



Paper to be presented at the
DRUID Society Conference 2014, CBS, Copenhagen, June 16-18

Nowcasting High Growth Entrepreneurship - A Methodology using Public

Records

Jorge Guzman
MIT SLoan

-
jorgeg@mit.edu

Scott Stern
MIT
Sloan School of Management
sstern@mit.edu

Abstract

We propose a novel approach for measuring and tracking growth entrepreneurship ? the founding and evolution of new companies with the intention and potential for growth. While most research in entrepreneurship has either focused on narrow samples of firms that have already achieved significant growth outcomes (e.g., receiving venture capital) or considered aggregate measures such as the rate of self-employment, we develop a new method for identifying the potential for growth among the population of firms that have met a simple but fundamental requirement for growth ? incorporation

NOWCASTING HIGH GROWTH ENTREPRENEURSHIP
A METHODOLOGY USING PUBLIC RECORDS

Jorge Guzman
MIT Sloan

Scott Stern
MIT Sloan & NBER

May, 2014

PRELIMINARY
PLEASE DO NOT DISTRIBUTE WITHOUT AUTHOR PERMISSION

Abstract

We propose a novel approach for measuring and tracking growth entrepreneurship – the founding and evolution of new companies with the intention and potential for growth. While most research in entrepreneurship has either focused on narrow samples of firms that have already achieved significant growth outcomes (e.g., receiving venture capital) or considered aggregate measures such as the rate of self-employment, we develop a new method for identifying the potential for growth among the population of firms that have met a simple but fundamental requirement for growth – incorporation. Our approach takes advantage of the fact that incorporation records are public and systematic. We bring together data about the medium-term growth outcomes of firms (e.g., whether they are able to grow to a particular size, whether they receive venture financing) with data about firms that is observable at the time of incorporation (ranging from the name of the firm, their stated purpose as a company, and auxiliary information such as whether they have received a patent or trademark). This procedure allows us to estimate, for each incorporated firm, the probability that a firm will achieve a growth-oriented outcome (i.e., all firms have some probability of achieving a growth outcome, but that probability varies significantly across firms, based on observables at the time of incorporation). This calculation then serves as the foundation for nowcasting and placecasting growth entrepreneurship. In our placecasting application, our approach allows us to track the changing locational patterns of growth entrepreneurs over time; in Massachusetts, we are able to document the transition from Route 128 growth entrepreneurship to clustering in Kendall Square and Boston. One advantage of our approach is that we can characterize growth entrepreneurship at any level of aggregation, from individual addresses to the level of a city or region. We are also able to nowcast growth entrepreneurship, and develop an index of current-period growth entrepreneurship activity based on the number of current-period incorporations that have observable attributes that are similar to companies in the past that have had a high likelihood of achieving a growth outcome.

1. Introduction

While it is trivial to find firms that become large, stay small, or dissolve, by looking at their financial information, it seems impossible to be able to do this prediction at the onset, when firms are still young and such result hasn't realized. There are many reasons for this, particularly, that unpredictable and uncertain events which cannot be accounted for would happen after firm birth (Knight, 1921). Public markets efficiently solve this problem by repeated transactions of shares of firms, most of the time leading to an efficient outcome, but such is not possible for private, new, firms. These problems in uncertainty and lack of liquidity are important and difficult to make progress on, but they are not the only reason why young firm's growth potential is not yet well characterized. There are at least three other reasons in the area of economic measurement, unrelated to uncertainty, on which, in principle, progress could be made. First, most measurement today requires a history to create information: at firm birth, even if firms are already different, most of the information used to differentiate firms – like TFP, industry, headcount, investments, trademarks, or patents – does not still exist and still needs to be acquired throughout the first years of firm life. Second, there are data lags: even if something could be inferred from the firm at birth, that information might take time to “trickle down” through data systems to be available to the researcher making most datasets have considerable lags between the current year and the latest year a researcher can observe. And third, data is far from comprehensive: even in the best datasets, firms might be required to achieve certain thresholds like having one employee before being observed therefore not being clear whether firms observed “at birth” at really at birth¹.

But, recent advances in the use of computers to digitize and process records allow progress to be made. We propose a new methodology and dataset, based on public records, that together make a significant step to address known measurement issues to characterize the growth potential of firms by only observing information available early in their existence. In some cases, we are able to use only information available right at the moment of founding, before any sales

¹ This is an issue with many of the most comprehensive research databases. Even the most comprehensive database, the US Census Longitudinal Business Database, requires firms to have at least one employee before recording them, and it is not clear to researchers how long it took from firm birth to one employee. Further, there is a lag of several years (about 3) between the data available in the Census and the current year.

or hiring, and are still able to characterize the heterogeneity in firm growth potential. By using a data driven approach, our methodology allows us to score firms by their growth potential without introducing researcher bias, about whether some firms are growth firms while others are not. Further, our significant use of public records highlights new datasets that are easily available to most researchers in firm dynamics, which are highly comprehensive and can be easily accessible without any security clearance and at a low cost.

While most economic models assume a profit maximizing firm (add something), most firms have other incentives besides profit or growth (Hurst and Pugsley, 2011). Our methodology does not assume a profit maximizing firm. In fact, by looking at what firms do at the time of registration (e.g. deciding to register in Delaware instead of their home state) can be understood, to some extent, to capture their intention of growth, rather than potential, but that is also correlated with later growth in the sample.

The datasets and methodology we highlight are most useful to scholars in economics of entrepreneurship, industrial organization, and management, and it has the possibility of applications to policy makers to optimize the portfolio of financial support to small businesses with an observable potential for growth which is still not reflected in sales growth or other later metrics.

2. Literature Review

The most ambitious effort to develop a foundation for researchers of firm dynamics is the US Census Longitudinal Business Database (LBD) (Jarmin and Miranda, 2002). The database uses tax information to create yearly profiles of all firms in the United States at the establishment level. Its longitudinal nature, length of the panel, coverage of all non-farm firms in the private economy with at least one employee, and depth of information from a high-quality source (firm tax records), has allowed substantial research in industrial organization, finance, productivity, economic geography, entrepreneurship, and others. In entrepreneurship, notable recent work including clusters (Delgado et al, 2009), employment contributions (Haltiwanger et al, 2013), and financial constraints (Nanda and Kerr, 2009), to name a few, has been done with this

database. As this is still only financial data, other studies have tried to get to the entrepreneurs themselves, their attitudes and aspirations, in the United States, most notably the Kaufman Firm Survey, a small panel of 5000 firms of all types that starts before the firm is founded. Since the LBD is highly specific to the United States, cross country research has tried to use other databases which, though providing less depth, it allows for more comparable results between countries. Two examples are, the World Bank entrepreneurship database (Klepper et al, 2008), which records the number of limited liability business registrations in 139 countries in a yearly panel, and the Global Entrepreneurship Monitor (Acs et al, 2006), which surveys at least 2000 individuals about their attitudes and aspirations in entrepreneurship each year in 100 countries, representing 89% of the world GDP in 2013. Other researchers do not try to cover all types of entrepreneurship, like the LBD, but a specific phenomena. Some researchers focus only on innovation-driven entrepreneurship and as such have found venture capital databases, particular Thomson Reuters VentureXpert, as a good sample², while many others have had to build their own datasets (a few examples of building their own datasets are Hsu, 2004; Kaplan et al, 2009; and Lerner and Malmendier, 2013).

While all of these databases have contributed to the study of entrepreneurship, choosing any of them presents a substantial tradeoff between coverage, quality and accessibility. The datasets that have high international coverage only offer information at the aggregate level or spread out thinly in the population, those with high US-specific coverage are challenging to access due to privacy concerns and might require working in specific locations, and the phenomena specific ones select often select on a variable (like venture capital) that is only specific to a few industries and geographies, or they might be quite laborious to put together. The new dataset we highlight, business registration records, makes progress on each of these dimensions. It has quasi population level coverage with an extremely long history - all business that ever register in a region³; it is easy to access, as these are public records it is often trivial to request them or just download them from the appropriate entity⁴; and it is easily replicable in any

² Many researchers in innovation based entrepreneurship have also found it worthwhile to create their own sample. See Hsu (2004) and Kaplan et al (2009) for two notable example.

³ In Massachusetts, we are able to see, for example the record for MIT, founded in 1851.

⁴ In fact, we have been able to collect data for six US states, California, Florida, Texas, Massachusetts, Washington, and Vermont, and the full country of Mexico, without much work.

Western economy in the world where limited liability firms have to register in a public records office. Compared to the LBD specifically, this database, while offering less depth, is substantially easier for any researcher to access. This is no small improvement as the complexity of accessing Census research centers is laborious. Also, it allows researchers to have a full list of all registered firms (not only those that hire an employee), and to establish a true date of birth of the firm (instead of when they hire their first employee).

The main contribution of this paper, however, is not highlighting a dataset of the methodology to use this dataset, by observing that there are signals a firm sends at founding that how does that relate to later stage growth. Several research papers have already identified aspects of new businesses which could be, at least in principle, observable or evaluated at firm birth and related to further economic growth. For example, Gompers et al (2005) find that founders who themselves come from venture backed public firms are more likely to raise venture capital, and Rajkamal and Schoar (2008) find that ethnic signals relate to future business outcomes. If we are able to observe characteristics of the firm founders, we might be able to make progress to characterize firm potential. There is also a long literature dating at least to Schmalensee (1984) on the highlighting the importance of industry in firm profitability. Kaplan et al (2009) find that industry is still relevant for high growth young firms⁵. In as much as we can observe industry at birth, we can further include that in our available information of firms. Location has also been known to matter for firm performance and its effect on young entrepreneurial firms has been considered since Chinitz (1961) and more recently updated by Delgado et al (2010), and Glaeser et al (2012). Further, theoretical models have long emphasized the importance of innovation to firm growth (for instance, Aghion and Howitt, 1992), and any measure of innovative activity early on in a firm's existence might be a useful indicator.

Our approach uses some of these insights from the literature, and looks for proxies to build a predictive model of firm high growth potential. We then show how this allows us able, to an extent, to characterize high growth firms by their founding characteristics and how the patterns of these firms are different than all firms. Importantly, we can characterize growth

⁵ Kaplan, Sensoy, and Stromberg (2009) actually do their analysis for venture backed firms instead of a more general group of high-growth firms.

potential for very new firms, before they display any financial outcome. While many researchers acknowledge there are many types of firms, to our knowledge, this is the first work that is able to provide a systematic quantitative approach to separate the set of high growth firms from a sample, and do so even easily when the sample is extremely large (in our analysis, we work with more than 400,000 firms).

