ecosystems. The EQI, RECPI, and REAI indexes offer quantitative measures that incorporate the *quality* of entrepreneurship. Each gives a better indication than possible under traditional methods about how skewed the distributions of growth potential and likely growth outcomes are (and whether and to what extent a greater number of small to medium-sized businesses could be expected to catalyze the same growth outcomes as a high-potential growth firm).¹⁵ Additionally, REAI systematically quantifies the ratio of *realized* to *expected* growth events for a given cohort of new firms, providing an indication of whether the ecosystem in which a cohort of new firms is located is conducive to growth (or not). As such, these indexes offer policymakers and stakeholders a better view of whether and to what extent their regions are generating startups with high-growth potential and to what extent they are helping or hampering these firms' efforts to realize their potential after founding.

¹⁵ The level of skewness of entrepreneurial quality is highly informative. It indicates how much more likely a startup at the high end of the entrepreneurial quality distribution is to grow than an average firm. If skewness were low, then adding several average firms could have as much regional impact as one high-growth-potential firm. But, if skewness is high (as the findings indicate), then a much larger number of firms with average growth potential is needed to generate the expected impact of one high-potential firm. Given the level of skewness observed, almost 4,000 local limited liability companies (average firm) are needed to generate the same potential as only one new Delaware corporation with an early patent and trademark. Put another way, initial ambition/potential for growth is a key dimension of heterogeneity across new firms. The subset of high-potential-growth startups is very small and fundamentally different than the vast majority of new firms

Aggregating Across Locations. Finally, we aggregate our estimates for four levels of locations—national, state, MSA, county and ZIP Code. For national and state level indexes, we aggregate all firms in our sample in each year from 1988 to 2016, while for county, MSA, and ZIP Code level indexes, we use all firms that have valid ZIP Code information to form the aggregation.¹⁶ Rather than changing the MSA definitions through time, we stick to the 2013 MSA definitions for consistency in our time-series.

Publicly Available Datasets. We make publicly available datasets aggregated at the national, state, MSA, county and ZIP Code level, which can be downloaded in tab-delimited text files, Stata compatible files, and R compatible files. These can be found on the Startup Cartography Project Harvard Dataverse and through the Startup Cartography Project website (<http://www.startupcartography.com>). In each, we provide SFR, EQI, RECPI and REAI estimates by year, state-year, MSA-year, county-year, and ZIP Code – year for 49 states (all except Delaware and Washington D.C.) from 1988-2014, and 46 states (all except Delaware, Michigan, Illinois, South Carolina and Washington D.C.) through 2016. The *Entrepreneurship_National.tab* includes 6 variables and 29 observations (one per year). The *Entrepreneurship_by_State.tab* includes 7 variables and 1,444 observations. The *Entrepreneurship_by_MSA.tab* includes 8 variables and 10,362 observations. The *Entrepreneurship_by_County.tab* includes 12 variables and 85,628 observations. The *Entrepreneurship_by_ZIP_Code.tab* includes 8 variables and 802,854 observations. While the U.S. Startup Map (discussed in Section VI) provides visualization down

¹⁶ Specifically for the county level index, we use the registered ZIP Code of each company to identify each county using the HUD USPS ZIP Code crosswalk files. We then aggregate across each county and year to estimate EQI, RECPI and REAI based on the observed outcomes in each one.

to the level of individual street addresses, confidentiality and data release restrictions do not allow us to release the data at the level of individual firms.

IV. SCP Ecosystem Statistics: Descriptive Findings

The calculation of SFR, EQI, RECPI and REAI at different levels of geographic agglomeration and across time enables researchers and policy makers to evaluate different entrepreneurial ecosystems and regional trends in ecosystem statistics. Table 4 reports summary statistics by region at the state, MSA, county and ZIP Code level. Across all region-years, there were on average 23,234 startups formed per state per year, 2,948 per MSA per year, 336 per county per year and 41 startups per ZIP Code per year, respectively. On average, the growth potential of an average start-up (or EQI) was low, with the probability of a growth outcome ranging from 1 in 1818 at the ZIP Code level to 1 in 1492 at the MSA level. The average expected number of growth outcomes (RECPI) ranged from .23 at the county level to 15.2 at the state level. RECPI correlated closely with the actual number of growth outcomes later observed (which ranged from .21 at the county level to 13.73 at the state level).

In Figures 2 and 3, we compare the estimated average SFR (entrepreneurial quantity) per capita and EQI (entrepreneurial quality) by MSA, to average GDP growth rate with a 5- year-lag. In Figure 2, we find that entrepreneurial *quantity* per capita and entrepreneurial *quality* are not highly correlated at the higher range of their respective distributions. We observe both a low SFR per capita and EQI for most MSAs (e.g., both a low number of new registered businesses and low estimated probability of achieving a growth outcome). The highest average EQI observed is in the Silicon Valley region, but the number of startups formed per capita is in the lower range of observations. Boulder, Missoula and Miami boast among the highest range of entrepreneurial quantity (on a per capita basis), but lowest in quality. In Figure 3A, we find a weak, but slightly

negative correlation between SFR per capita and GDP growth by MSA, when fitted by Real GDP of MSAs. On the other hand, as shown in Figure 3B, the correlation between EQI and GDP growth is significantly positive. We observe MSA with a higher EQI also have higher average GDP growth in the 5 year period following the firm founding.¹⁷ In comparison with Figure 3A, we can see that the GDP growth rate is more correlated with the quality, than the quantity, of regional entrepreneurship.

Table 5 shows the ranking of average RECPI (quality-adjusted quantity) per capita and EQI (entrepreneurial quality) between 2010 and 2014 by MSA. We first select the top 100 MSAs in average population between 2010 and 2014. From that group, we select the top and bottom 30 MSAs, respectively. We next rank the average RECPI and REAI of these two groups. We can see several of these MSAs perform well in both measures. Specifically, the San Francisco Bay Area ranks 1st in average RECPI per capita as well as average quality. MSAs of San Francisco, Bridgeport, Boston, Los Angeles all sit in the top 10 of these rankings. MSA of Miami ranks 3rd in average RECPI per capita, but falls out of top 30 in the ranking of average quality ranking, this also happens in several Florida MSAs, which ranks higher in RECPI per capita, but not much in the latter. Besides, we can also find MSAs of New York, Chicago, Dallas, San Diego, Madison, Durham, Raleigh, Austin, Columbia, Oxnard, Tulsa in both top lists. On the other hand, we can also find similar pattern in the bottom lists of these two. MSAs of Augusta and McAllen are both within the bottom 10 MSAs in terms of average RECPI per capita and average quality. Similarly, 11 other MSAs (e.g. Albuquerque, Dayton, Tucson, etc.) fall at the lower end of both lists of these

¹⁷ Specifically, this figure considers a panel data where, for each observation, we include the average quality of firms founded in the region-year and the GDP growth over the five subsequent years. We only include years since 2001 because MSA GDP estimates are not available before and stop in 2013 to be able to observe GDP growth. The plotted point in the scatterplot is the average of each variable for all years in our data.

entrepreneurial measures. Interestingly, Charleston-North Charleston ranks 28th in the top average RECPI per capita, but ranks 27th in the bottom average quality list.

In Figure 4, we select six high-achieving MSAs in terms of EQI and find striking differences in MSA entrepreneurial quality across MSAs and across time from 1988 to 2014. The movement of the line indicates the level of entrepreneurial quality for the region, while the line thickness indicates the average real MSA GDP within these years. Interestingly, these six MSAs all begin at similar levels in 1988 and follow a similar trajectory—starting to grow in the early 1990s and then peaking in 2000. The San Francisco Bay Area MSAs have consistent high EQI compared to other regions through years. The San Jose-Sunnyvale-Santa Clara, CA MSA has the highest entrepreneurial quality every year and nearly double its 1988’s level in 2014, while the San Francisco-Oakland-Fremont, CA MSA starts from a second lowest EQI in 1988 and gradually grows to the second highest MSA of EQI. The Boston-Cambridge-Quincy, MA-NH MSA, however, starts at a second-highest EQI level in 1988 but is superseded by San Francisco after 2003. It has a level at basically half of San Francisco’s EQI by 2014. On the other hand, the New York-Northern New Jersey-Long Island, NY-NJ-PA MSA—which has the largest average real MSA GDP of all, shares a similar level of EQI with the Austin-Round Rock, TX MSA both in 1988 and 2014. The EQI of Austin MSA reaches the highest level in 2007 before plummets to New York’s level of EQI since 2008. Among these six MSAs, the area of Miami-Fort Lauderdale-Pompano Beach, FL, with the second highest business registration records, has the lowest EQI every year across these regions.

Trends in the Effect of the US Entrepreneurial Ecosystem (REAI). Regional performance depends on not simply founding new high potential enterprises, but also scaling of those enterprises to generate employment and economic activity. In Figure 5, we assess the performance

of the US ecosystem by plotting REAI for our time-period to examine “ecosystem” performance of the United States during our time period, using predicted values for the years 2011 to 2014.¹⁸ We estimate confidence intervals by repeating our procedure on thirty bootstrap samples, and also including the maximum and minimum of each value in the graph. REAI captures the relative ability of a given start-up cohort to realize its potential, relative to the expectation for growth events as measured by RECPI. A value above 1 indicates a positive ecosystem effect, and a value under 1 indicates a negative effect. In contrast to RECPI, this index reflects the impact of the economic and entrepreneurial environment in which a start-up cohort participates (i.e., the “ecosystem” in which it participates).

Three distinct periods stand out in this graph. The early portion of our sample saw a significant increase in REAI from a slight negative level to a peak of 1.66 for the 1995 cohort. Relative to the rest of the years that we observe, startups born in 1995 were 66% more likely to achieve an equity growth outcome conditional on their estimated quality. This peak was followed by a steady decline of REAI over the subsequent decade, and the index turns negative in the year 2001. It continues this negative decline from 2001 to 2007, with REAI moving from 0.89 down to 0.63. Putting these values together creates meaningful differences for startup cohorts: a start-up at a given estimated quality level was 3 times more likely to experience a growth event if it was founded in 1995 rather than in 2007. Our index then begins to recover after 2007, and turns once again positive in 2011. Though these last few years are only preliminary estimates due to the natural time-lags inherent in observing startup growth, it would appear that there is a significant

¹⁸ Because our approach requires that we observe the *realized* growth firms we can only measure our index with a 6 year lag, thus, up to 2010. For years 2011 to 2014, we estimate our model with a varying lag of $n = 2016 - year$ and calculate RECPI using such lag.

increase in the performance of the US entrepreneurial ecosystem, reaching a level higher than all prior estimates by 2014.

V. The U.S. Startup Map

The U.S. Startup Map is an interactive visualization and map of SFR and EQI from 1988-2016 (in states and years where data is available). The map enables users to geographically explore SCP data and analysis in a web browser with familiar mouse click and touch gestures. It broadens the impact of the Startup Cartography measures by allowing users to better see the growth potential of entrepreneurship in their ecosystem. As mentioned earlier, based on feedback from policy users, the U.S. Startup Map uses the policy model as its basis for assessment of any given location, in order to focus on a consistent and understandable set of digital markers of start-up quality.¹⁹

Assigning colors by these measures does not take advantage of all available startup characteristics leveraged in our academic model. For example, founder-firm eponymy does not impact color assignment. This is a deliberate choice to make the color palette, and the map, more accessible to all stakeholders. Earlier map iterations assigned colors according to a more complex algorithm. This inflicted a burden on users to understand the algorithm before using the map. We found this burden to be counter-productive to the map's goals.

Entrepreneurial *quantity*—SFR—is a necessary foundation for the map. But it is not sufficient. Alone, entrepreneurial *quantity* produces an image that closely matches a general population map. For the map to bring the Startup Cartography Project data to life it must also show entrepreneurial *quality*. Panel A of Figure 6 presents a national view of the U.S. Startup Map. The

¹⁹ It is useful to note that the use of the Policy Model means that we are not using all the information available in our data. For example, founder-firm eponymy does not impact the color assigned on the map. This is a deliberate choice to make the color palette, and the map, more accessible to all stakeholders.

larger section on the right displays the map. Users can explore the map with zoom, pan, search, and select interactions. To the map's left is the legend section. The legend introduces the map and contains dynamic controls. These controls include a timeline scrollbar filter and an option to add various contextual data to the map. Only a single year is displayed in a single view.

The map places new business registrations at the location associated with their registration. New businesses are represented as circles. Quantity of registrations is visualized with circle (or *bubble*) size. Larger bubbles represent greater numbers of newly registered firms. The color of the circle corresponds to the quality percentile of new business registration(s) at that location. There is a direct correlation between the number of new businesses and the number of pixels in the displayed circle (i.e. circle area).