While we just highlighted multiple regressors related to growth, we do not include all possible regressors as predictors. Specifically, we do not include location, ethnicity or gender effects on our regression, even while we have an expectation that the effects of growth from these parameters would be significant. The reason is that our application tries to capture the variance in innovativeness that is not directly related to these effects, we might be able to observe some of the changes that happen endogenously. Our geography examples in this paper highlight that. Similarly, if we concluded the effect was driven by ethnicity or gender we might not be able to observe the full effect of innovation-based variables, which is what we are trying to get to (even if some of them could have different means at different genders). We must stress, however, that these are decisions we made for our specific application but that these variables could be included in other cases where it was considered relevant.

3. Business Registration - who registers their business and why?

Business registration is the act of formally registering a company in the official government business records office. In common law countries, registered businesses are, generally, of two types: corporations and partnerships. Partnerships are the simplest type of social organization, and it simply means a group of people organizing for a specific interest. Business partnerships are created from the organization of people to execute on a specific profit motive. Partnerships have emerged organically in most societies⁶. Depending on the features of the partnership it can be of different types. General partnerships are a specific type of business partnership. They have two notable features: (1) all partners share equal rights on the decision

⁶ Legal historians Crane and Bloomberg (1968) note that partnerships as profit seeking arrangements date back through Babylonia and continue through classical Greece and Rome.

making process of a business, and (2) all partners have full liability on any litigation aspects the business might be involved in. In general partnerships decisions need to be made in by consensus and there are no often detailed governing laws that can help resolve disputes.

General partnerships are the original form of business organization, and provided a framework for businesses to organize throughout history, but it was a type of firm that catered well only to businesses run by owner-managers. With the advent of mercantilism in the early 1600s, a new business form started to be needed that catered to investors and allowed them to exist as a separate group from managers, and hence the corporation was born. The first corporations were chartered (by a royal charter) to lead colonial ventures, notably the British East India Company, the Dutch East India Company, and the Hudson's Bay Company. From the beginning these early corporations already offered two advantages over partnerships which will persist through time: investors were given paper certificates as proof of their ownership, which they could sell to others, and they were explicitly granted limited liability (Prakash, 1998). The ability to sell and buy shares became quickly important and, by 1719, this feature precipitated the South Sea Bubble, the first stock-driven market crash in history, caused by increasing speculation on the South Sea Company's shares. After the South Sea Bubble, the Bubble Act of 1720 was established prohibiting corporations to form without a royal charter. It wasn't until 1844 that corporations were allowed again to form independently and limited liability was provided again until 1855, stipulating that shareholders are still liable to creditors, but only for the unpaid portion of their shares. In the United States, it wasn't until the late 1800s that a process similar to that of UK after 1855 became possible.

Besides corporations and partnerships, there is another type of organization in the US, limited liability companies (LLCs). Limited liability companies, however, are a quite recent phenomenon created to offer the best of both partnerships and corporations to owner-managers that seek limited liability but do still want to be managed as partnership⁷. LLCs offer limited liability to managers, and allow investors to also act as managers, while, like partnerships, they offer pass-through income to avoid double taxation, are managed under the one partner one vote

⁷ Though a recent phenomenon in the United States and other common law countries, an equivalent to the LLC has existed in Germany – Gesellschaft mit beschränkter Haftung (GmbH) – and other civil law countries since the 19th.

principle, and do not offer tradable shares. The first state to allow LLCs was Wyoming in 1977, then Florida in 1982. Today, LLCs, corporations and LPs are the three existing types of registered companies in the United States. Though there does exist further classification at the tax code level, it is important to keep in mind that these are tax classifications, rather than legal statuses. The main tax classifications are non-profit corporations, C Corporations, and S Corporations, the S Corporation being a small business tax status that allows owner-managers to avoid double taxation.

Still, it is a fact that most businesses are not registered businesses. The alternative to a registered firm is a sole proprietorship. A sole-proprietorship is any individual that makes income by themselves. In the United States, any individual who makes an income outside of a paid job and files a Schedule C is considered a sole-proprietor. It is the most common type of business organization. While common, there are at least six clear advantages for having a registered firm instead of a sole proprietorship:

1) Limited liability is one important one. Through a registered company managers can have less exposure to business risk than through a sole proprietorship, as any claims that exist toward the company will not affect the personal assets of the owners. This allows owners to take on businesses with more inherent risk knowing that only the capital that they have already invested in the business is at risk⁸.

2) The life of the business is also more secure in a registered business. As these become a separate legal entity, they exist beyond specific founders, and the business can continue after the death or inactivity of one founder, which is not the case in a sole proprietorship. This has direct consequences in terms of the risk that potential clients face when buying from this firm, as the relationships and product support that they invest in has less risk and a longer expected horizon in registered firms.

⁸ As was remarked to us, potentially, this could also increase a firm's borrowing costs since there is less liability a bank can claim in the case of unpaid debt. However, the owner has full control on this situation, as they can offer specific personal assets as a co-signer of a loan. For example, if it was really a better option for them, they could be (as an individual) a co-signer of the loan and include their house as collateral.

3) In the case of both corporations and LLCs, liquidity is also a benefit for owners – i.e. they have companies they can sell. Corporations have a straightforward process through which owners can make their investment in the company liquid, by selling their shares. Partnerships (LLCs) also allow for that even if the process is not a simple. In terms of liquidity, corporations can also use these shares to get capital, as investors can buy shares of the company, as do venture capital funds.

4) The fact that the firm is registered and not a sole proprietorship also serves as a signal for customers. Firms are required to state their type in their name. For example, in Massachusetts, corporations must say “Corporation”, “Incorporated”, or “Company” or an abbreviation of that in their name. Having “Inc” in the name could serve as a signal of quality to customers.

5) Registered firms are also allowed to document governing rules and operating procedures in the case of multiple partners, and allow for these partners to work together in the same entity. In the case of corporations, it also provides a mechanism for oversight through the corporate board and offers clear guidance on what minority shareholder right could be.

6) Finally, the specific details of US tax law offer tax benefits to firms over sole proprietorships that further the incentives to register the firm. I offer two examples of these. Tax deferral policy allows registered firms to carry forward any amount of losses in income for any number of years, but long-term capital losses are capped at \$3000 for sole proprietorships. Similarly, registering a business allows firms to take advantage of the small business tax deduction.

All of this, we suggest, creates strong incentives for any entrepreneur to register their business, either as a corporation or as an LLC. Added to it, the cost of registering a business is cheap, for example, only \$89 in Delaware. Given that it is still a fact that many people file taxes as sole proprietors, it naturally leads to questioning the counterfactual: who, then, is not registering their business. It is easy to see that it must be people that do not benefit from any of these features. These are people who are not interested in having partners, do not need investors,

take little credit and therefore do not benefit from reduced collateral exposure, hire none or very few people and therefore do not benefit from reduced liability, do not expect to incur any big losses in any year that they can carry forward in taxes, whose clients do not benefit from the longevity of the firm beyond the entrepreneur, and who do not find value in legally document the mission and governing structure of the firm through articles of organization and company by-laws. Sole proprietorships are likely so common because many self employed individuals fall in this group, including the consulting work of university professors, independent graphic designers, journalists, painters, landscapers, babysitters, and many others. However, once a business goes beyond a single professional charging for their individual work, it is easy to imagine that each of the features outlined pushes the incentives to register a business quickly.

But, even if it is true that incentives quickly raise for more complex business than independent individuals, this does not suggest that there is a strict breakpoint after which firms to registered. We cannot say that all registered firms are fundamentally different from all non registered firms. Yet, we do show that the marginal firm that does register is very far away from our selected group of growth firms, therefore suggesting growth firms are not at the margin – and thus we are unlikely to miss them. It is also possible it is still the case in other countries, where, for example, small business credit could be non-existent making founders always use personal savings. Our results and argument also lead us to assume that those sole proprietorships that might find themselves growing, even if not by much, will also quickly decide to register as they need to hire people, manage finances separate from the business, and take certain types of credits. For example, convenience stores, dry cleaners, and pizza shops would benefit from registration because these businesses often need to borrow large amounts (relative to the private wealth of the owner) for machinery or inventory. These loans are easier to get since they are implicitly collateralized by the inventory and machines themselves. We find that these types of business tend to be registered businesses in the period of 1995 to 2013 in Massachusetts.

4. Econometric Methodology

In our econometric specification, we predict a firm's outcome of growth from the information that we can observe from business registration records or other data which is

generated close to that time. While in principle we could predict any continuous outcome, we settle instead for simple binary outcomes of growth and predict the probability of achieving that growth for a firm. In our most comprehensive specification, we consider growth an IPO or a merger with a valuation of over \$10 million. This approach of regressing a later outcome with firm characteristics early on requires that we are rigorous and consistent in the time allowed to achieve such outcome. We estimate the following stacked logit model for all firms where enough time has elapsed:

$$P(\mathit{growth}_{i,t+k} | \mathbf{X}_{i,t}, \mathbf{Z}_{i,t}) = \alpha + \beta' \mathbf{X}_{i,t} + \gamma' \mathbf{Z}_{i,t} + \varepsilon_{i,t}$$

Where $\mathbf{X}_{i,t}$ and $\mathbf{Z}_{i,t}$ are characteristics from business registration data and other sources, respectively, for firm i in cohort registered in year t . This is information that is available at time t but which we believe could be correlated with later growth outcomes. $\mathit{growth}_{i,t+k}$ is a binary outcome observed k years after business registration. The specific value of one or zero varies depending on each of our specifications. In a cohort by cohort setup like the one we have there is an intrinsic balance in selecting the value of the time lag k . While a large k allows more years for firms to achieve the growth outcome and thus have our sample look like the eventual long term realization, a large k also causes the sample to reduce since, as it increases, there will be more and more firms for whom the necessary time window has not elapsed. Our approach to select k is to use the 1995 data, where we have outputs up to 17 years later, and look at the distribution of IPO and acquisitions in those 17 years for the cohort. We conclude that a value of $k=6$ strikes a good balance between comprehensiveness and value. We use robust standard errors though we also try clustering at the year level and there is no difference in the significance of any result.