The U.S. Startup Map zoom level determines how registrations aggregate into bubbles. Panel B of Figure 6 shows the four map zoom levels: State, Metro, City and Address. Zoomed all the way out, the contiguous United States are seen and registrations aggregate at the state level. Zoom in and businesses aggregate into metropolitan statistical areas. Zoom in further and businesses aggregate into cities. Zoom all the way in, to a neighborhood level, and businesses aggregate at individual addresses. At the neighborhood level each bubble represents an individual address. A larger address bubble often represents a large building where many businesses were registered during a given year.

Selecting a jurisdiction circle reveals a pop-up *tooltip* that lists the jurisdiction (state, metro, or city), quality percentile, quantity of new business registrations, and year displayed on the map. Address-level tooltips are not displayed. Panel C of Figure 6 provides an example of tooltip detail for Palo Alto California.

The color palette for the map is *grouped* into two quality classes. Blues are associated with high growth potential entrepreneurship. Oranges are associated with local entrepreneurship, businesses not expected to experience a growth outcome.

Each individual business registration's specific color is determined by the presence of specific or multiple impactful measures included in our policy model: LLC (pale orange) or corporation (orange), Delaware registration (pale blue), patent or trademark (blue), and a combination of at least two high quality measures (dark blue). Figure A2 shows the semantic color table for the map.

As shown in Figure A2, the map legend displays colors on a percentile scale (below), sizing color buckets according to the distribution of registrations. Color buckets are not regularly spaced across the percentile spectrum (e.g. with breakpoints at 25, 50, and 75%) because the actual portion of impactful measures is not regular. For example, the pale orange color that corresponds to the lowest 56% of the palette directly represents the observed 56% of new business registrations that are LLCs (with no other impactful measure).

The color palette breakpoints were determined at the individual registration level. These same numerical breakpoints are extended across all aggregation levels (city, metro, state). The direct association with impactful measures (Patent, Trademark, etc.), however, does not similarly extend. Each aggregate bubble is colored by its quality percentile, relative to the rest of the nation. A pale orange city represents a city somewhere in the lower 56% of all cities, not a city only composed of LLCs.

Aggregating data, in this map's case into jurisdiction bubbles, is a necessary way of viewing a field as large as the United States. However, summary aggregation poses a risk of

missing significant outliers. In this case we are concerned that an interesting cohort of high quality businesses might go undetected, lost in a large city.

A new bubble design addresses this concern. It employs nested rings to reveal the entrepreneurial composition of individual cities. Each registration is now represented by its associated color. This ring design closes the semantic gap mentioned above between aggregate color assignments and impactful measures. Figure 7 compares city ring views of Silicon Valley and Phoenix, both at the same zoom level.²⁰

VI. Conclusion

This paper presents the Startup Cartography Project (SCP), which offers a new set of entrepreneurial ecosystem statistics (SFR, EQI, RECPI and REAI) for the United States from 1988-2016. The SCP includes both a public-access dataset at the state, MSA, county, and zip code level, as well as an interactive map, the U.S. Startup Map, that permits academic and policy users to assess entrepreneurial ecosystems at an arbitrary level of granularity (from the level of states down to individual street addresses). The SCP's consistent, transparent and accessible data and visualization inform debate around the design and structure of policies intended to promote regional "entrepreneurial ecosystems" (Feldman and Francis, 2004; Lerner, 2009; Audrestch and Lehmann, 2005; Stam, 2015) by addressing the issues that make systematic measurement challenging and enabling evaluation on a granular (as well as aggregated) and timely (as well as retrospective) basis. By estimating the growth potential (or entrepreneurial quality) of startups at

²⁰ Real world testing was an important part of the map development process. We used a design thinking approach where we developed solutions for visualization through observation of users and product iteration. As cartographers familiar with the data, it is important not be blinded by our own knowledge. Real users showed us what was confusing and which aspects of the map did not work as expected. For example, the city rings now a focal point of our design were arrived at through observing stakeholders interacting with and questioning the map.

or near the time of founding, SCP indexes provide a view of the skew of entrepreneurship most correlated with later regional economic growth. SCP indexes enable the assessment of entrepreneurial potential prior to the emergence of outcomes through predictive analytics (and the study of impact without selecting on desired outcomes). They permit evaluation of entrepreneurial ecosystems at multiple levels of geographic analysis, empowering academics and policymakers to consider the power of place in novel ways. Taken together, our quality-oriented approach can yield significant insights in both research and policy, and has set the stage for a more nuanced and comprehensive understanding of entrepreneurial ecosystems and the role that entrepreneurship policy plays in economic development and regional resilience.

Research insights. Relative to quantity and outcome-based measures, research leveraging the SCP's quality-oriented approach has already offered significant insight into both entrepreneurship and entrepreneurial ecosystems at the macro, regional and firm level. At a macro-level, Guzman and Stern (2020) document that the skew of high-growth potential startups is sensitive to economic conditions and capital markets, including the rate of expansion of the economy, and not in a 30-year secular decline suffered by more local firms. Fazio et al (2018) further emphasize the key importance and variation of the U.S. entrepreneurial ecosystem in allowing firms to scale. Consistent with other research on the size distribution of firms across economies (Hsieh and Klenow, 2014), these results indicate that it is not enough for the United States (or regional ecosystems for that matter) to produce high potential firms, the United States must also foster an environment that allows them to grow.

Moving to the regional level, our simple correlations of quality and GDP growth document the critical role that entrepreneurial quality (though not quantity) plays in predicting economic performance and presents a significant opportunity for follow-on work. Recent research has begun

to use SCP measures to study the impact of R&D tax policy on new firm formation (Fazio et al, 2019), the role of universities and other knowledge intensive institutions in shaping local entrepreneurship (Tartari and Stern, 2019), the returns to entrepreneurial migration (Guzman, 2018), and the consequences of regional entrepreneurship for local economic inequality (Marinoni and Voorheis, 2019). This research appears to us only the beginning of a rich avenue of further inquiry, which we hope our public dataset will supports in developing.

At a firm level, a quality-oriented approach lends itself to answering some of the main questions in entrepreneurship research by allowing progress in the separation of entrepreneurial quality from the process of selection. Existing work has focused on understanding how the process of selection shapes the gender gap in entrepreneurship (Guzman and Kacperczyk, 2019), and the benefits of investment by venture capitalists (Catalini et al, 2019).

Policy use. With our streamlined policy model and sharper focus on measures that more concretely differentiate between the startup formation of local and high-growth potential firms, the SCP provides stakeholders with a much clearer view of the potential and trajectory of startup formation in their respective ecosystems. Given the possibility that entrepreneurial quality is a leading indicator for other outcomes in regional performance, tracking EQI, for example, would allow government analysts to measure and support entrepreneurial quality, and to observe entrepreneurial dynamics in a more proactive and informed way. Not simply a tool for direct measurement, our methodology allows government organizations (e.g., the Small Business Administration) to design and evaluate interventions that focus on the quality of entrepreneurship rather than only increasing rates of firm formation, thus facilitating an approach that could potentially increase the impact of entrepreneurship interventions.

The U.S. Startup Map will assist policy makers in forming consensus with ecosystem stakeholders in evaluating and designing entrepreneurship initiatives. By dynamically visualizing entrepreneurial quality, the map “sets the table” for policy makers and other stakeholders to come together. Our hypothesis: if stakeholders can *see* the status of new firms founded over time, they, in turn, may more easily achieve a productive consensus on the state of their entrepreneurship ecosystem and its potential to fuel economic growth.

The U.S. Startup Map is valuable in many ways. It attracts attention as a salient introduction to the Project. Everyone arrives to the map with a wealth of geographic knowledge and is usually delighted to see how a new layer of information intersects with their ready understanding. It engages users at all levels with millions of data observations and sophisticated empirical analysis. There is simply something wonderful about playing with novel data on a map. On first encounter a new user is likely to investigate what the data look like around their hometown or current office. From there, user engagement often proceeds to visually testing little hypotheses, eager to see if reality matches expectations.

The U.S. Startup Map also serves as a crude (and intuitive) validator of data, displaying that the data exists and it is rich. Seeing tens of millions of business registrations dynamically snap into formation according to your interaction impresses the eye. Likewise, the map is a crude quality check on the data system. Unlike a lonely typo in a book, a single map bubble out of place (e.g. an address in the middle of a body of water) casts doubt upon the entire project.

At its best, the U.S. Startup Map acts as a central object of discussion between anyone engaged with the project. As a shared artifact, it provides a common ground for people with different models of understanding to come together, discuss, and imagine a better vision together. In this sense, the map is a coordination mechanism that fosters discussion about local and high growth firms. In

addition to facilitating discussion, the map also creates opportunity for insight discovery. Comparisons are possible crosstown and across regions. Patterns can be detected, especially across time as one scrolls through the years and sees the entrepreneurial activity of a location change. These insights are most powerful at the intersection of the map's display and the local knowledge of an interested stakeholder. Their special context and vested interest bring the map's data to life.

Looking forward. We believe the opportunities for the SCP and the U.S. Startup Map to help advance research and policy understand and improve entrepreneurship ecosystems are just emerging. Our approach highlights the significant potential of business registration records, a data source that has been used sparingly and only in an aggregated form by economists. We encourage further efforts by states to collect somewhat more granular information about the objectives of an enterprise (e.g., industry codes or founder addresses) in connection with business registration and to make business registration records more easily accessible. The lack of standardization and the uneven level and scope of digitization of business registration records across states remains a significant barrier to scaling business registration analysis across the entire United States. We look forward to further use of SCP measures and the U.S. Startup Map by researchers, policymakers and other stakeholders and the insights that work will bring in seeding and scaling entrepreneurship ecosystems and fostering the growth of regional economies.

References

- Aghion, Philippe and Peter Howitt. (1992) "A Model of Growth Through Creative Destruction" *Econometrica*, 60(2): 323–351.
- Amoros, Jose E., and Neils Bosma. (2014) "Global Entrepreneurship Monitor: 2013 Executive Report" London, GB: London Business School, and Wellesley, MA: Babson College.
- Arzaghi, Mohammad, and J. Vernon Henderson. (2008) "Networking of Madison Avenue" *Review of Economic Studies*. 75 (4): 1011-1038
- Audretsch, David B & Maryann P. Feldman. (1996) "R&D Spillovers and the Geography of Innovation and Production," *American Economic Review*. 86(3): 630-640.
- Audretsch, D. B. & Lehmann, E. E .2005. "Does the knowledge spillover theory of entrepreneurship hold for regions?" *Research Policy*, 34(8), pp. 1191–1202. doi:10.1016/j.respol.2005.03.012
- Balasubramanian, Natarajan, and Jagadeesh Sivadasan. (2009) "NBER Patent Data-BR Bridge: User Guide and Technical Documentation". Working Paper.
- Barnes, Beau, Harp, Nancy, Oler, Derek. (2014) "Evaluating the SDC Mergers and Acquisitions Database" *SSRN Working Paper #2201743*
- Belenzon, Sharon, Chatterji, Aaron and Brendan Daley. (2014) "Eponymous Entrepreneurs" *Working Paper*
- Carlino, Geraldo, Catterjee, Satyajit, and Robert Hunt. (2007) "Urban Density and the Rate of Invention". *Journal of Urban Economics*. 61 (3): 389-419
- Davis, Steven, and John Haltiwanger. (1992) "Gross Job Creation, Gross Job Destruction, and Employment Reallocation". *The Quarterly Journal of Economics*. 107 (3): 819-862
- Decker, Ryan, Haltiwanger, John, Jarmin, Ron and Javier Miranda. (2014) "The Role of Entrepreneurship in US Job Creation and Economic Dynamism". *Journal of Economic Perspectives*. 28(3): 3-24`
- Delgado, Mercedes, Porter, Michael, and Scott Stern. (2016) "Defining Clusters in Related Industries." *Journal of Economic Geography*. 16 (1): 1-38.
- EDA. 2010. "Regional Innovation Strategies". Economic Development Agency. Available at: <https://www.eda.gov/oie/ris/> (Accessed on January 21, 2020)
- Fairlie, Robert W.. (2014) "Kaufman Index of Entrepreneurial Activity: 1996-2013." *Ewing Marion Kaufman Foundation*, Kansas City, MS.
- Fazio, Catherine, Jorge Guzman and Scott Stern. (2019). "The Impact of State-Level R&D Tax Credits on the Quantity and Quality of Entrepreneurship". NBER Working Paper #26099