5. Data

We build our dataset using the business registration records from the Massachusetts Corporations Division, and then merge that dataset by name and location of firm with the USPTO patent and trademark database and Reference USA yearly files. The business registration database contains all firms that have ever registered in Massachusetts, starting from

as early as the 17th century for some cemetery corporations⁹. From business registration records we record the founding year, name, main office location, entity type, and jurisdiction. We drop all firms incorporated before 1995, those whose main location is not in Massachusetts, and those whose main jurisdiction is another state than Massachusetts or Delaware¹⁰. Since some evidence suggest high-growth firms might change jurisdiction to Delaware when approaching some liquidity events like an IPO, we connect those firms that merge from a local jurisdiction to a firm in Delaware as one firm. We merge our business registration information with two other data sources. From the United States Patent Office (USPTO) we get both trademarks and patents and Reference USA yearly files¹¹ for employment and sales outcomes.

We start by presenting summary information on our Massachusetts data. We also present information for another geography that we used in analysis but we do not run in our regressions here, the San Francisco Bay region. We show this with the goal of allowing researchers to see a preview of what another area that is highly innovative could look like as well as a state that is much larger in its GDP share. We also note that this analysis could be done by merging multiple states and without having to pick one. Figure 1 shows the total number of local firms registered in each region.

Table 1 contains a breakdown of count by type of firm in the two regions we use. Since each registry can vary in the way they catalog firm, the type and detail of firm can vary and we can see there is a lot more detail in the data from Massachusetts than California. It contains a mix of domestic (local) entities, which register as a firm in the state, and foreign entities, which are firms registered originally in other states but who open business in the ones we analyze. It also includes a set of non-profit and religious or civil associations (e.g. cemetery corporations). We drop all non-profit and religious associations before analysis. We also drop all foreign firms

⁹ Corporations and their nature has changed substantially through the years. Early on they where only used for cemeteries, then required state legislature to be approved and became the main mode of business organization in the twentieth-century. In the last decades, corporations have been superseded by the newly developed limited liability company (LLC).

¹⁰ We keep Delaware to account for anecdotal evidence that many high-growth firms register in Delaware as opposed to their local state. For example, there is anecdotal information that venture capitalists will require firms to register in Delaware before providing them with funds.

¹¹ Reference USA is also known as Infogroup, a close competitor of them that is more commonly used in research is Dunn and Bradstreet. For our purposes, the Reference USA dataset can be considered an equivalent of Dunn and Bradstreet.

registered in a state that is not Delaware. Delaware poses a particular problem since firms that seek institutional investors are often required by venture capital funds or other to register in that state, even if business done in another. While these firms are a small percentage of the total, they are often highly innovative, and in dropping them we could also drop an important part of the growth firms in the state. Therefore, we do drop all Delaware firms that have a registered main office address in another state, but keep all those whose main office address is within the state.

We build five regressors from the business registration data: a dummy equal to one if the firm's jurisdiction is Delaware, a dummy for eponymy, a dummy equal to one if the firm's entity type is as a corporation and not an LLC or a partnership, a continuous measure between 0 and 1 for innovativeness in the name, and a set of simple industry dummies. We consider a firm eponymous if either the first or last name of the president or the person managing the firm¹². For entity type as corporation or LLC (there are very few LPs), we have a challenge in our dataset, as LLCs only became possible in 1996 in California and 1994 in Massachusetts and their share as total of firms has kept steadily growing since. We do not make any effort to control for this. Therefore, when we move to our results and find a significant positive coefficient associating corporations (and not LLCs) to growth, this coefficient will be calculated from a model with early data (up to 2005), but we expect the division between LLCs and corporations to be more useful at separating growth firms today since, as outlined in earlier, the incentives align in such a way that only the most growth oriented firms are likely to become corporations. Also, we should note that our treatment of corporations, LLCs and LPs differs from the analysis in entrepreneurship done with tax records including the Longitudinal Business Database (Jarmin and Miranda, 2002), which do not allow for LLCs (see footnote)¹³. Similarly, the tendency of firms to register in Delaware has changed through time and the influence of the estimated parameter is likely to be stronger with more recent data. Figure 2 shows the yearly share of LLCs and Delaware firms in both states.

¹² While Massachusetts calls the individuals managing an LLC “managers” the term should not be considered similar to the business term manager, often used for people in the middle of an organization. It is instead an operating partner of the firm.

¹³ Specifically, in terms of tax treatment, LLCs are allowed to choose whether they want to be taxed under a corporation regime or a partnership regime, and therefore appear as either one or the other in tax files.

Our measure of innovativeness in name is built through the Naïve Bayes natural language processing algorithm, training with a sample of venture and non-venture backed firms as innovative and non-innovative and then predicting for each new firm. To exclude selection effects, we make sure to train prediction for firms in year t only with firm events occurring in $t-1$, that is, only events before the firm was registered. Appendix 2 contains more information on how we build this measure, and Appendix 3 contains information on how we build our industry dummies from name.

We match our data to patent assignment data. We use a name matching algorithm with location filters, which is virtually the same that Balasubramanian and Sivadasan (2008) and Kerr and Fu (2008) use to match the business registry and the Survey of Industrial R&D to the NBER patent database, except we only allow exact matches after multiple types of name cleaning. We decide not to use the NBER patent database but instead build our own by downloading the all assignment XMLs from the Google USPTO data stores, which contains all assignments between 2002 and 2013. Building our own dataset allows us to improve upon previous approaches and patents in a way that is specifically tailored to our question. First, the NBER database ends in 1999 while we are able to use the most recent data in prediction, therefore being able to understand more current patterns of entrepreneurship. Second, the approach of Balasubramanian and Sivadasan (2008) filters with MSA location of inventors, while our dataset allows us to use the location of the assignee specifically, leading to possibly better matching. Finally, we are able to track the same patent as it is reassigned to multiple firms and, therefore, can observe firms that get assigned a patent even if they don't hold it today. This is the most important aspect for our purposes: since the NBER patent database does not have the specific date on which the patents are assigned, we would not be sure if observed patent activity was early stage activity, by looking at assignment dates of applications, we can fix the problem. For all patents assigned, we match 80% of them to a firm, the same number as Balasubramanian and Sivadasan (2008). Also, while not matching all patents will bias our estimate, we can only expect an attenuation bias from missing some patents.

We do the same process for trademarks from the Google USPTO trademark files. We then create four dummies which take the value of one if the patent (or trademark) was assigned

between 0-6 months after the business was registered, or 6-12 months. In this way, we guarantee we are only observing early stage innovation instead of confounding other elements.

Finally, for our output variables, we use Reference USA files and SDC and perform the same matching process as with patents. Reference USA is a firm (a close competitor of Dunn and Bradstreet) that collects information on other firms to be used by marketers and researchers, the files we use are snapshots of the Reference USA database at the end of every year. Using yearly files give us the benefit of being able to control the time elapsed from founding to outcome as we need, particularly making sure we are allowing firms at different founding years the same time to achieve the growth outcome. In our matching tests we are able to match 51% of all firms with 5 or more employees in Reference USA. As a quality comparison, Balasubramanian and Sivadasan (2008) get 65% of firms in their process of matching patent assignees. Since we have a more stringent match that only allows exact and not fuzzy string match, and since we match firms on location while they allow multiple locations (one for each inventor), we believe the efficacy of our algorithm is largely comparable to theirs but has more stringent filters. When we consider growth outcomes we code as 0 all firms that do not match to a Reference USA firm. Finally, for all our matching, we consider valid all the names a firm has had in its history. For mergers, we include the names of all firms acquired but do not include the name of the acquirer in merged firm's history. Table 2 shows summary statistics for our Massachusetts data.

6. Firm Growth Regressions in Massachusetts

We run two models of firm growth in Massachusetts. In the first model, our dependent variable is a dummy with a value of 1 if a firm gets 100 or more employees in six years. In the second model, it is a dummy with the value of 1 if a firm has an outcome of an IPO or a merger with a valuation over \$10 million USD in at most six years after founding. We use data for years 1995 to 2005. We truncate at this year to allow us to observe the outcome six years later in the Reference USA and IPO/merger files.

Table 3 has the results for the employment regressions. We run three models. Models (1) and (2) only use information available in the business registration record, while model (3) includes USPTO information for patents and trademarks. In each model, the first column is the regression coefficients and the second is the marginal effects. The base probability is at the bottom. We do want to stress that this research is not focused on understanding causality of any effect, and does not make any statements about causality, rather it is an empirical assessment of different signals which because measurable at birth of the firms, or close to it, and what can they tell us about the firms and their eventual outcome. We use only four covariates in the first model, all of which turn out to be significant. With only these four covariates, and a base probability of high growth of 2.1%, we find that firms registered in Delaware increase 2 percentage points the probability of achieving high employment, and eponymous firms reduce it by 1.4 percentage points¹⁴. We create the name innovativeness by training on the Reference USA files as they existed the year before a firm was founded. While this could implicitly have a bias of that a firm that is scored could already be in the Reference USA, we believe we control for this by using a file that was created before the firm registered in the corporations office. Since our first file starts in 1997, we start our regression with firms in 1998. Also, because this probability is less skewed than the VC one in the IPO or merger outcome, we use machine-learned score directly as our innovativeness score. The marginal effect implies that moving from the 10th percentile to the 90th percentile in name innovativeness would increase success probability by about 2.1 percentage points. Finally, we find corporations are one percentage point less likely to achieve high employment. This is surprising contrast to our IPO or merger model, where corporations are positively associated with the outcome. We do not present any reason for this to be the case.

In model (3) we add dummies that are 1 if a firm registers a trademark or is assigned a patent early in the process. We see that all coefficients stay very similar except for Delaware jurisdiction, which goes down by one percentage point, almost of all of which is then picked up

¹⁴ Our results for eponymy differ from Chatterji and Belenzon (2013) for multiple reasons. First, we only look at new firms in a recent time period while they look at all firms, and it appears that the highly successful eponymous firms (Ford, HP, Dell) were mostly founded in that a previous period. Second, they establish a difference in profitability for the average of all firms, in new firms, it is entirely possible that eponymous firms are more profitable on average but less likely to achieve very high levels of growth. Finally, our sample is for Massachusetts while theirs is in Europe, and these signals could mean different things in each region.

by registering a trademark in the first six months. Registering a patent within the first year appears to have about an equal effect of registering a trademark (1.5 percentage points) though it is a little less robust for patents.

Table 4 does the same analysis with a different outcome variable: a dummy that is one if a firm has an IPO or a merger with a transaction value of over \$10M dollars. In this case, the name innovativeness variable is trained by using venture capital events that happened the year before the firm was founded. In this case, this variable, while very skewed, is standardized with a mean of zero and standard deviation of 1. Many of the coefficients are quite similar in magnitude of the marginal effects. In particular, a Delaware jurisdiction increases success probability by a little over 2 percentage points, which goes down to 1.3 percentage points when patents and trademarks are added. Patents and trademarks applied for within the first year both add 1.5 percentage points. With the base probability being around 0.7%, any of these effects is quite substantial. The effect of eponymy is weaker and, different from the employment result, being a corporation positive and significant, adding between 0.54 and 0.38 percentage points to the predicted probability of success. For the name innovativeness, moving from the 10th percentile to the 90th percentile adds 0.2 percentage points. However, it is highly skewed and mean marginal effects are hard to interpret. Table A1 shows the value of each of the name innovativeness measures at different percentiles.

7. Prediction with Massachusetts Regressors

Besides estimating the importance of a signal, our logit model also allows us to do prediction on firms. We investigate the results using predicted probability of growth on our sample of firms. We analyze three aspects, contribution and distribution of growth, prediction and regional patterns.

Before moving into specific analyses, it is important to be clear about the meaning of the predicted probability of growth in each firm. Given that we run a regression with growth as a dependent variable, the predicted probability is the probability that a firm will achieve growth based on observables at the time of founding. We propose this value can be interpreted as the

firm's growth potential. This approach has at least three advantages. First, instead of making the researcher select a small sample of firms with growth potential we allow all firms to have some growth potential, even if that potential might be very small for many. Second, it allows us to score a firm before many later actions happen in the "real world". And third, econometrically, the expectation of the number of growth firms in a sample can be inferred as the sum of the probabilities of the firms in that sample.

Distribution of growth

We first look at the distribution of growth potential in both models from 1995 (1998 for employment) to 2005. We order our data by growth probability, divide it in bins of 5% size and get the sum for each bin. Then show two separate graphs, the sum for each bin and the Lorenz curve for cumulative distribution. This analysis helps us understand the heterogeneity in growth potential. If all firms have the same growth potential, then we should see the cumulative distribution grow in a straight 45 degree line while the share of each bin should be roughly constant at 1/20. In a population with heterogeneous potential, we should see the share vary significantly between the lowest and the highest bins and the cumulative distribution be under the 45 degree line. As in our previous analysis, we use the full sample for the IPO and merger outcome and restrict the employment analysis to the matched sample between 1998 and 2005.

Figure 3 has the graphs for both outcomes. As might be expected, the IPO/merger outcome is a little more skewed than the employment outcome, though the results are very similar. In the IPO/merger model, which is the one that uses all firms ever registered in Massachusetts in the 11 year window, we find that the top 5% of the firms contributes 36% of all growth potential in a region at founding, and the top 10% contributed 45% of the growth potential. We believe this is the first comprehensive empirical result showing the heterogeneity of firms at founding, and the differences in potential that sets of them can have in the economy.

It is also of interest to consider the distribution within the top 5%. Given that the incidence of our growth outcomes is much lower than 5% (the number of firms with IPO or merger is only about one percent), it is entirely possible that there is still substantial

heterogeneity within the top 5%. Figure 5 shows the share of growth predicted within the top 10%, by separating firms into 0.5% bins. Of all the growth in the economy, 20% is carried out by the top 1%. It is important to note that this is not a small number of firms such that any firm could bias the estimate. While 0.5% bins might sound like a very small set, our large sample of firms (251,726 between the years 1995 and 2005), allows us to still have more than one thousand firms in each bin. With the number of firms registered each year hovering around 30000, a researcher that wanted to focus only on the top 1-percent firms by potential, registered in the last three years might still need to focus in as many as one thousand firms. Compared to venture backed firms, this is roughly triple the amount of venture backed firms in VentureXpert in the state of Massachusetts in the last three years. This is a substantial amount to do research with as well as any other type of analysis.

8. Predicting Growth Probabilities on Firms

One main application of our methodology is be to predict a growth probability for new firms in a region as they emerge. In fact, our predictions allow researchers to do two things previously not possible at a similar population-like level. A prediction allows researchers to create a score of growth potential before a firm faces “real world” events. If done to firms where the growth outcome is already realized, we can start separating potential early on and the effect of later events in the realized outcome. If the prediction is done on firms where the outcome is not realized, we can have an statistical expectation of the outcome. While this might miss many firms with real potential, it would certainly allow analysis to differentially target firms for policy intervention in ways that are quantitatively rigorous and can look at all firms at the same time.

We do two types of analysis, regional and firm-specific. In our regional analysis we estimate the number of growth firms in small regions inside Massachusetts. We select to aggregate firms at the zipcode-year level and see how the density of firms moves geographically. In particular, we are interested in seeing if our methodology is able to track the geographical shift of innovative entrepreneurship from Route 128 to Cambridge and Kendall Square. In our firm-specific analysis, we are interested in finding specific firms with growth potential and separating them from the rest. There are multiple ways to do it. We do all further analyses with

the IPO or merger regression as that allows us to use the full population of firms. For the firm-specific analysis, we establish a cutoff value the split growth and non-growth firms. This value is the probability value exactly at the 95th percentile in the pooled 1995 to 2005 estimates. This approach allows us to have a different number of firms in different years as the cutoff is the same for all years and estimated from empirical evidence of firms before the more recent years ones. Other researchers using this methodology, however, could prefer to use the top 5% of firms in each year if the analysis they are doing suggests that this is a better approach. For the regional analysis, we have three options to characterize growth potential in a region. We could sum up the probabilities of all firms in each region-year to get an expectation of growth firms in the region, we could do an unweighted count of the number of firms above a cutoff, or we could do we probability (potential) weighted count of the number of firms above a cutoff. We choose to do an unweighted count for consistency in methodology and because the results of the three options were virtually the same. While we would have liked to be more stringent with the cutoff (i.e. have an even higher cutoff) we are constrained by the number of firms that we can find above the 95th percentile cutoff in only a few cities. Therefore, we preferred a cutoff with enough firms to believe there is a pattern than only a few (the number of firms per year in either Cambridge or Route 128 above a 99-percentile cutoff is never more than 20).

Regional Analysis

For this exposition of our methodology, we track the number of growth firms in two regions in Massachusetts, Route 128 and Cambridge. Our goal is to see if we can see a growth in growth entrepreneurship in Cambridge and a decline in Route 128, which anecdotal evidence suggests. While it is somewhat arbitrary to decide where exactly Route 128 ends, we consider Route 128 all firms in the cities of Waltham, Wayland, Lexington, Burlington and Woburn, which are, in a sense, the “core” of Route 128. However, we did test for robustness by adding further cities including Maynard, Lincoln, Weston, and others without any real difference in results. We present all our results for the ten year period of 2002 to 2013. While more years are interesting the dot-com bubble in the late 1990s is a different phenomenon and this exposition allows us to better show the post 2000 patterns that we want to highlight or analyze. While we decide to leave in the year 2013, we should mention that 2013 firms are calculated without patent or trademark

information as patent applications are not public until 18 months after application and the trademark datafiles only go until 2012.

Figure 6 has all our graphs in this regional analysis. We first show all firms registered in both locations, Cambridge and Route 128. We can see that business registration has increased in both locations, as has for the whole state. Though somewhat faster in Cambridge, reflecting urbanization patterns of the last decade, Route 128 still registers more businesses than Cambridge in any year. The second graph shows only the count of high growth firms (firms above the growth potential cutoff), it is apparent that the locus growth firms has changed. While in the beginning of the time period Route 128 substantially surpassed Cambridge in growth firms, the opposite is true in the later years. There is a consistent drop in innovative businesses in Route 128 while there is a consistent rise in innovative businesses in Cambridge.

These two results highlight the importance of looking at the correct set of businesses when making inference about different types of entrepreneurship. While Route 128 has more businesses created over the period, and there does not appear to be any much shift in the gap between Route 128 and Cambridge when looking at all businesses, the pattern is strikingly different when we look only at high-growth entrepreneurship. While we do not imply that we are the first ones to observe this geographical movement of growth firms our result is, as far as we know, the first empirical way to track it for the whole economy.

We can also start looking within the city for the locus of innovation. We use the zip code as a unit of analysis in our data, which comes directly from the registration records, but could also be possible to do much smaller geographies by using geo-coded data. We look for the dynamics inside Cambridge, specifically, how high growth entrepreneurship has changed between the two poles of innovation in Cambridge: Kendall Square (zip code 02142), in the vicinity of MIT, and Harvard Square (zip code 02138), the area around Harvard University. The last chart in Figure 6 shows these results. While the data is inherently noisy given the low number of events it shows a pattern where both locations were producing a comparable level of high-growth entrepreneurship up to 2010, when there is a displacement from Harvard Square to Kendall Square.

This is just one example of use of our methodology. In the appendix, Figure A1, we offer graphs from a second similar example, now in California, tracking the shift of high growth entrepreneurship from Silicon Valley to San Francisco. The results are very similar in nature.