- Feld, Brad. 2012. *Startup Communities: Building an Entrepreneurial Ecosystem in Your City*. Wiley. pp. 224.
- Feldman, Maryann, and David Audrestch. (1999) “Innovation in cities: Science-based diversity, specialization and localized competition”. *European Economic Review*. 43(2): 409-429.
- Feldman, Maryann. 2001. “The Entrepreneurial Event Revisited: Firm Formation in a Regional Context”. *Industrial and Corporate Change*. 10 (4): 861-891
- Feldman, Maryann, and Johanna Francis. 2004. “Homegrown solutions: Fostering cluster formation.” *Economic Development Quarterly*. 18 (2): 127-137.
- Feldman, Maryann, Johanna Francis, and Janet Bercovitz. 2005. “Creating a cluster while building a firm: Entrepreneurs and the formation of industrial clusters.” *Regional Studies*. 39 (1): 129-141.
- Furman, Jeffrey, Porter, Michael and Scott Stern. (2002) “The determinants of national innovative capacity”. *Research Policy*. 31:899-933
- Graham, Stuart, Hancock, Galen, Marco, Alan, and Amanda Fila Myers. The USPTO Case Files Dataset: Descriptions, Lessons and Insights. United States Patent and Trademark Office. SSRN Working Paper #2188621 (2013)
- Glaeser, Edward, Sari Pekkala Kerr, and William Kerr. 2015. “Entrepreneurship and Urban Growth: An Empirical Assessment with Historical Mines.” *The Review of Economics and Statistics*. MIT Press. 2 (97): 498-520.
- Gompers, Paul, Kovner, Anna, Lerner, Josh, and David Scharfstein. (2010) “Performance Persistence in Entrepreneurship.” *Journal of Financial Economics*. 96(1): 18-32.
- Guzman, Jorge, and Scott Stern. (2015) “Where is Silicon Valley?” *Science*. Vol. 347. Issue #6222.
- Hamilton, Barton. (2000) “Does Entrepreneurship Pay? An Empirical Analysis on the Returns to Self-Employment”. *Journal of Political Economy*. 108(3): 604-631.
- Hastie, Trevor, Tibshirani, Robert, and John Friedman. (2001) “The Elements of Statistical Learning”. Second Edition. Springer.
- Hathaway, Ian, and Robert Litan. (2014) “Declining Business Dynamism in the United States: A Look at States and Metros”. *Economic Studies at Brookings*. Brookings Institution.
- Hathaway, Ian, and Robert Litan. (2014) “Declining Business Dynamism: It’s For Real” *Comment*.
- Hurst, Erik, and Benjamin Pugsley. (2011) “What do Small Businesses Do?”. *Brookings Papers on Economic Activity*.
- Jaffe, Adam, Trajtenberg, Manuel, and Rebecca Henderson. (1993) “Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations.” *Quarterly Journal of Economics*. 108 (3): 577-598.

- Jarmin, Ron, and Javier Miranda. 2002. "The longitudinal business database." SSRN Working Paper #2128793.
- Jovanovic, Boyan. (1982) "Selection and the Evolution of Industry." *Econometrica*. 50(3): 649–670.
- Katz, Bruce, and Julie Wagner. (2014) "The Rise of Innovation Districts: A New Geography of Innovation in America". *Brookings Institution – Metropolitan Policy Program*.
- Kauffman Foundation. 2019. "2018 State Report on Early-stage Entrepreneurship". *Kauffman Foundation – Kauffman Indicators for Entrepreneurship*.
- Kauffman Foundation. 2019. "ESHIP Communities". Available at: <https://www.kauffman.org/what-we-do/entrepreneurship/entrepreneurial-communities> (Accessed on January 21, 2020)
- Kerr, William, Nanda, Ramana, and Matthew Rhodes-Kropf. (2014) "Entrepreneurship as Experimentation". *Journal of Economic Perspectives*. 28 (3): 25-48.
- Kerr, William, and Shihe Fu. (2008) "The Survey of Industrial R&D--Patent Database Link Project." *Journal of Technology Transfer*. 33 (2):173-186
- Kerr, William, and Scott Kominers. (2015) *Review of Economics and Statistics*. forthcoming
- Klapper, Leora, Amit, Raphael, Guillen, Mauro. (2010) "Entrepreneurship and Firm Formation across Countries". *NBER Volume: International Differences in Entrepreneurship*. Edited by Josh Lerner and Antoinette Schoar.
- Kuznets, Simon. (1941) "National Income and its Composition, 1919-1938. Volume I" p. 3.
- Lafontaine, Francine, and Kathryn Shaw. (2014) "Serial Entrepreneurship: Learning by Doing?" *NBER Working Paper #20312*
- Lerner, Josh. 2009. *Boulevard of Broken Dreams: Why Public Efforts to Boost Entrepreneurship and Venture Capital Have Failed—and What to Do about It*. Princeton University Press. p. 240.
- Levenshtein, V. I. (1965), "Binary codes capable of correcting deletions, insertions, and reversals.", *Doklady Akademii Nauk SSSR* 163 (4): 845–848
- Levine, Ross, and Yona Rubinstein. (2013) "Smart and Illicit: Who Becomes an Entrepreneur and Does it Pay?" *NBER Working Paper #19276*
- McFadden, Daniel. (1974) "Conditional logit analysis of qualitative choice behavior". *Frontiers in Econometrics. Chapter 4*. Academic Press. p. 105-142.
- Moskowitz, T., & Vissing-Jorgensen, A. (2002). "The Returns to Entrepreneurial Investment: A Private Equity Premium Puzzle?" *American Economic Review*. 92(4): 745-778.
- Murray, Fiona, and Scott Stern. "Linking and Leveraging." *Science*. 348 (6240): 1203

- Nanda, Ramana, and Matthew Rhodes-Kropf. (2014) "Financing Risk and Innovation". *HBS Working Paper* 11-013.
- Nanda, Ramana, and Matthew Rhodes-Kropf. (2013) "Investment Cycles and Startup Innovation." *Journal of Financial Economics* (110): 403–418.
- Pohlman, John, and Dennis Leitner. (2003) "A Comparison of Ordinary Least Squares and Logistic Regression." *The Ohio Journal of Science*. 103 (5): 118-125
- Reister, Shane. (2014) "Why Should we Actively Track and Measure Startup Communities." Available in: *Kauffman Thoughtbook 2015 – Entrepreneurship: New Directions for a New Era*. Retrieved from: <http://www.kauffman.org/thoughtbook2015/paths-to-entrepreneurship#startupcommunities> on December, 2014.
- Saxenian, AnnaLee. 1994. *Regional Advantage: Culture and Competition in Silicon Valley and Route 128*. Harvard University Press.
- Schoar, Antoinette. (2010) "The Divide Between Subsistence and Transformational Entrepreneurship". *Innovation Policy and the Economy, Volume 10*. Edited by Josh Lerner and Scott Stern.
- Schumpeter, Joseph A. (1942) "Capitalism, Socialism, and Democracy". New York: Harper.
- Stam, Erik. 2015. "Entrepreneurial Ecosystems and Regional Policy: A Sympathetic Critique." *European Planning Studies*. 23 (15): 1759-1769.
- Stangler, Dane, and Bell-Materson. 2015. "Measuring Entrepreneurial Ecosystems". Report.
- Taddy, Mathew. (2013) "Big Data Analysis". *Lecture on NBER Summer Institute: Econometric Methods for High-Dimensional Data*.

TABLE 1
Summary Statistics

Measure	Source	Description	Mean	Std. Dev.
<i>Outcome Variables</i>				
Equity Growth (IPO or Acquisition)	SDC Platinum IPO and M&A.	1 if firm has an equity growth event in the first 6 years.	0.0007	0.026
<i>Corporate Form Observables</i>				
Corporation	Business Reg.	1 if a firm is a corporation (not an LLC or partnership)	0.443	0.497
Delaware	Business Reg.	1 if the firm's jurisdiction is Delaware	0.025	0.157
<i>Name-Based Observables</i>				
Short Name	Business Reg.	1 if the firm's name length is 3 words or less (including firm type (e.g. "inc."))	0.467	0.499
Eponymous	Business Reg.	1 if the firm's name includes the president or CEO first or last name.	0.087	0.282
<i>Intellectual Property Observables</i>				
Patent	USPTO	1 if the firm acquires for a patent application within 1 year of founding.	0.0020	0.045
Trademark	USPTO	1 if the firm acquires for a trademark within 1 year of founding.	0.0016	0.040
<i>Industry Measures (US CMP Clusters)</i>				
Local Industry	Estimated from name	If firm name is associated to a local industry.	0.194	0.395
Traded	Estimated from name	If firm name is associated to a traded industry.	0.538	0.499
Resource Intensive Industry	Estimated from name	If firm name is associated to a resource intensive industry.	0.125	0.331
<i>Industry Measures (US CMP High-Tech Clusters)</i>				
Biotechnology	Estimated from name	If firm name is associated to the Biotechnology industry cluster.	0.002	0.044
E-Commerce	Estimated from name	If firm name is associated to the E-Commerce industry cluster.	0.049	0.216
Medical Devices	Estimated from name	If firm name is associated to the Medical Devices industry cluster.	0.027	0.163
Semiconductor	Estimated from name	If firm name is associated to the Semiconductor industry cluster.	0.0004	0.019
Observations			38,506,776	

This table represents our full dataset, comprised of all registered firms registered within the years 1988 and 2014 in Washington D.C. and 49 US states (excluding Delaware), and 46 states (excluding Delaware, Illinois, Michigan, South Carolina) within the year 2014 and 2016. These states account for 99.6% of US GDP in 2014. All measures defined in detail in Section III of this paper. Business registration records are public records created endogenously when a firm registers as a corporation, LLC, or partnership. IP observables include both patents and trademarks filed by the firm within a year of founding, as well as previously filed patents assigned to the firm close to founding. All business registration observables, IP observables, are estimated at or close to the time of firm founding. Further information on all measures and our approach more generally, can be found in Guzman and Stern (2018). Growth IPOs include only 'true' startup IPOs; we exclude all financial IPOs, REITs, SPACs, reverse LBOs, re-listings, and blank check corporations.

TABLE 2
 Academic Model
 Predictive Analytics Model of Equity Growth
 Dependent Variable: Equity Growth
 Logit model. Incidence Rate Ratios Reported

	(1)	Preliminary Models		Full Model
		(2)	(3)	(4)
<i>Corporate Governance Measures</i>				
Corporation	3.826*** (0.0758)			3.202*** (0.0650)
Delaware	23.61*** (0.402)			
<i>Name-Based Measures</i>				
Short Name		1.928*** (0.0245)		1.786*** (0.0208)
Eponymous		0.216*** (0.0104)		0.312*** (0.0150)
<i>Intellectual Property Measures</i>				
Patent			44.27*** (1.223)	
Trademark			8.369*** (0.439)	4.288*** (0.200)
<i>Patent - Delaware Interaction</i>				
Patent Only				20.24*** (0.847)
Delaware Only				15.26*** (0.294)
Patent and Delaware				84.08*** (2.720)
US CMP Clusters				Yes
US CMP High-Tech Clusters				Yes
N	26,051,461	26,051,461	26,051,461	26,051,461
R-squared	0.146	0.059	0.100	0.194

We estimate a logit model with Growth as the dependent variable. Growth is a binary indicator equal to 1 if a firm achieves IPO or acquisition within 6 years and 0 otherwise. Growth is only defined for firms born in the cohorts of 1988 to 2010. This model forms the basis of our entrepreneurial quality estimates, which are the predicted values of the model. Incidence ratios reported; Robust standard errors in parenthesis. * p<0.05 ** p<0.01 *** p<0.001

TABLE 3
 Policy Model
 Predictive Analytics Model of Equity Growth
 Dependent Variable: Equity Growth
 Logit model. Incidence Rate Ratios Reported

	(1)	(2)	(3)
<i>Independent Effects</i>			
Delaware	25.49*** (0.614)		
Patent	18.01*** (0.558)		
Trademark	5.343*** (0.287)	5.497*** (0.267)	
<i>Delaware, Patent Interactions</i>			
Delaware = 1, Patent = 0		34.22*** (0.839)	
Delaware = 0, Patent = 1		74.78*** (3.279)	
Delaware = 1, Patent = 1		402.9*** (14.71)	
<i>Delaware, Patent, Trademark Interactions</i>			
Delaware = 1, Patent = 0, Trademark = 0			35.53*** (0.902)
Delaware = 0, Patent = 1, Trademark = 0			89.25*** (3.894)
Delaware = 0, Patent = 0, Trademark = 1			42.40*** (3.142)
Delaware = 1, Patent = 1, Trademark = 0			528.2*** (17.97)
Delaware = 1, Patent = 0, Trademark = 1			350.6*** (20.25)
Delaware = 0, Patent = 1, Trademark = 1			163.4*** (25.26)
Delaware = 1, Patent = 1, Trademark = 1			950.4*** (68.52)
Corporation	2.071*** (0.0428)	2.490*** (0.0569)	2.670*** (0.0631)
N	26051461	26051461	26051461
pseudo R-sq	0.132	0.136	0.139

Robust standard errors reported in parenthesis. * p < .1, ** p < .05, *** p < .01

TABLE 4
Summary Statistics by Regions

Measure	<u>State</u>		<u>MSA</u>		<u>County</u>		<u>Zip Code</u>	
	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.
Firm Quantity (SFR)	23233.77	36996.34	2947.73	9407.27	336.20	1850.17	40.77	100.35
Firm Quality (EQI)	0.00048	0.00050	0.00067	0.0010	0.00040	0.00063	0.00055	0.0019
Quality-adjusted Quantity (RECPI)	15.20	47.22	1.86	6.56	0.23	2.39	0.27	0.13
Equity Growth (IPO or Acquisition)	13.73	39.86	1.84	8.24	0.210	2.13	0.24	0.23
Number of Regions	50	N.A.	362	N.A.	3,127	N.A.	37,662	N.A.
Observations	1,444	N.A.	10,362	N.A.	85,628	N.A.	802,854	N.A.