Looking at Individual Firms

While there are many ways to assess the growth potential of a firm, most of them are time consuming. With 30000 firms registered only in Massachusetts every year, it is virtually impossible to assess all of them, or even a large percentage, due to the large cost of evaluating each one. The challenge would be even harder for the whole country. However, our methodology allows us to assign a growth score (the probability of being high-growth) to each of the firms in a systematic way and there is information to be learned by looking at those firms individually. We also believe it is important to show the actual firms predicted in growth or no growth as consistency check of our algorithm. For our predicted probability to be a measure of growth potential of firms, we would need to see firms that are clearly have a higher growth potential in the higher scores. We do two things to test this, both for Cambridge. We present a random sample of the firms above and below the cutoff, and we map the high growth ones and offer researchers an ability to look at the online map directly, for the 2011 cohort. The map is available at <http://web.mit.edu/jorgeg/www/highgrowth.html>

Table 5 shows twenty random firms above and below our high growth cutoff. While not everything can be learned from the name it is easy to see that the list above the cutoff represents more high-growth innovative firms while the list below represents more local firms. This results are highly congruent with the type of firms that you would expect either in a Schumpeterian framework (Aghion and Tirole, 1992) or in a local vs traded framework (Porter, 2000). Importantly, there is very little opportunity for exogenous differences. This list was selected for firms in the same city and year of registration, therefore the economy faced by all of them is very equivalent¹⁵. The difference in the lists is noticeable. For example, the first firm in the non high

¹⁵ In fact, added to that Cambridge is notoriously small at only 7.13 sq miles in size, therefore suggesting all local economic environment was mostly similar.

growth list is a corporation setup to manage a property (possibly between the property owners), while the first one for the high-growth is a venture backed firm (by venture capital group Flagship Ventures) developing new air delivery systems. One thing to note is that Cambridge has a high concentration of innovation-based firms which might not be the case in other locations in which this algorithm is applied, and it is possible that in many regions there are not enough firms above the cutoff to guarantee adequate sampling for a similar analysis¹⁶.

Figure 7 shows a snapshot of the map we developed to track these firms at the location level. We use an automatic cluster algorithm in a mapping technology called Leaflet and in this case we only show a snapshot of Cambridge, but a map of all firms that allows the researcher to see each one is available at our website. The area of Kendall Square is highlighted. The map shows similar geographical patterns to what we have seen in other sections of this paper, with a high concentration of growth firms in the Kendall Square area, and some on the neighboring Central Square, then many firms much more scattered throughout the rest of the city.

9. Conclusion

We offer a solution to two problems facing entrepreneurship researchers today. First, we highlight a publicly available dataset that researchers can use for the analysis of entrepreneurship, public business registration records. While it offers much less information than the state of the art dataset used today, the Census Longitudinal Business Database (LBD), it offers several unique advantages. First, it is a dataset that can be found in almost any geography and region including all states of the United States and almost every country in the world¹⁷, therefore allowing entrepreneurship research methods to be truly global and ease constraints stemming from issues with lack of data and transferability of approach. Second, it is easily available for any researcher that needs this information in the United States and, while most states charge a delivery fee, it is usually possible for researchers to get the full of a state dataset for a few hundred dollars in a couple of weeks. Third, as comprehensive as the LBD is, it only

¹⁶ In fact, Cambridge contributes substantially to the set of high-growth firms in Massachusetts. For the year 2011, while it contributed less than 3% of total firms in Massachusetts it contributed close to 12% of all high-growth firms.

¹⁷ As an example of this point, we have been able to download the data for all firms in Mexico relatively easily.

includes firms that employ at least one person, but there are likely to be many firms that do not employ anyone. And fourth, it allows the researchers to establish a true “date of birth” of the company where, since the age of business is a critical dimension to understand the dynamics (Davis and Haltiwanger, 1994), less noise from analysis would be included.

We then created a methodology to use this information collated with other public information. In fact, for the regressions we use the most almost all information, except for the mergers and IPOs, is retrieved from public sources and can be freely distributed. While collating these public datasets might appear daunting, our collation approach based on the one used by Balasubramanian and Srivadasan (2008) for patents appears to work reasonably well. We showed that, depending on the definition of growth, some covariates might change in magnitude while other stay roughly similar.

We then moved on to use a prediction from our regression as a score of high growth potential and showed that substantial understanding of a region can be performed with this assumption. While this analysis was only demonstrated we Massachusetts data for exposition, we have found our methodology to also work well at least with data for California, Washington, and Texas, which are the only other states we have tried up to now. We have no reason to believe that other states within the United States will not work as well, or other countries outside the US.

This methodology promises to be a useful start for entrepreneurship researchers that are looking for broad populations of firms and for policy-makers looking to understand entrepreneurship quantitatively at each region.

Bibliography

Acs, ZJ, P Arenius, M Hay, M Minniti -[Global entrepreneurship monitor](#) Executive Report, London Business School, Babson, 2004

Aghion, P., & Howitt, P. (1992). A Model of Growth Through Creative Destruction. *Econometrica*, 60(2), 323–351.

Balasubramanian, N., & Sivadasan, J. (2008). NBER Patent Data-BR Bridge: User Guide and Technical Documentation. Working Paper. Retrieved from <ftp://tigerline.census.gov/ces/wp/2010/CES-WP-10-36.pdf>

Bhide, A. (2003). *The Origin and Evolution of New Businesses* (p. 432). Oxford University Press.

Catalini, C., Lacetera, N., Oettl, A. (2014). An Analysis of Positive and Negative Citations. Working Paper.

Chinitz, B. (1961). Contrasts in Agglomeration: New York and Pittsburgh. *The American Economic Review*, 151(3712), 867–8. doi:10.1126/science.151.3712.867-a

Davis, S., & Haltiwanger, J. (1992). Gross Job Creation , Gross Job Destruction , and Employment Reallocation. *The Quarterly Journal of Economics*, 107(3), 819–863. Retrieved from <http://qje.oxfordjournals.org/content/107/3/819.short>

Delgado, M., Porter, M. E., & Stern, S. (2010). Clusters and entrepreneurship. *Journal of Economic Geography*, 10(4), 495–518. doi:10.1093/jeg/lbq010

Ericson, R., & Pakes, A. (1995). Markov-perfect industry dynamics: A framework for empirical work. *The Review of Economic Studies*, 62(1), 53–82. Retrieved from <http://restud.oxfordjournals.org/content/62/1/53.short>

Glaeser, E. L., Kerr, S. P., & Kerr, W. R. (2012). Entrepreneurship and Urban Growth: An Empirical Assesment with Historical Mines. NBER Working Paper.

- Gompers, P., Lerner, J., & Scharfstein, D. (2005). Entrepreneurial Spawning: Public Corporations and the Genesis of New Ventures, 1986 to 1999. *Journal of Finance*, 60(2), 577–614. Retrieved from <http://doi.wiley.com/10.1111/j.1540-6261.2005.00740.x>
- Haltiwanger, J., Jarmin, R., & Miranda, J. (2013). WHO CREATES JOBS? SMALL VERSUS LARGE VERSUS YOUNG. *Review of Economics and ...*, (1999). Retrieved from http://www.mitpressjournals.org/doi/abs/10.1162/REST_a_00288
- Hurst, E., & Pugsley, B. (2011). What do small businesses do? NBER Working Paper. Retrieved from <http://www.nber.org/papers/w17041>
- Iyer, R., & Schoar, A. (2010). Are there cultural determinants of entrepreneurship? *International Differences in Entrepreneurship*, (May). Retrieved from <http://www.nber.org/chapters/c8219.pdf>
- Jarmin, R., & Miranda, J. (2002). The longitudinal business database. Retrieved from <http://www.vrdc.cornell.edu/info7470/2007/Readings/jarmin-miranda-2002.pdf>
- Kaplan, S. N., Sensoy, B. A., & Stromberg, P. (2009). Should Investors Bet on the Jockey or the Horse? Evidence from the Evolution of Firms from Early. *Journal of Finance*, LXIV(1), 75–115.
- Kerr, W. & Nanda, R. 2010. "Banking Deregulations, Financing Constraints, and Firm Entry Size," *Journal of the European Economic Association*, MIT Press, vol. 8(2-3), pages 582-593, 04-05.
- Kerr, William R., and Shihe Fu. "The Survey of Industrial R&D--Patent Database Link Project." *Journal of Technology Transfer* 33, no. 2 (April 2008): 173–186.
- Knight, F. (1921). *Risk, Uncertainty and Profit*. Boston, MA: Hart, Schaffner & Marx; Houghton Mifflin Co.
- Klapper, L., Amit, R., & Guillén, M. (2010). Entrepreneurship and Firm Formation across Countries. ... *differences in entrepreneurship*, (May). Retrieved from <http://www.nber.org/chapters/c8220.pdf>

Lucas, R. J. (1978). On the size distribution of business firms. *The Bell Journal of Economics*.
Retrieved from <http://www.jstor.org/stable/3003596>

Porter, M.E, 2000, "Location, Competition, and Economic Development: Local Clusters in a
Global Economy." *Economic Development Quarterly* 14 (1), pp. 15-34

Schmalensee, R. (1985). Do markets differ much. *American economic review*, 75(3), 341–351.
Retrieved from <http://kendlevidian.pbworks.com/f/Schmalensee-1985.pdf>

Table 1

Breakdown by type of firm for years 1995 to 2013

Massachusetts			Bay Area		
	Count	% of Total		Count	% of Total
Domestic Profit Entities					
Domestic Limited Liability Company (LLC)	163,027	34.2%	DE Corporation	24,739	3.8%
Domestic Limited Partnership (LP)	8,031	1.7%	DE LLC	23,413	3.6%
Domestic Profit Corporation	179,189	37.6%	Domestic Corporation	313,807	48.8%
Professional Corporation	7,543	1.6%	Domestic LLC	280,442	43.6%
Other Domestic Entities			Nonprofit Corporation	532	0.1%
Nonprofit Corporation	29,174	6.1%	Firms in California but not in Bay Area	1,797,074	
Registered Domestic Limited Liability Partnership (LLP)	1,310	0.3%			
Religious (Chapter 180)	3,093	0.6%	Total	2,440,007	
Voluntary Associations and Trusts	2,662	0.6%			
Foreign Entities					
Foreign Corporation	28,916	6.1%			
Delaware firm in MA	26,192	5.5%			
Foreign Limited Liability Company (LLC)	25,037	5.3%			
Foreign Limited Partnership (LP)	2,222	0.5%			
Total	476,396	100%			

Table 2

Variable	Obs	Mean	Std. Dev.	Min	Max
Industry Realtor	481809	0.0541376	0.2262893	0	1
Industry Restaurant	481809	0.0089724	0.0942971	0	1
Industry Law	481809	0.0063511	0.0794402	0	1
Industry Dental	481809	0.0026068	0.0509907	0	1
IPO Date	480	14553.82	2191.158	10995	19766
Merger Date	6462	16146.53	2499.635	10975	19788
Employees	39578	10.9951	60.17641	1	5000
Employment Code	39578	1.886856	1.304252	1	10
Sales Code	36702	2.352597	1.528901	1	10
Trademark in 6mo	481809	0.0120546	0.1091297	0	1
Trademark in 6-12mo	481809	0.0016957	0.0411439	0	1
Patent in 6mo	481809	0.007393	0.0856641	0	1
Patent in 6-12mo	481809	0.00165	0.0405871	0	1
Innovativeness in Name	447471	0.1012467	0.205025	2.91E+15	1
Delaware Firm	481809	0.1152967	0.31938	0	1
Eponymous	481809	0.070117	0.2553444	0	1
Is Corporation	481809	0.5401041	0.4983896	0	1
Inc Date	481809	16540.69	1962.854	12784	19723
Inc Year	481809	2004.82	5.359871	1995	2013
log(Innovativeness in Name)	447471	13.27737	3.497498	15.05149	18.42068
Has IPO in 6 years	481809	0.0009195	0.0303085	0	1

Table 3

Dependent Variable: Dummy with 1 if employment after 6 years is over 100
 Sample: Massachusetts, years 1998 to 2005, only firms matched with Reference USA

	(1)		(2)		(3)	
	Logit Model	Marg. Effects	Logit Model	Marg. Effects	Logit Model	Marg. Effects
Delaware Jurisdiction	0.774***	0.0217***	0.765***	0.0213***	0.420***	0.00841**
	(0.0944)	(0.00380)	(0.0946)	(0.00377)	(0.107)	(0.00271)
Is Corporation	-0.466***	-0.0107***	-0.467***	-0.0106***	-0.503***	-0.00962***
	(0.0859)	(0.00222)	(0.0871)	(0.00224)	(0.0893)	(0.00199)
Name innovativeness (trained w RefUSA)	1.425***	0.0292***	1.411***	0.0288***	1.332***	0.0226***
	(0.103)	(0.00224)	(0.0974)	(0.00217)	(0.108)	(0.00214)
Eponymous	-0.936***	-0.0137***	-0.944***	-0.0137***	-0.761***	-0.00975***
	(0.135)	(0.00124)	(0.136)	(0.00121)	(0.121)	(0.00101)
Patent in 6mo					0.317	0.00624
					(0.184)	(0.00425)
Patent in 6-12mo					0.492**	0.0106*
					(0.164)	(0.00447)
Trademark in 6mo					2.043***	0.0926***
					(0.116)	(0.00877)
trademark in 6-12mo					1.452***	0.0522*
					(0.317)	(0.0204)
Industry Realtor			-0.132	-0.00255	-0.0406	-0.000676
			(0.191)	(0.00343)	(0.202)	(0.00329)
Industry Restaurant			-2.136*	-0.0188***	-2.000*	-0.0152***
			(1.031)	(0.00272)	(1.016)	(0.00254)
N	29304	29304	29106	29106	29106	29106
Base Probability		0.0210		0.0209		0.0172

Marginal effects; Standard errors in parentheses. Standard errors clustered at the year level. Sample: All firms registered between 1995 and 2005 which also match by name a firm in the Reference USA file 6 years after registration.
 * p<0.05 ** p<0.01 *** p<0.001

Table 4

Dependent Variable: Dummy with 1 if merger over \$10M of IPO after at most 6 years
 Sample: Massachusetts, years 1995 to 2005, all firms

	(1)		(2)		(3)	
	Logit Model	Marg. Effects	Logit Model	Marg. Effects	Logit Model	Marg. Effects
Delaware Jurisdiction	1.497***	0.0223***	1.484***	0.0217***	1.212***	0.0136***
	(0.0409)	(0.000939)	(0.0410)	(0.000923)	(0.0462)	(0.000782)
Is Corporation	0.752***	0.00540***	0.726***	0.00515***	0.599***	0.00375***
	(0.0489)	(0.000309)	(0.0491)	(0.000309)	(0.0505)	(0.000288)
Name innovativeness (trained w VC events)	0.196***	0.00155***	0.188***	0.00146***	0.154***	0.00104***
	(0.0169)	(0.000135)	(0.0171)	(0.000134)	(0.0195)	(0.000133)
Eponymous	-1.611***	-0.00731***	-1.606***	-0.00720***	-1.483***	-0.00599***
	(0.150)	(0.000345)	(0.150)	(0.000344)	(0.151)	(0.000322)
Patent in 6mo					0.810***	0.00834***
					(0.105)	(0.00157)
Patent in 6-12mo					0.552*	0.00497
					(0.220)	(0.00257)
Trademark in 6mo					2.820***	0.0936***
					(0.0718)	(0.00652)
trademark in 6-12mo					0.842**	0.00886
					(0.307)	(0.00474)
Industry: Realtor			-0.709***	-0.00411***	-0.594***	-0.00314***
			(0.146)	(0.000607)	(0.146)	(0.000583)
Industry: Restaurant			-0.892**	-0.00465***	-0.871**	-0.00399***
			(0.334)	(0.00110)	(0.331)	(0.000965)
Industry: Law			-0.552	-0.00332	-0.470	-0.00256
			(0.414)	(0.00188)	(0.408)	(0.00175)
Industry: Dental			-0.806	-0.00433	-0.724	-0.00351
			(0.706)	(0.00249)	(0.721)	(0.00240)
N	251726	251726	251726	251726	251726	251726
Base Probability		0.00796		0.00785		0.00683

Marginal effects; Standard errors in parentheses. Standard errors clustered at the year level. Sample: All firms registered between 1995 and 2005.

* p<0.05 ** p<0.01 *** p<0.001

Table 5**Below Cutoff (Not high-growth)**

Firm Name	Zip
1154-1166 MASS. AVE. LLC	02138
4 ALCOTT, LLC	02138
835 N. 6TH AVENUE LLC	02140
A2E PARTNERS LLC	02139
ADMINISTRATIVE ASSISTANCE, LLC	02138
ANTHA GROUP, INC.	02138
ANYBODY TECHNOLOGY INC.	02142
BIKE RIDES FOP LLC	02139
BYTEJUNGLE LLC	02139
CDM INTERNATIONAL ENTERPRISES LLC	02138
CM FARKAS, LLC	02138
EPSTEIN JOSLIN ARCHITECTS, INC.	02138
HARRY + LILI, LLC	02138
INSTANT LLC	02139
INTELLIGENT HUMAN VEHICLE INTERFACE, INC.	02142
LYNCON LLC	02141
MAPMYAPPS, INC.	02138
MY HOUSE WELLNESS AND REJUVENATION CENTER, INC.	02141
ONE PLEASANT STREET LLC	02139
WINDWARD ASSET MANAGEMENT LLC	02138

Above Cutoff (high-growth)

Firm Name	Zip
AERODESIGNS, INC.	02139
APPTOPIA, INC.	02142
BIOSTREAM THERAPEUTICS, INC.	02142
BLUEPRINT MEDICINES CORPORATION	02142
CLEARINSIGHT, INC.	02139
CO3 SYSTEMS, INC.	02140
COMPARZ, INC.	02139
COMPARZ, INC.	02139
GENEPEEKS, INC.	02138
INSIGHTSQUARED, INC.	02139
LEAF HOLDINGS, INC.	02142
NARA LOGICS, INC.	02142
NIMBUS DISCOVERY, INC.	02141
PAPERWORK.PRO, INC.	02141
PUNCHEY, INC.	02141
RANA THERAPEUTICS, INC.	02139
RECORDED FUTURE, INC.	02138
ROCKEFELLER CONSULTING TECHNOLOGY INTEGRATION, INC.	02141
THE FOUNDATION FOR RELIGIOUS LITERACY	02140
TRANSATOMIC POWER CORPORATION	02142

Figure 1



Figure 2

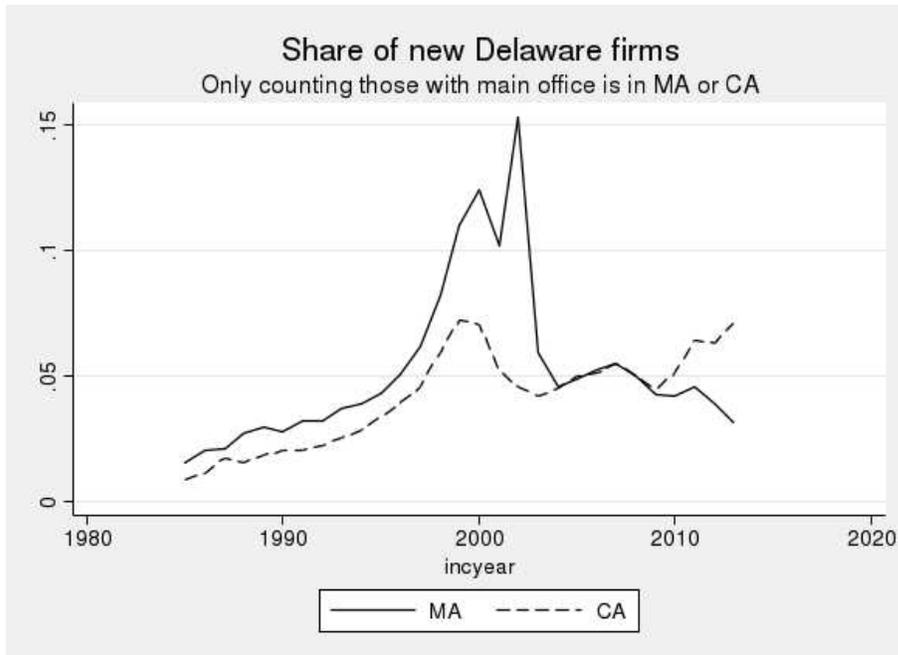
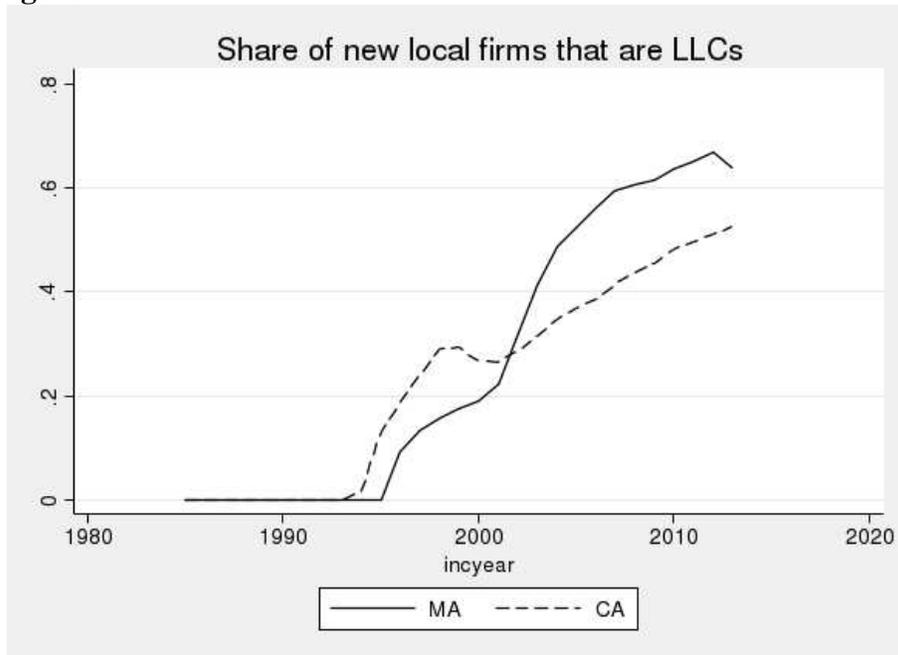