This table represents our full dataset, of all registered firms registered within the years 1988 and 2014 in Washington D.C. and 49 US states (excluding Delaware), and 47 states (excluding Delaware, Illinois, Michigan, South Carolina) within the year 1988 and 2016. These states account for 99.6% of US GDP in 2014. MSA is Metropolitan Statistical Area. All measures defined in detail in Section III of this paper.

TABLE 5

Rankings of Average RECPI / Population and Quality between 2010 and 2014 by MSA

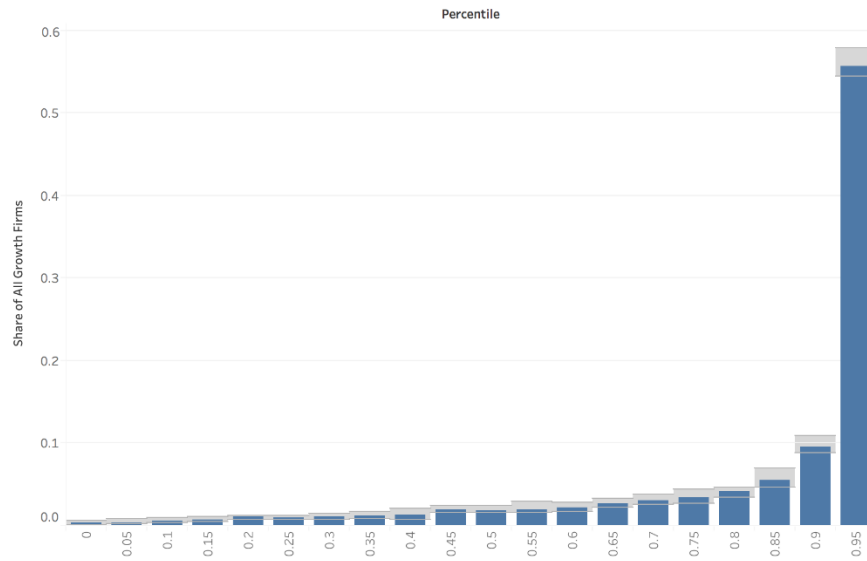
Top MSAs of Average RECPI / Population	Top MSAs of Average Quality	Bottom MSAs of Average RECPI / Population	Bottom MSAs of Average Quality
1 San Jose-Sunnyvale-Santa Clara, CA	San Jose-Sunnyvale-Santa Clara, CA	McAllen-Edinburg-Mission, TX	Tucson, AZ
2 Columbia, SC	San Francisco-Oakland-Fremont, CA	El Paso, TX	Phoenix-Mesa-Scottsdale, AZ
3 Miami-Fort Lauderdale-Pompano Beach, FL	Columbia, SC	Bakersfield, CA	Baton Rouge, LA
4 San Francisco-Oakland-Fremont, CA	San Diego-Carlsbad-San Marcos, CA	Stockton, CA	New Orleans-Metairie-Kenner, LA
5 Bridgeport-Stamford-Norwalk, CT	Boston-Cambridge-Quincy, MA-NH	Fresno, CA	Augusta-Richmond County, GA-SC
6 Madison, WI	Los Angeles-Long Beach-Santa Ana, CA	Modesto, CA	McAllen-Edinburg-Mission, TX
7 Provo-Orem, UT	Bridgeport-Stamford-Norwalk, CT	Spokane, WA	Albuquerque, NM
8 Denver-Aurora, CO	Oxnard-Thousand Oaks-Ventura, CA	Augusta-Richmond County, GA-SC	Milwaukee-Waukesha-West Allis, WI
9 Boston-Cambridge-Quincy, MA-NH	Omaha-Council Bluffs, NE-IA	Scranton--Wilkes-Barre, PA	Colorado Springs, CO
10 Los Angeles-Long Beach-Santa Ana, CA	Durham, NC	Jackson, MS	Des Moines-West Des Moines, IA
11 Austin-Round Rock, TX	Chicago-Naperville-Joliet, IL-IN-WI	Springfield, MA	Jackson, MS
12 Orlando-Kissimmee, FL	Austin-Round Rock, TX	Providence-New Bedford-Fall River, RI-MA	Dayton, OH
13 San Diego-Carlsbad-San Marcos, CA	Worcester, MA	Tucson, AZ	Akron, OH
14 Salt Lake City, UT	New York-Northern New Jersey-Long Island, NY-NJ-PA	Knoxville, TN	Toledo, OH
15 Cape Coral-Fort Myers, FL	Raleigh-Cary, NC	Youngstown-Warren-Boardman, OH-PA	Youngstown-Warren-Boardman, OH-PA
16 Sarasota-Bradenton-Venice, FL	Sacramento--Arden-Arcade--Roseville, CA	Wichita, KS	Virginia Beach-Norfolk-Newport News, VA-NC
17 Atlanta-Sandy Springs-Marietta, GA	Madison, WI	Albuquerque, NM	Indianapolis-Carmel, IN
18 New York-Northern New Jersey-Long Island, NY-NJ-PA	Providence-New Bedford-Fall River, RI-MA	Buffalo-Niagara Falls, NY	Wichita, KS
19 Tampa-St. Petersburg-Clearwater, FL	Tulsa, OK	Winston-Salem, NC	Boise City-Nampa, ID
20 Durham, NC	Riverside-San Bernardino-Ontario, CA	Lancaster, PA	Columbus, OH
21 Washington-Arlington-Alexandria, DC-VA-MD-WV	Dallas-Fort Worth-Arlington, TX	Riverside-San Bernardino-Ontario, CA	Greenville-Mauldin-Easley, SC
22 Oxnard-Thousand Oaks-Ventura, CA	Nashville-Davidson--Murfreesboro--Franklin, TN	Portland-Vancouver-Beaverton, OR-WA	Richmond, VA
23 Jacksonville, FL	Charlotte-Gastonia-Concord, NC-SC	Memphis, TN-MS-AR	Deltona-Daytona Beach-Ormond Beach, FL
24 Oklahoma City, OK	Albany-Schenectady-Troy, NY	Pittsburgh, PA	Ogden-Clearfield, UT
25 Palm Bay-Melbourne-Titusville, FL	Rochester, NY	Harrisburg-Carlisle, PA	Cincinnati-Middletown, OH-KY-IN
26 Tulsa, OK	Memphis, TN-MS-AR	Rochester, NY	Lakeland, FL
27 Raleigh-Cary, NC	Fresno, CA	Syracuse, NY	Charleston-North Charleston, SC
28 Charleston-North Charleston, SC	Modesto, CA	Dayton, OH	Kansas City, MO-KS
29 Dallas-Fort Worth-Arlington, TX	Houston-Sugar Land-Baytown, TX	Chattanooga, TN-GA	Spokane, WA
30 Chicago-Naperville-Joliet, IL-IN-WI	Philadelphia-Camden-Wilmington, PA-NJ-DE-MD	Toledo, OH	El Paso, TX

This table represents the ranking of top and bottom 30 MSAs in average quality-adjusted quantity (RECPI) / Population and quality aggregated by MSA between 2010 and 2014. We only consider those regions that ranks top 100 in average population between 2010 and 2014 and excluded fracking regions (Midland, Odessa, Casper, Tyler), Arkansas, Alabama and Michigan.

FIGURE 1

Validating Entrepreneurial Quality Predictions
10-Fold Out of Sample Test (Academic Model)

Top 1% includes 37% of all growth firms [36%, 39%]
Top 5% includes 56% of all growth firms [54%, 58%]
Top 10% includes 65% of all growth firms [64%, 67%]



Notes: We evaluate the predictive quality of our estimates by undertaking a tenfold cross-validation test, and report the out-of-sample share of realized growth outcomes at different portions of the entrepreneurial quality distribution, divided in 5-percent bins.

FIGURE 2

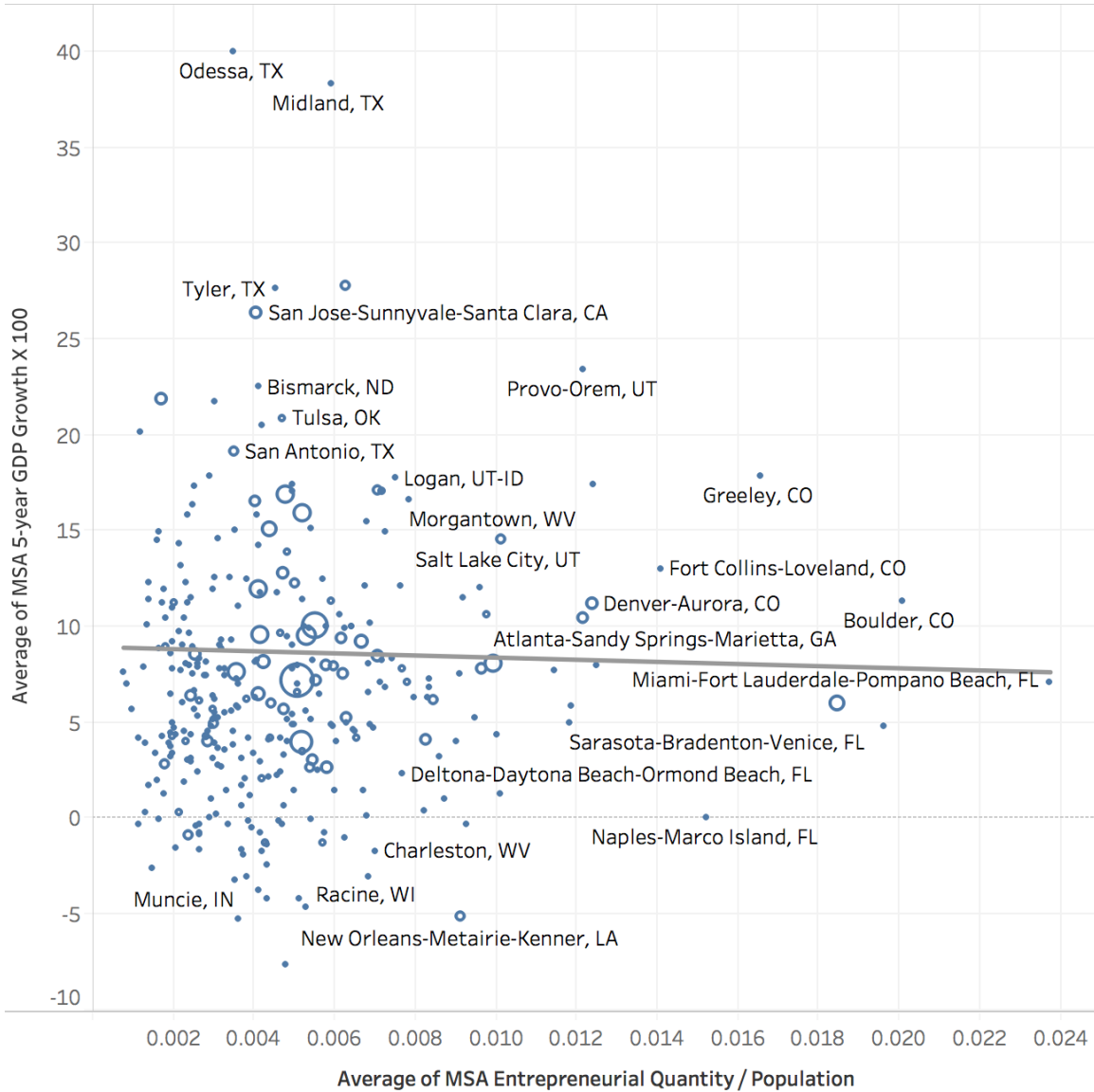
Population-adjusted Quantity and Estimated Quality by MSA



Notes: This figure compares the two dimensions of regional entrepreneurship produced by our data by U.S. Metropolitan Statistical Area (MSA). On the Y axis we include the number of companies by population of each, this is a measure of the intensity of entrepreneurial quantity. On the X-axis we instead include the estimated average quality for firms using our entrepreneurial quality approach. MSAs are defined using the U.S. Census 2013 MSA definitions.

FIGURE 3A

Does Quantity Predict Economic Growth?
Population-adjusted Quantity and Average GDP Growth by MSA



Notes: This figure presents the relationship between the average intensity of firm formation in an MSA (quantity over population), and the average real GDP growth of that MSA over the subsequent five years. The fitted line is the fit weighted by the average GDP of the MSA for the whole time period.

FIGURE 3B

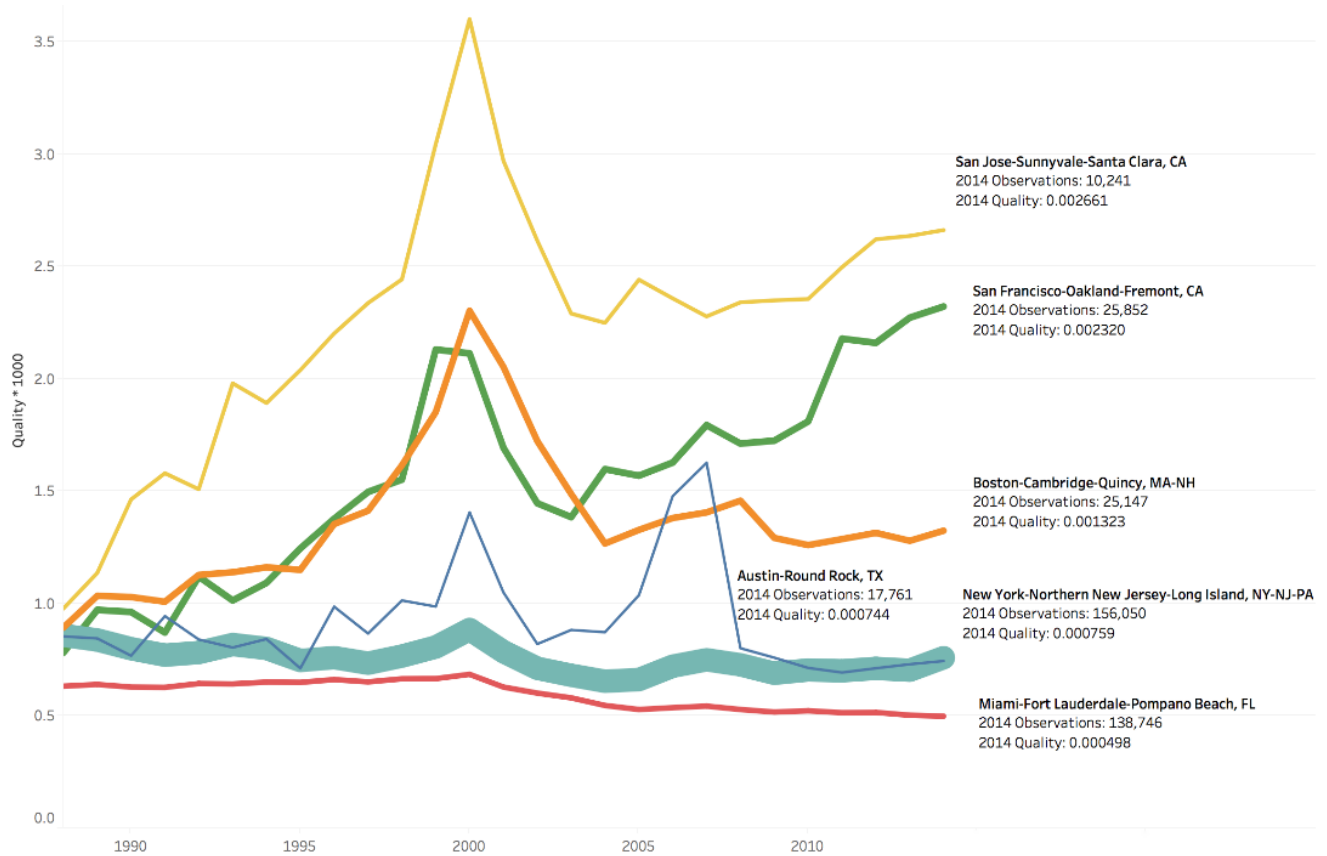
Does Quality Predict Economic Growth?
Entrepreneurial Quality and Average GDP Growth by MSA



Notes: This figure presents the relationship between the average entrepreneurial quality in an MSA (quantity over population), and the average real GDP growth of that MSA over the subsequent five years. The fitted line is the fit weighted by the average GDP of the MSA for the whole time period.

FIGURE 4

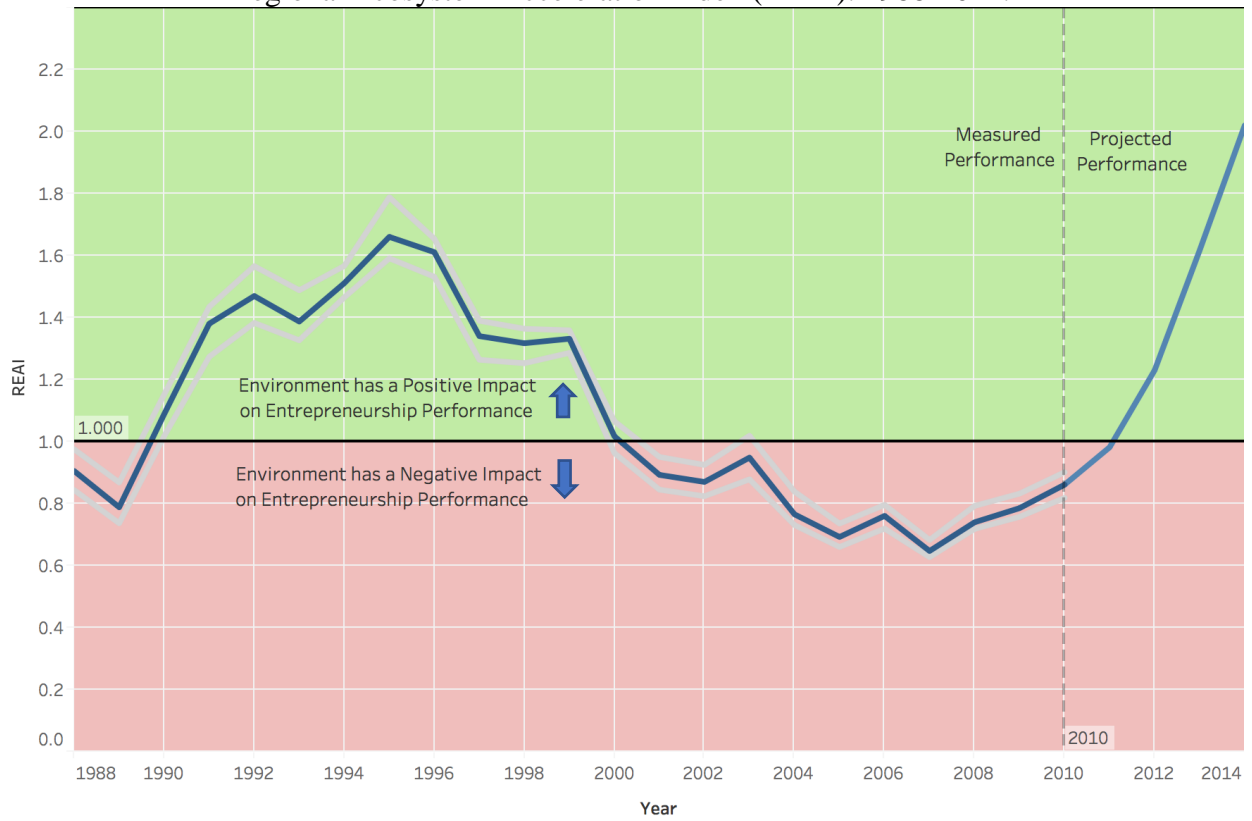
Evolution of Entrepreneurial Quality for Select MSAs, 1988-2014



Notes: This figure plots the evolution of the estimated quality by MSA for a few select MSAs in the US economy. The line represents the level of average entrepreneurial quality in a region, while the width of the line is the region's average GDP to indicate its overall economic importance. The raise of Silicon Valley during the late 1990s, and of the San Francisco area in the later-on period are particularly noticeable.

FIGURE 5

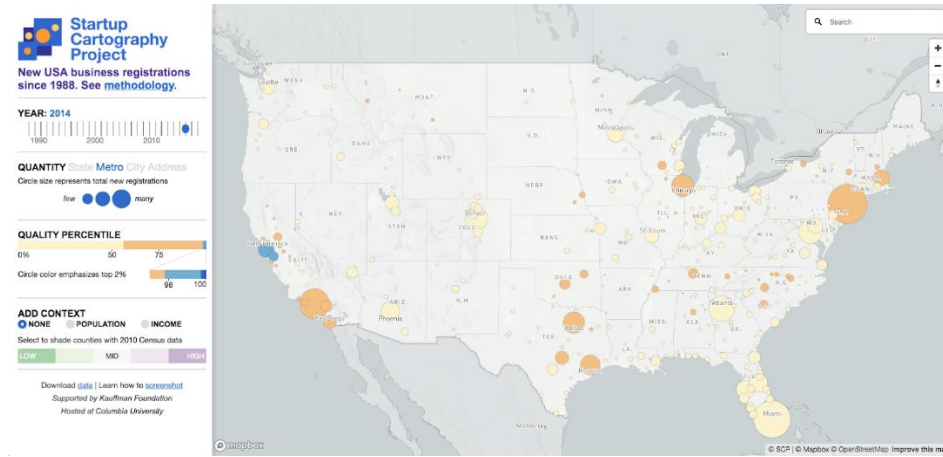
Regional Ecosystem Acceleration Index (REAI). 1988-2014.



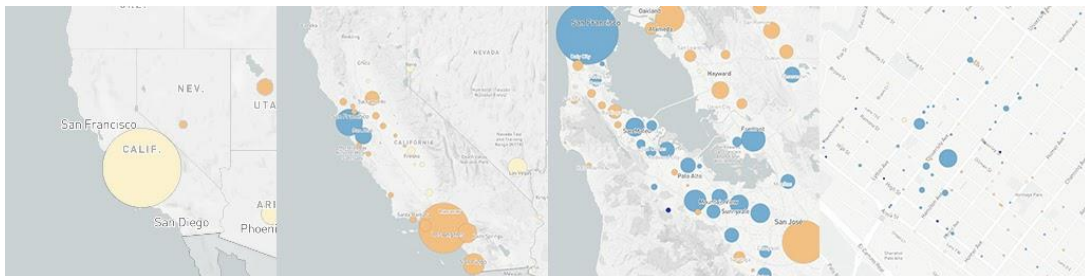
Notes: This figure plots the estimated REAI from 1988 to 2014 using our methodology. REAI is the number of growth events in a cohort divided by the expected number of growth events. Grey lines indicate a 95% confidence interval estimated by 30 bootstrap iterations.