Figure 3

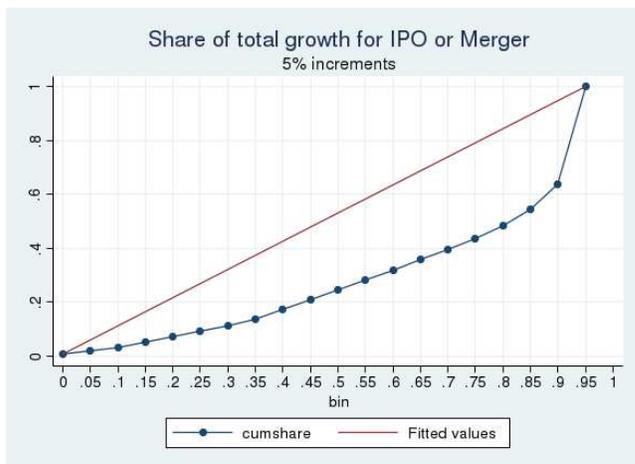
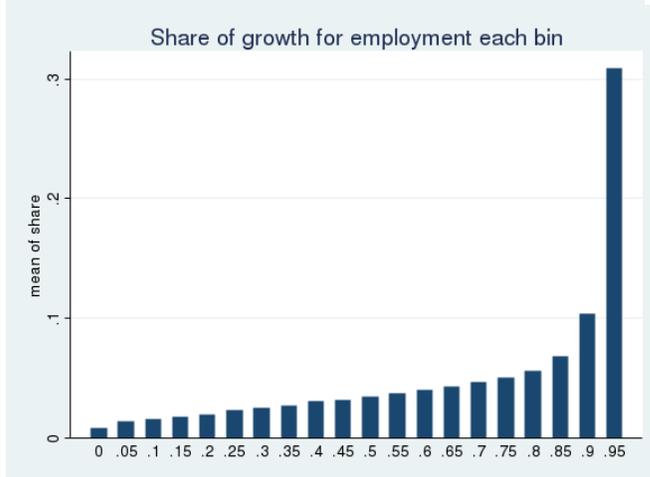
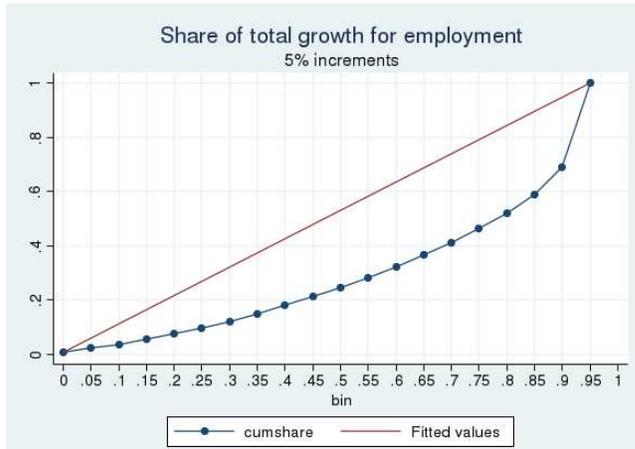


Figure 5

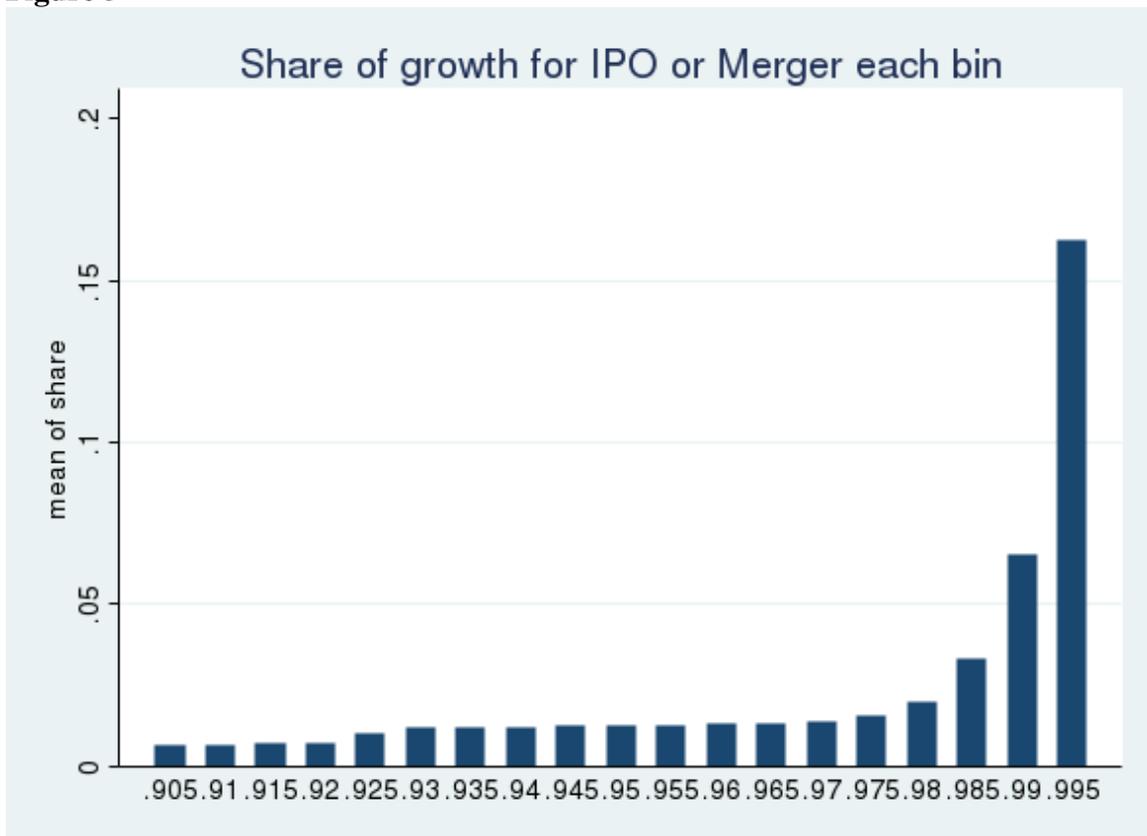


Figure 6

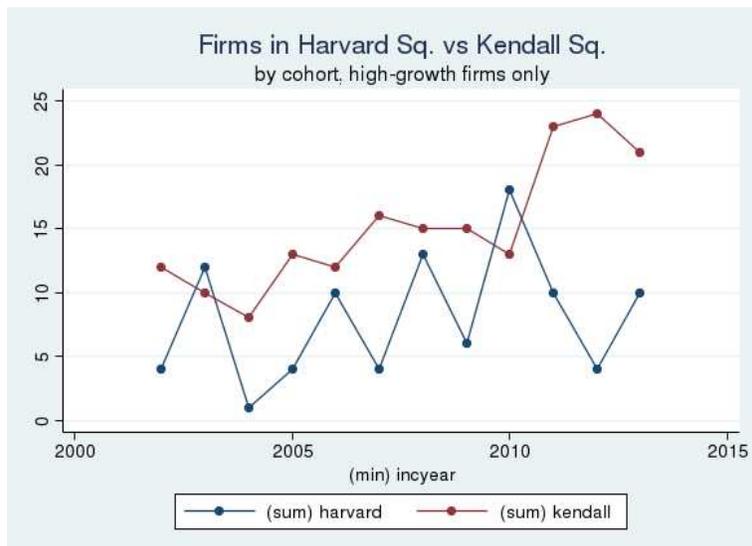
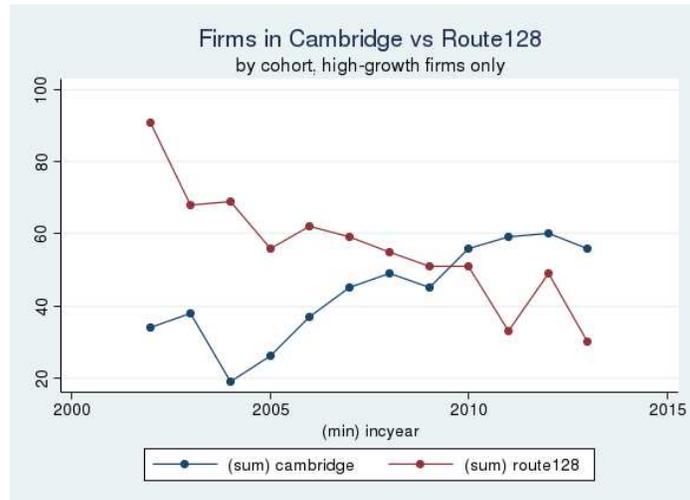
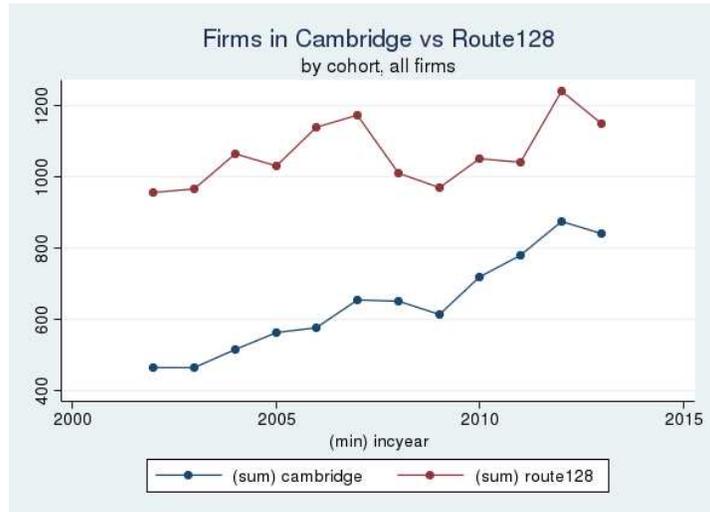