FIGURE 6

Panel A. Startup Cartography Project Map



Panel B. Map zoom levels (L-R): State, Metro, City, Address



Panel C. Tooltip Detail: Palo Alto, CA

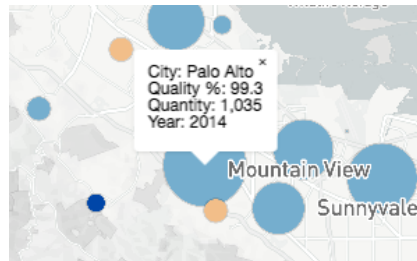
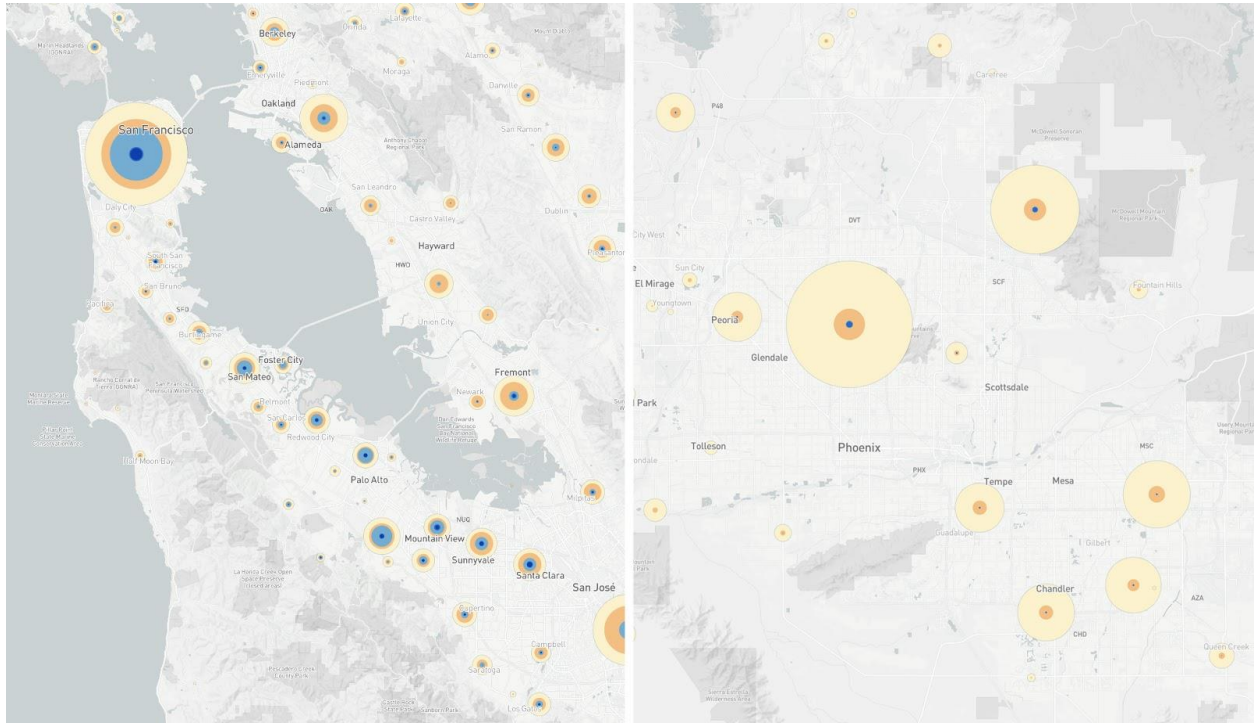


FIGURE 7

Two City Ring Views at Same Zoom Level: Silicon Valley vs. Phoenix, AZ



Notes: This figure presents two maps from the Startup Cartography project, using the concentric circles to highlight differences in the amount of entrepreneurship across cities in the San Francisco Bay area, and the Phoenix metropolitan area.

APPENDIX

TABLE A1
Summary Statistics of industry measures

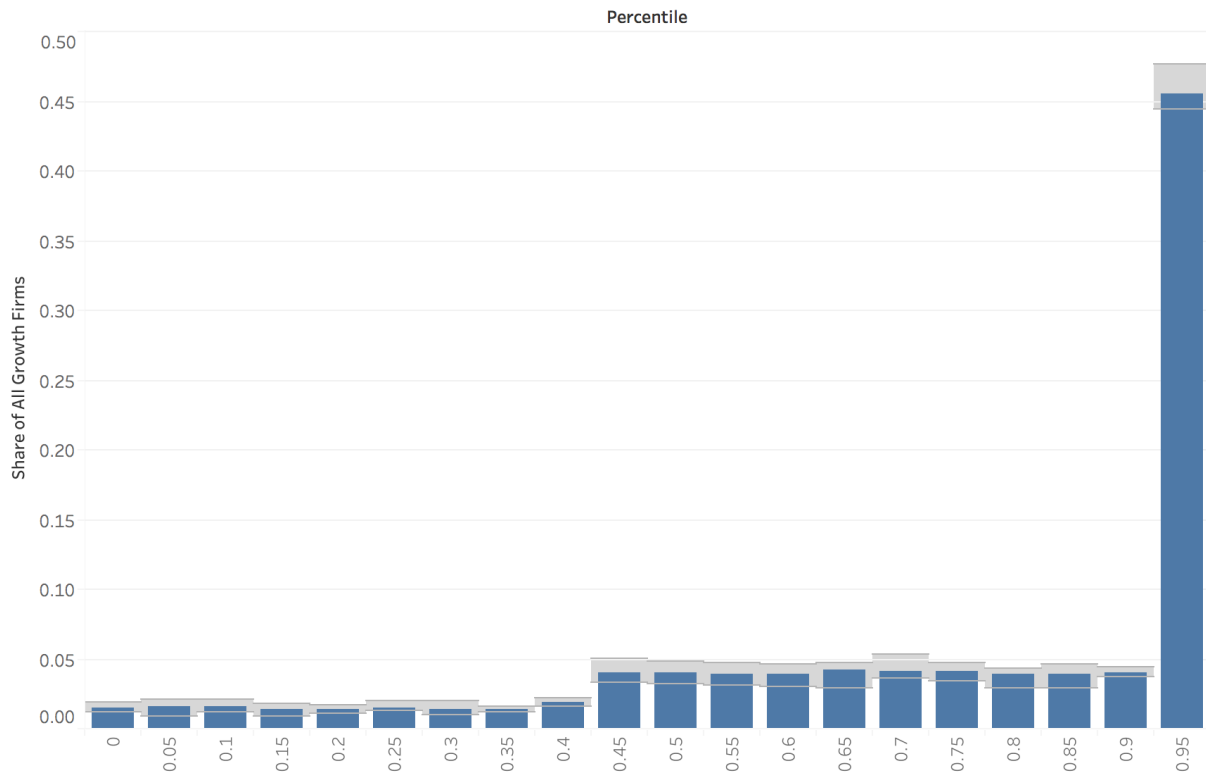
Measure	Source	Description	Mean	Std. Dev.
<i>USCMP Name Based Industry Measures</i>				
Local Industry	Business Reg.	If firm name is associated to a local industry.	0.194	0.395
Traded Industry	Business Reg.	If firm name is associated to a traded industry.	0.538	0.499
Resource Intensive Industry	Business Reg.	If firm name is associated to a resource intensive industry.	0.125	0.331
IT	Business Reg.	If firm name is associated to the IT industry cluster.	0.021	0.145
Biotechnology	Business Reg.	If firm name is associated to the Biotechnology industry cluster.	0.002	0.044
E-Commerce	Business Reg.	If firm name is associated to the E-Commerce industry cluster.	0.049	0.216
Medical Devices	Business Reg.	If firm name is associated to the Medical Devices industry cluster.	0.027	0.163
Semiconductor	Business Reg.	If firm name is associated to the Semiconductor industry cluster.	0.0004	0.019
Observations			38,506,776	

This table represents our full dataset, comprised of all registered firms registered within the years 1988 and 2014 in Washington D.C. and 49 US states (excluding Delaware), and 47 states (excluding Delaware, Illinois, Michigan, South Carolina) within the year 1988 and 2016. These states account for 99.6% of US GDP in 2014. All measures defined in detail in Section III of this paper. Business registration records are public records created endogenously when a firm registers as a corporation, LLC, or partnership. IP observables include both patents and trademarks filed by the firm within a year of founding, as well as previously filed patents assigned to the firm close to founding. All business registration observables, IP observables, are estimated at or close to the time of firm founding. Further information on all measures and our approach more generally, can be found in Guzman and Stern (2018). Growth IPOs include only ‘true’ startup IPOs; we exclude all financial IPOs, REITs, SPACs, reverse LBOs, re-listings, and blank check corporations.

FIGURE A1

Validating Entrepreneurial Quality Predictions
10-Fold Out of Sample Test (Policy Model)





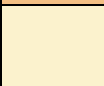
10-Fold Out of Sample Test of Predictive Quality
Top 1% includes 25% of all growth firms [23%, 26%]
Top 5% includes 46% of all growth firms [44%, 48%]
Top 10% includes 50% of all growth firms [48%, 52%]



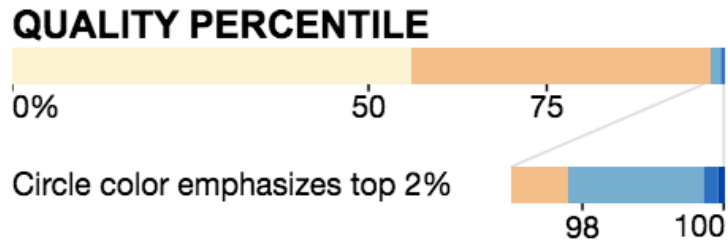
Notes: We evaluate the predictive quality of our estimates by undertaking a tenfold cross-validation test, and report the out-of-sample share of realized growth outcomes at different portions of the entrepreneurial quality distribution, divided in 5-percent bins.

FIGURE A2

Semantic Color Table

Quality	Color	Name	Registration attributes
Highest		Dark Blue	<i>At least two:</i> Delaware Registration, Patent, Trademark
		Blue	Patent <i>or</i> Trademark
		Pale Blue	Delaware Registration
		Orange	Corporation
Lowest		Pale Orange	LLC

Legend Color Detail



SUPPLEMENTARY MATERIALS TO:

**THE STARTUP CARTOGRAPHY PROJECT:
Measuring and Mapping Entrepreneurial Ecosystems**

RJ Andrews

Catherine Fazio

Jorge Guzman

Yupeng Liu

Scott Stern

Online Appendix

APPENDIX A
DATA APPENDIX

I. Overview of Data Appendix

This data appendix to the paper *The State of American Entrepreneurship*, by Jorge Guzman and Scott Stern, outlines in detail the use of business registration records in the United States, the steps and decisions we took when converting those records into measures for analysis, and robustness tests we ran to validate the potential for bias both due to specific assumptions about each measure as well as heterogeneity in our sample across geography and time. It serves the dual purpose of serving as an introduction for future users of business registration data while also providing detailed robustness verification and explaining the logic of specific decisions on many aspects of our data.

Section II of this appendix explains the development of our measures and dataset, including how we matched multiple datasets for analysis, how we built our measures using the merged dataset, and the economic rationale for the production of each one. Section III explains the differences between business registration records across the United States, their ease of access, and variation in the data they provide. It also highlights the potential for bias given the time when different data is observed (i.e. whether we observe the most recent value of a business or the original one) and performs numerous robustness tests to rule out the potential for bias driving our results given these differences. Section IV analyzes the potential for bias in our aggregate RECPI with a focus on guaranteeing that the predictive value of our indexes is high across geographies and time, and is not driven by a particularly large startup period (e.g. the dot-com bubble) nor driven by a particular area with many growth startups (e.g. Silicon Valley).

II. Using Business Registration Records to Find Signals of Quality

Our data set is drawn from the complete set of business registrants in thirty two states from 1988 to 2014. Our analysis draws on the complete population of firms satisfying one of the following conditions: (i) a for-profit firm whose jurisdiction is in the source state or (ii) a for-profit firm whose jurisdiction is in Delaware but whose principal office address is in the state. The resulting data set is composed of 27,976,477 observations. For each observation, we construct variables related to (i) the growth outcome for the startup, (ii) measures based on business registration observables and (iii) measures based on external observables that can be linked to the startup.

Growth outcome. The growth outcome utilized in this paper, *Growth*, is a dummy variable equal to 1 if the startup achieves an initial public offering (IPO) or is acquired at a meaningful positive valuation within 6 years of registration. Both outcomes, IPO and acquisitions, are drawn from Thomson Reuters SDC Platinum¹. Although the coverage of IPOs is likely to be nearly comprehensive, the SDC data set excludes some acquisitions. SDC captures their list of acquisitions by using over 200 news sources, SEC filings, trade publications, wires, and proprietary sources of investment banks, law firms, and other advisors (Churchwell, 2016). Barnes, Harp, and Oler (2014) compare the quality of the SDC data to acquisitions by public firms and find a 95% accuracy (Netter, Stegemoller, and Wintoki (2011), also perform a similar review). While we know this data not to be perfect, we believe it to have relatively good coverage of ‘high value’ acquisitions. We also note that none of the cited studies found significant false positives,

¹ Thomson Reuters’s SDC Platinum is a commonly used database of financial information transferred to Refinitiv in 2018. More details are available at <https://www.refinitiv.com>.

suggesting that the only effect of the acquisitions we do not track will be an attenuation of our estimated coefficients.

We observe 13,406 positive growth outcomes for the 1988–2008 start-up cohorts), yielding a mean for *Growth* of 0.0007. In our main results, we assign acquisitions with an unrecorded acquisitions price as a positive growth outcome, because an evaluation of those deals suggests that most reported acquisitions were likely in excess of \$5 million. We perform a series of robustness tests on different outcomes in the next section of this data appendix.

Start-up characteristics. The core of the empirical approach is to map growth outcomes to observable characteristics of start-ups at or near the time of business registration. We develop two types of measures: (i) measures based on business registration observables and (ii) measures based on external indicators of start-up quality that are observable at or near the time of business registration. We review each of these in turn.