Figure 7

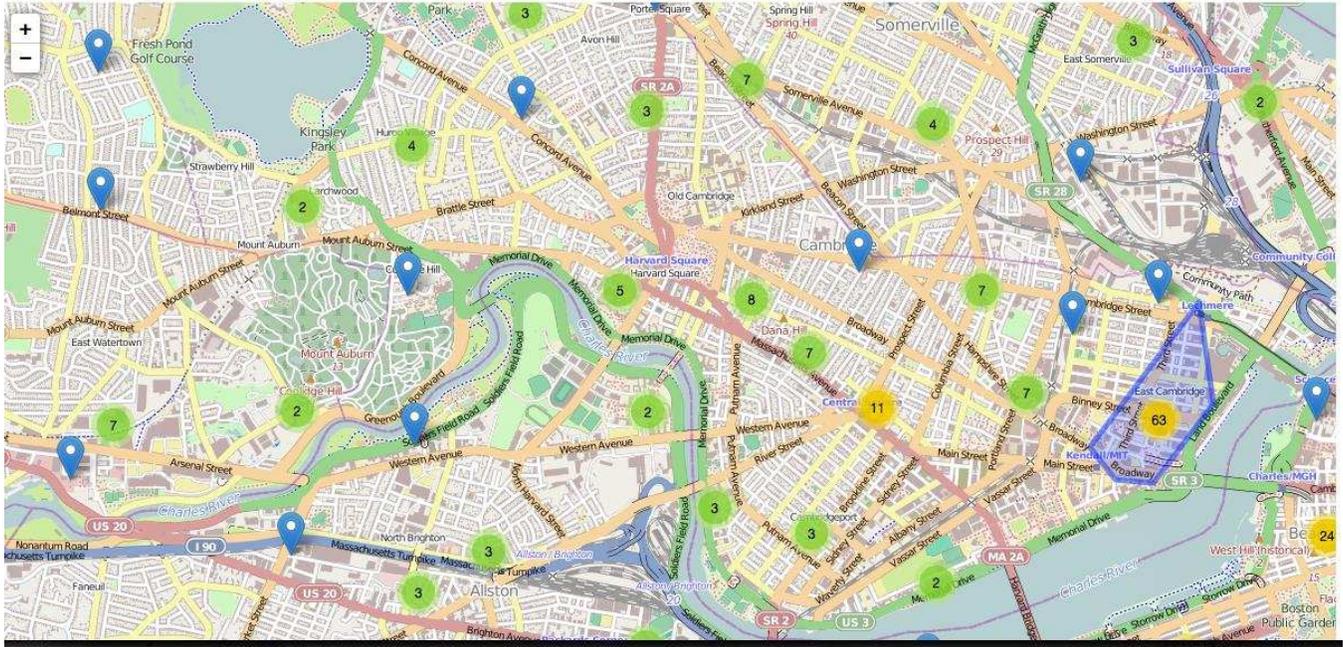
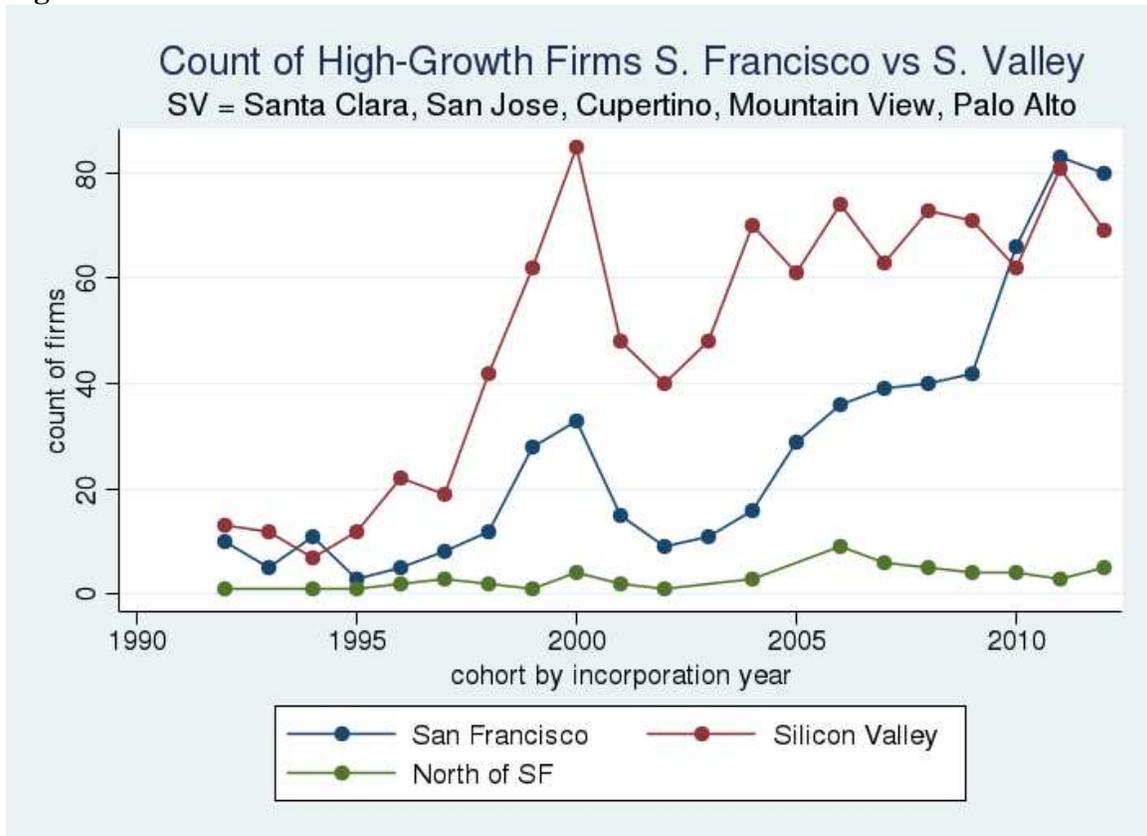


Table A1

Percentiles	IPO/Merger Innovative Name	Employment Innovative Name
1%	-0.4973786	0.0025452
5%	-0.4973309	0.0245783
10%	-0.4971061	0.0615577
25%	-0.4938647	0.1974802
50%	-0.4486014	0.4359153
75%	-0.0932632	0.6706466
90%	1.180011	0.8546112
95%	2.596365	0.9224964
99%	4.181873	0.9819102
Variance	1.016677	0.0812925
Skewness	2.717092	0.1195748
Kurtosis	9.910052	1.863652

Figure A1



Appendix 1

Snapshot of information of one firm (Digital Equipment Corporation) in the Massachusetts public record.

us/CorpWeb/CorpSearch/CorpSummary.aspx?FEIN=042226590&SEARCH_TYPE=1

Registers ▾ Classes ▾ NextBus Route S... Seminars ▾ NBER Groups ▾ blogs ▾ establishments ▾ SIIPP Data Dictio...

Corporations Division

Business Entity Summary

ID Number: 042226590 [Request certificate](#) [New search](#)

Summary for: **DIGITAL EQUIPMENT CORPORATION**

The exact name of the Domestic Profit Corporation: DIGITAL EQUIPMENT CORPORATION
Merged into COMPAQ COMPUTER CORPORATION on 12-31-1999
Merged with MAYNARD DEVELOPMENT CO., INC. on 06-27-1974
Merged with MAYNARD INDUSTRIES, INC. on 06-27-1974
Merged with APL SOFTWARE SYSTEMS, INC.(PA) on 06-27-1975
Merged with DEC REALTY TRUST(MA TR) on 08-13-1981
Merged with COMPAQ MERGER, INC. on 06-11-1998
Entity type: Domestic Profit Corporation
Identification Number: 042226590
Date of Organization in Massachusetts: 08-23-1957
Last date certain:
Current Fiscal Month/Day: 12/31 Previous Fiscal Month/Day: 06/30
The location of the Principal Office:
Address: 40 OLD BOLTON RD.
City or town, State, Zip code, Country: STOW, MA 01775 USA

Appendix 2

Note: Please email the authors if you wish to use our Python code or our innovative/non-innovative data sample.

We build our innovativeness in name metric by using the Natural Language Toolkit (NLTK) in Python. NLTK is a vast toolkit that allows to process normal language and classify, predict similarity, build grammars, get text corpora, and other uses. Other work using NLTK in social science includes Catalini et al (2014).

We use a specific part of it, the Naïve Bayes classifier. The Naïve Bayes classifier allows us to use text analysis to predict the probability that a firm is in different classes. While we could have many classes in principle, we only do two, innovative and non innovative. A classifier works by extracting features from text and storing the relative incidence of that feature f in each class. In a prediction, the classifier updates a base prediction continuously depending on each of the probabilities of new features.

This can be done by using Bayes rule: $P(innov|f_1 \dots f_n) = \frac{P(innov)P(f_1 \dots f_n|innov)}{P(f_1 \dots f_n)}$ but then we still have a problem because we don't know the joint probability $P(f_1 \dots f_n|innov)$. The Naïve Bayes algorithm solves this problem by assuming each feature is independent (hence the name). While apparently simplistic, this algorithm has been shown to be one of the most efficient and effective ones (see Zhang, 2004 for a technical discussion on the reasons). Our measure is the predicted probability of innovativeness.