Measures based on business registration observables. We construct six measures of start-up quality based on information directly observable from the business registration record. First, we create binary measures related to how the firm is registered, including *corporation*, whether the firm is a corporation (rather than partnership or LLC) and *Delaware jurisdiction*, whether the firm is incorporated in Delaware. *Corporation* is an indicator equal to 1 if the firm is registered as a corporation and 0 if it is registered either as an LLC or partnership.² In the period of 1988 to 2008, 0.10% of corporations achieve a growth outcome versus only 0.03% of noncorporations. *Delaware jurisdiction* is equal to 1 if the firm is registered under Delaware, but has its main office in the source state (all other foreign firms are dropped before analysis). Delaware jurisdiction is favorable for firms which, due to more complex operations, require more certainty in corporate

² Previous research highlights performance differences between incorporated and unincorporated entrepreneurs (Levine and Rubinstein, 2013).

law, but it is associated with extra costs and time to establish and maintain two registrations. Between 1988 and 2008, 2.4% of the sample registers in Delaware; 37% of firms achieving a growth outcome do so.

Second, we create four measures that are based on the name of the firm, including a measure associated with whether the firm name is eponymous (named after the founder), is short or long, is associated with local industries (rather than traded), or is associated with a set of high-technology industry clusters.

Drawing on the recent work of Belenzon, Chatterji, and Daley (2017) (BCD), we use the firm and top manager name to establish whether the firm name is eponymous (i.e., named after one or more of the president, CEO, chairman, or managers (in the case of LLCs and partnerships)). *Eponymy* is equal to 1 if the first, middle, or last name of the top managers is part of the name of the firm itself.³ We require names be at least four characters to reduce the likelihood of making errors from short names. Our results are robust to variations of the precise calculation of eponymy (e.g., names with a higher or lower number of minimum letters). We have also undertaken numerous checks to assess the robustness of our name matching algorithm. Not all states include the name of top managers⁴. Within those that do, 7.7% of the firms in our training sample are eponymous [an incidence rate similar to BCD], though only 2.4% for whom *Growth* equals one. It is useful to note that, while we draw on BCD to develop the role of eponymy as a useful start-up characteristic, our hypothesis is somewhat different than BCD: we hypothesize that eponymous firms are likely to be associated with lower entrepreneurial quality. Whereas BCD evaluates whether serial entrepreneurs are more likely to invest and grow companies which they name after

³For corporations, we consider top managers only the current president, for partnerships and LLCs, we allow for any of the two listed managers. The corporation president and two top partnership managers are listed in the business registration records themselves.

⁴These, and other, institutional differences are taken care of in our specifications through the inclusion of state fixed effects.in

themselves, we focus on the cross-sectional difference between firms with broad aspirations for growth (and so likely avoid naming the firm after the founders) versus less ambitious enterprises, such as family-owned “lifestyle” businesses.

Our second measure relates to the length of the firm name. Based on our review of naming patterns of growth-oriented start-ups versus the full business registration database, a striking feature of growth-oriented firms is that the vast majority of their names are at most two words (plus perhaps one additional word to capture organizational form (e.g., “Inc.”). Companies such as Google or Spotify have sharp and distinctive names, whereas more traditional businesses often have long and descriptive names (e.g., “Green Valley Home Health Care & Hospice, Inc.”). We define *short name* to be equal to one if the entire firm name has three or less words, and zero otherwise. 46% of firms within the 1988-2008 period have a short name, but the incidence rate among growth firms is more than 73%. We have also investigated a number of other variants (allowing more or less words, evaluating whether the name is “distinctive” (in the sense of being both non-eponymous and also not an English word). While these are promising areas for future research, we found that the three-word binary variable provides a useful measure for distinguishing entrepreneurial quality.

We then create four measures based on how the firm name reflects the industry or sector that the firm is operating. To do so, we take advantage of two features of the US Cluster Mapping Project (Delgado, Porter, and Stern, 2016), which categorizes industries into (a) whether that industry is primarily local (demand is primarily within the region) versus traded (demand is across regions) and (b) among traded industries, a set of 51 traded clusters of industries that share complementarities and linkages. We augment the classification scheme from the US Cluster Mapping Project with the complete list of firm names and industry classifications

contained in Reference USA, a business directory containing more than 10 million firm names and industry codes for companies across the United States. Using a random sample of 1.5 million Reference USA records, we create two indices for every word ever used in a firm name. The first of these indices measures the degree of localness, and is defined as the relative incidence of that word in firm names that are in local versus non-local industries (i.e., $\rho_i = \frac{\sum_{j=\{\text{local firms}\}} 1[w_i \subseteq \text{name}_j]}{\sum_{j=\{\text{non-local firms}\}} 1[w_i \subseteq \text{name}_j]}$). We then define a list of Top Local Words, defined as those words that are (a) within the top quartile of ρ_i and (b) have an overall rate of incidence greater than 0.01% within the population of firms in local industries (see Guzman and Stern, (2015, Table S10) for the complete list). Finally, we define local to be equal to one for firms that have at least one of the Top Local Words in their name, and zero otherwise. We then undertake a similar exercise for the degree to which a firm name is associated with a traded name. It is important to note that there are firms which we cannot associate either with traded or local and thus leave out as a third category. Just more than 19% of firms have local names, though only 5% of firms for whom growth equals one, and while 54% of firms are associated with the traded sector, 59% of firms for whom growth equals one do.

We additionally examine the type of traded cluster a firm is associated with, focusing in particular on whether the firm is in a high-technology cluster or a cluster associated with resource intensive industries. For our high technology cluster group (Traded High Technology), we draw on firm names from industries include in ten USCMP clusters: Aerospace Vehicles, Analytical Instruments, Biopharmaceuticals, Downstream Chemical, Information Technology, Medical Devices, Metalworking Technology, Plastics, Production Technology and Heavy Machinery, and Upstream Chemical. From 1988 to 2008, while only 5% firms are associated with high technology, this rate increases to 16% within firms that achieve our growth outcome. For our resource

intensive cluster group, we draw on firms names from fourteen USCMP clusters: Agricultural Inputs and Services, Coal Mining, Downstream Metal Products, Electric Power Generation and Transmission, Fishing and Fishing Products, Food Processing and Manufacturing, Jewelry and Precious Metals, Lighting and Electrical Equipment, Livestock Processing, Metal Mining, Nonmetal Mining, Oil and Gas Production and Transportation, Tobacco, Upstream Metal Manufacturing. While 14% of firms are associated with resource intensive industries, and 13% amongst growth firms.

Finally, we also repeat the same procedure to find firms associated with more narrow sets of clusters that have a closer linkage to growth entrepreneurship in the United States. We specifically focus on firms associated to Biotechnology, E-Commerce, Information Technology, Medical Devices and Semiconductors. It is important to note that these definitions are not exclusive and our algorithm could associate firms with more than one industry group. For Biotechnology (Biotechnology Sector), we use firm names associated with the US CMP Biopharmaceuticals cluster. While only 0.19% of firms are associated with Biotechnology, this number increases to 2.2% amongst growth firms. For E-commerce (E-Commerce Sector) we focus on firms associated with the Electronic and Catalog Shopping sub-cluster within the Distribution and Electronic Commerce cluster. And while 5% of all firms are associated with e-commerce, the rate is 9.3% for growth firms. For Information Technology (IT Sector), we focus on firms related to the USCMP cluster Information Technology and Analytical Instruments. 2.4% of all firms in our sample are associated with IT, and 12% of all growth firms are identified as IT-related. For Medical Devices (Medical Dev. Sector), we focus on firms associated with the Medical Devices cluster. We find that while 3% of all firms are in medical devices, this number increases to 9.6% within growth firms. Finally, for Semiconductors (Semiconductor Sector), we focus on the sub-

cluster of Semiconductors within the Information Technology and Analytical Instruments cluster. Though only 0.04% of all firms are associated with semiconductors, 0.5% of growth firms are.

Measures based on external observables. We construct two measures related to start-up quality based on information in intellectual property data sources. Although this paper only measures external observables related to intellectual property, our approach can be utilized to measure other externally observable characteristics that may be related to entrepreneurial quality (e.g., measures related to the quality of the founding team listed in the business registration, or measures of early investments in scale (e.g., a Web presence).

Building on prior research matching business names to intellectual property (Balasubramanian and Sivadasan, 2010; Kerr and Fu, 2008), we rely on a name-matching algorithm connecting the firms in the business registration data to external data sources. Importantly, because we match only on firms located in California, and because firms names legally must be “unique” within each state’s company registrar, we are able to have a reasonable level of confidence that any “exact match” by a matching procedure has indeed matched the same firm across two databases. In addition, our main results use “exact name matching” rather than “fuzzy matching”; in small-scale tests using a fuzzy matching approach [the Levenshtein edit distance (Levenshtein, 1965)], we found that fuzzy matching yielded a high rate of false positives due to the prevalence of similarly named but distinct firms (e.g., Capital Bank v. Capitol Bank, Pacificorp Inc v. Pacificare Inc.).

Our matching algorithm works in three steps.

First, we clean the firm name by:

- expanding eight common abbreviations (“Ctr.,” “Svc.,” “Co.,” “Inc.,” “Corp.,” “Univ.,” “Dept.,” “LLC.”) in a consistent way (e.g., “Corp.” to “Corporation”)
- removing the word “the” from all names
- replacing “associates” for “associate”
- deleting the following special characters from the name: . | ’ ” - @ _

Second, we create measures of the firm name with and without the organization type, and with and without spaces. We then match each external data source to each of these measures of the firm name. The online appendix contains all of the data and annotated code for this procedure.

This procedure yields two variables. Our first measure of intellectual property captures whether the firm is in the process of acquiring patent protection during its first year of activity. *Patent* is equal to 1 if the firm holds a patent application in the first year. All patent applications and patent application assignments are drawn from the Google U.S. Patent and Trademark Office (USPTO) Bulk Download archive. We use patent applications, rather than granted patents, because patents are granted with a lag and only applications are observable close to the data of founding. Note that we include both patent applications that were initially filed by another entity (e.g., an inventor or another firm), as well as patent applications filed by the newly founded firm. While only 0.2% of the firms in 1988–2008 have a first-year patent, 14% of growth firms do.

Our second intellectual property measure captures whether a firm registers a trademark during its first year of business activity. *Trademark* is equal to 1 if a firm applied for a trademark within the first year, and 0 otherwise. We build this measure from the Stata-ready trademark DTA file developed by the USPTO Office of Chief Economist (Graham et al, 2013). Between 1988 and 2008, 0.11% of all firms register a trademark, while 4.7% of growth firms do.

APPENDIX B

ANALYZING SOURCES OF BIAS USING MASSACHUSETTS DATA

We now turn to analyzing the potential for bias in our estimates due to the specific nature of business registration records. We specifically comment on six specific areas where there exists the possibility of bias: the impact of unobserved name changes, the role of re-incorporations on our data, the impact of spin-offs vs new firms, changes of ownership, changes in firm location, and the role of subsidiaries as separate corporate entities. We review each one in turn.

Changes in Firm Location. A main concern in our analysis is the potential of bias from changes in firm location. The data we receive from business registries holds the *current* location of the firm, but our goal in understanding entrepreneurial quality geography is to understand the *initial* location of the firm. (Importantly this does not impact our firm-level quality estimates, and hence we can analyze variation across different unbiased ex-ante quality levels of firms.) Firms are likely to move for many reasons. Ex-ante better firms might be more likely to start close to the center of an entrepreneurial cluster as it might have more value for the local externalities and move out of high potential clusters if unsuccessful, while ex-post successful firms (with lower quality ex-ante) might be more likely to move into such clusters. The potential direction and effect of this bias is in principle unclear.

While we are unable to study the extent of this bias in all states, we are able to perform a sub-sample study in Massachusetts. Using Massachusetts offers several important benefits that support the robustness of any forthcoming conclusions. First, our samples are beneficial: We are able to obtain two samples in Massachusetts that are almost exactly two years apart (one from January 06, 2013 (Commonwealth of Massachusetts, 2013), and one from November 24, 2014 (Commonwealth of Massachusetts, 2014)); furthermore, a sample from January 2013 provides the earliest possible snapshot that includes all 2012 firms (the most recent firms for which we estimate our full quality model, and the data we use for our full US snapshot), and hence includes the

address in the firm’s actual registration. Second, Massachusetts requires firms to update their address (among other things) in a yearly annual report guaranteeing we observe the new address for all firms that move. In other states, such annual report is not necessary. If a firm doesn’t report its new address, we would continue to observe the original business address even after it moves, and our analysis will hold no bias. And third, the period we consider is a period in which there is considerable geographic migration of high-quality firms within Massachusetts, from Route 128 to the Cambridge and Boston area (see Guzman and Stern, 2017 for further details). Each of these details guarantees that our estimate is most likely to be an upper bound, and the extent of bias identified in this analysis is, if anything, likely to be lower in our national sample.

For this analysis, given that the ZIP Code is the smallest unit of geographic measurement that we use in this paper, we focus all of our analysis in ZIP Code level variation⁵. First, for each firm, we keep their 2013 ZIP Code (observed in January 06, 2013) their 2015 ZIP Code (observed in November 24, 2014). We also geocode each ZIP Code to assess the distance of any geographic move and remove all firms that have an invalid ZIP Code (e.g. due to typos)⁶. Finally, we estimate the leave-self-out quality of each ZIP Code for each firm using the average quality of all firms from 1988-2012 in our sample period.

We begin by documenting the extent to which a firm changes location at all. Table B3 presents the rates of change in ZIP Code for each 2-year group in our data. The first column indicates the age of the firm in 2013, when we first observe it, and the second column the share of firms that stay in the same ZIP Code in the next two years for the group. These estimates are not conditional on survival, and thus capture the share of total firms that will change from one category

⁵ This also helps protect from noise that could occur from “fuzzy” address matching approaches rather than exact ZIP Code matching.

⁶ We consider all ZIP Codes we cannot geocode through the Google API to be invalid.

to the next in the total sample (i.e. it controls for changes in survival probability), the quantity we are interested on. Firms under 4 years or less (at 2013) are most likely to change address, with a probability of change between 2.9% and 3.6%. This probability then drops quickly, and in the 26-year-old cohort the probability of change is only 0.3%. Because our measure implicitly also includes likelihood of survival at different cohorts, we can estimate the overall likelihood that a firm record will have a different address after N years by simply doing the running product of the probability of same ZIP Code (under the assumption the migration dynamics have been the same historically). Column 3 includes this result. For the cohort of 10-year-old firms, we estimate 95% of the records to still contain the original ZIP Code, and for 26-year-old firms we estimate this share at 88%. We repeat this exercise with only the top 10% of quality firms in the distribution. While the likelihood of change of ZIP Code for a high-quality firm is higher, even within this group, we estimate 89% of records still contain the original ZIP Code by 10 years and 81% by 26 years. In unreported tests, we find the share of firms that move in the top 1% is not meaningfully higher than the top 10%.

In our paper, most of our micro-geography results are done based on spatial visualizations. We therefore would also like to know *how far* are the firms moving. If firms are moving to contiguous ZIP Codes around the same high-quality cluster, perhaps due to small relocations or even ZIP Code redistricting, then the impact of those moves on our maps is small. On the contrary, if they move over large distances, then the impact is large. Using geocodings for each ZIP Code we estimate the distance of each ZIP code to another. We find 25% of all firms move less than 4 miles (25th percentile is 3.5), 50% of all firm moves are on less than 8 miles (50th percentile is 7.2), and 90% of all moves are 30 miles or less (90th percentile is 28.7).

Finally, any firm movement across ZIP Codes can only bias our results if it is systematic. If the moves are instead random, then average ZIP Code quality (our measure) would be constant even after there is firm migration. We estimate the difference in ZIP Code quality before and after a firm move (ZIP Code quality is estimated using all firms in that ZIP Code in November 24, 2014, without the moving firm included in either the source or destination ZIP Codes), and present a histogram of this measure in Figure B1. This difference in ZIP Code quality has a mean and median both basically centered at zero, therefore suggesting these moves are unbiased.

As a final test, we investigate whether this difference can vary by firm quality or age – i.e. if firms of higher or lower quality (or age) can systematically move to higher or lower average quality ZIP Codes. To do so, we run an OLS regression of firm quality on difference in ZIP Code (both in natural logs to account for the substantial skewness in entrepreneurial quality measures and be able to interpret this as an elasticity). The coefficient is -0.007 with a p-value of .61 using robust standard errors and an R^2 of .0001. This effect is (basically) indistinguishable from zero. We also regress log-age on difference in ZIP Code quality to get a coefficient of -.0014 with a p-value of .94 and R^2 of .000.

Name changes. As mentioned in section I of this appendix, we receive the original name for only some states in our dataset and only the current name in the rest of the states. While changes in name that correlate to growth could bias the relationship between our name-based measures and growth, it is unlikely to bias our most important measures. Specifically, changes in name cannot impact firm legal type (corporations vs non-corporations) or firm jurisdiction (Delaware). Our name-matching algorithm to match patents and trademarks uses firm names and assumes that the name we use is the same name as in the patent. While this can result in bias, it is only a bias that would work against our results – since we look for patents around the registration date, we can

have false negatives for firms where we are looking for the wrong (new) name in the patent record but the firm had a previous name, but false positives are much less likely. These governance and intellectual property measures are, in fact, the most important in our study, and we find the fact that they cannot be affected by name changes assuring. Perhaps a risk in using only original names in some states is that the rate of false negatives will change depending on states. In unreported robustness tests, we have found the variation in results from using always the final name for all states (and hence implicitly having the same bias for all states) to be immaterial for our results.

Change of Ownership. Our dataset differs from other datasets in what is a firm and how it changes depending on ownership. The Longitudinal Business Database is built using tax records from corporate entities. As such, establishments that change ownership might bias the sample in different way and users of this data take substantial care to make sure changes in ownership do not drive their results (e.g. see the data appendix of Decker, Haltiwanger, Jarmin, and Miranda, 2014). Our data is different. Changes in ownership do not affect the registered firm and, unless the firm is closed down and re-incorporated, changes in ownership do not change anything in registration records.

The potential for re-incorporations. We argue in our analysis that we identify the extent to which firms are born with different quality, which is observed to the entrepreneur. An alternative hypothesis would be that entrepreneurs change their firm type once they observe their potential, at which point they re-incorporate the firm differently (e.g. as a Delaware corporation). To study the possibility of this bias we take advantage of institutional details of the process through which firms re-incorporate to observe the instances when it occurs. When a low potential firm (e.g. a Massachusetts LLC) re-incorporates as a high-quality firm (e.g. a Delaware corporation), it is done in two steps. First, a new firm is registered under the high quality regime; then, the old firm is

merged into the new firm so that the new firm holds the old firm's assets and other matters (note that it is not possible to just "convert" the firm among firm types without creating a new target firm).

Once again, we use our Massachusetts data, which also includes a list of all mergers that have occurred among registered firms and the date of each merger. Obviously, firms can merge for many reasons and re-incorporation is only one of them. We create a measure *Re-registration*, which is equal to 1 only when the target firm was registered close to the merger date (90 days window). The facts we identify are included in Table B4. We review each in turn.

We identify a total of 6,767 mergers where the target firm is in Massachusetts (we drop all other firms earlier in our data, including firms registered before 1988 and firms with domicile outside Massachusetts). Of those, 3,041 firms (44.94%) are re-registrations, which are 2,847 new firms (sometimes multiple firms merge into one), while the rest are not. This total is low relative to the total firms in our sample for Massachusetts, 518,921 firms, suggesting that at most 0.55% of firms can potentially have a bias. We identify 1,905 cases in which both the source and target are in our dataset, with the rest likely being firms either registered before 1988 or with a foreign domicile.

We now proceed by studying our five most significant variables in this transition: patent, trademark, Delaware Jurisdiction, Corporation). Our main goal is to understand the extent to which founders of low-quality firms might later on re-register as high quality firms. To do so, we estimate the number firms that "gain" each of these observables, where a "gain" means the source firm did not have the observable, but the new firm does (e.g. the source firm is not a Corporation but the new firm is). We also compare this number with the total number of firms with this measure equal to 1 in our Massachusetts sample. As can be seen in Table B5, in all cases, the share of firms

that gain a positive observable is always less than 3%. In Delaware, the observable which might hold the most bias, only 0.84% of all Delaware firms are re-registrations of firms changing corporate form, while the other 99.2% is not.

REFERENCES

- N. Balasubramanian, J. Sivadasan, (2010) "NBER Patent Data-BR Bridge: User guide and technical documentation" *SSRN Working paper #1695013*
- B. Barnes, N. Harp, D. Oler, (2014) "Evaluating the SDC mergers and acquisitions database" *The Financial Review*. 49(4): 93-822.
- Belenzon, Sharon, Chatterji, Aaron and Brendan Daley. 2017. "Eponymous Entrepreneurs" *American Economic Review* 107(6):1638-55, June 2017
- C. Churchwell. (2016). "Q. SDC: M&A Database". *Baker Library – Fast Answers*. Url: <http://asklib.library.hbs.edu/faq/47760>. (Accessed on January 17, 2017.)
- Commonwealth of Massachusetts. 2013. *Business Registration Database*. Commonwealth of Massachusetts, Corporations Division. <https://www.sec.state.ma.us/cor/coridx.htm> (Received: January 06, 2013).
- Commonwealth of Massachusetts. 2014. *Business Registration Database*. Commonwealth of Massachusetts, Corporations Division. <https://www.sec.state.ma.us/cor/coridx.htm> (Received: November 24, 2014).
- M. Delgado, M. Porter, S. Stern, (2016) "Defining clusters in related industries" *Journal of Economic Geography*. 16 (1): 1-38
- S. Graham, G. Hancock, A. Marco, A. F. Myers, (2013) "The USPTO case files data set: Descriptions, lessons and insights" *SSRN Working Paper #2188621*
- W. R. Kerr, Shihe Fu, (2008) "The Survey of Industrial R&D--Patent Database Link Project." *J. Technol. Transf.* 33, no. 2
- V. I. Levenshtein, (1965) "Binary codes capable of correcting deletions, insertions, and reversals." *Doklady Akad. Nauk SSSR* 163(4): 845–848
- R. Levine, Y. Rubinstein, (2013) "Smart and illicit: Who becomes an entrepreneur and does it pay?" *NBER Working Paper #19276*
- J. Netter, M. Stegemoller, and M. B. Wintoki. (2011) "Implications of Data Screens on Merger and Acquisition Analysis: A Large Sample Study of Mergers and Acquisitions from 1992 to 2009" *The Review of Financial Studies*. 24 (7): 2316–2357.

St. Louis Fed. 2017. “Real Gross Domestic Product” *Federal Reserve Economic Data (FRED)*. <https://fred.stlouisfed.org/series/GDPCA> (Accessed on July of 2017)

ReferenceUSA. 2014. “ReferenceUSA Business Historical Data Files.” Harvard Dataverse. <https://doi.org/10.7910/DVN/GW2P3G> (Accessed on the summer of 2016)

United States Census. 2017. “Legacy Firm Characteristics Tables 1977-2014”. *Business Dynamics Statistics*. <https://www.census.gov/programs-surveys/bds.html>. United States Census Bureau (Accessed on July of 2017).

TABLE B3**Test of changes of address using a Massachusetts subsample****P(Address Change) by Age**

Lifespan	<i>All Firms</i>		<i>Top 10% of Quality</i>	
	P(Address Change) in Two Years	Lifetime Probability	P(Address Change) in Two Years	Lifetime Probability
0-2	3.6%	96.4%	6.2%	93.8%
2-4	2.9%	89.5%	5.2%	83.5%
4-6	2.0%	86.3%	3.7%	77.6%
6-8	1.5%	84.8%	2.2%	75.4%
8-10	1.2%	83.6%	1.8%	73.6%
10-12	1.0%	95.4%	1.3%	92.4%
12-14	1.0%	94.5%	1.4%	91.1%
14-16	0.8%	93.7%	1.0%	90.0%
16-18	0.8%	92.9%	0.7%	89.4%
18-20	0.6%	92.3%	0.6%	88.7%
20-22	0.5%	89.0%	0.6%	82.9%
22-24	0.4%	88.6%	0.9%	82.0%
24-26	0.3%	88.3%	0.7%	81.3%

Cohort of Age 0 is the 2012 Cohort

Lifetime probability of address change is the implied probability of changing address for a firm

TABLE B5**Re-Registrations in Massachusetts***General Statistics*

Total Massachusetts Firms in Sample	518,921
Firms founded through a re-registration	2,847
Share of Firms Founded through re-registration	0.55%
Re-incorporations with source and destination firm in sample	1,905

Corporations

Firms that Gain Corporation = 1	573
Total Corporations in Sample	310,061
Share	0.18%

Delaware Jurisdiction

Firms that Gain Delaware = 1	259
Total Delaware Firms in Sample	30,781
Share	0.84%

Patents

Firms that Gain Patent = 1	51
Total Patent Firms in Sample	2,670
Share	1.91%

Trademark

Firms that Gain Trademark = 1	36
Total Trademark Firms in Sample	1,463
Share	2.46%

Short Name

Firms that Gain Short Name = 1	234
Total Short Name Firms in Sample	250,212
Share	0.09%

A firm is coded as gaining an observable if the source firm of the re-registration did not have such observable at birth but the new firm does.

FIGURE B1

