

# **Online Appendix**

## **Age and High-Growth Entrepreneurship**

Pierre Azoulay, Benjamin F. Jones, J. Daniel Kim, Javier Miranda

### **Appendix I: Further Description of Data Sets and Methods**

This appendix provides additional details regarding the datasets used in this paper. Table A2 provides a summary of the datasets and their key variables. Many data sets are available to researchers through Census approved projects and accessible through Federal Statistical Research Data Centers (FSRDC), as further indicated in the table. The Schedule K-1 and Form W-2 datasets are currently accessible only by U.S. Census employees who have been granted access through approved internal projects.

#### **The Longitudinal Business Database (LBD)**

The LBD is an establishment-level longitudinal database tracking all establishments and firms in the US with at least one employee. Starting in 1976 and updated annually, the LBD currently covers years through 2015. The LBD is sourced from administrative income and payroll filings and enhanced with Census collections, including the Economic Census and the Company Organization Survey.

Key variables in the LBD include payroll, employment, industry, location (including state and county), and legal form of organization. Establishment and firm identifiers allow us to aggregate establishment-level information to the firm to identify firm-level employment and payroll. The ability to track establishments over time makes it possible to identify de novo firms (startups) distinctly from firms that emerge from corporate restructuring or M&A activity. Startups are defined as single-unit firms during the year in which the firm first appears in the LBD with at least 1 employee. In order to identify M&A activity using the LBD, we manually track changes in firm ownership. More specifically, we flag a firm ownership change when all of the existing establishments in a firm simultaneously receive a new firm identifier in the following year. In order to ascertain that the firm ownership changes are the result M&A rather than corporate expansions, we impose the following conditions: (1) the owning firm is an incumbent firm that exists in the LBD prior to the ownership change; (2) the original EIN and names of the

establishments prior to the acquisitions differ from those of the new owner's prior to the acquisition. For additional information regarding the LBD, see Jarmin and Miranda (2002).

### **Schedule K-1 (Form 1065/1120)**

Schedule K-1 is a tax form used to report business income or loss for owners of S-corporations and partnerships. Partnerships and S-corporations are pass-through entities, meaning that their profits are not taxed at the entity level but rather as they flow through to the owners. The Schedule K-1 reports the amount of income passed through to each party. Partnerships and S-corporations file a separate K-1 form for each of their owners and are required to account for 100% of profits. The availability of both employer identification numbers (EIN) and person identification numbers (SSN) allow us to identify all the owners of pass-through entities. The data start in 2007 and currently cover years through 2015. These data are confidential and currently can only be accessed through an internal U.S. Census project.

Key variables in Schedule K-1 include the income, deduction, and ownership share of partners and shareholders as well as the name, location, and employer identifier of the company. Unlike S-corporations, partnerships can be owned by other legal entities including partnerships and corporations. These tiered entities can make it hard to identify the ultimate owners of these enterprises when there are circular references. For more information see Goldschlag et al. (2017) and Cooper et al. (2015).

### **Form W-2**

Form W-2 is a tax form used to report the income paid to employees in remuneration for services rendered to an employer. Employers must file a W-2 for each of their employees for services performed during the year. The availability of both employer identification numbers (EIN) and person identification numbers (SSN) allow us to identify all the salaried workers associated with employer businesses in the US. The data start in 2005 and currently cover years through 2016. Key variables in Form W-2 include the income, social security taxes, and Medicare taxes as well as individual and employer identifiers. For more information about the W2 see (<https://www.irs.gov/forms-pubs/about-form-w2>).

### **Longitudinal Employer-Household Dynamics- Employment History File (LEHD-EHF)**

The LEHD-EHF is one of the core infrastructure files of the LEHD program. The EHF is sourced from quarterly unemployment insurance earnings records collected by labor market

information systems across the country for unemployment insurance purposes. The EHF provides a time series of all jobs held by individuals each quarter in each state. Key variables in the EHF include employer and individual identifiers, employment quarter and year, quarterly earnings, and industry of activity. The unit of analysis is the job or an employer-employee combination. A crosswalk between the state employer identifier (SEIN) and the federal employer identifier (EIN) is available. For additional information see the LEHD infrastructure file documentation ([https://lehd.ces.census.gov/doc/technical\\_paper/tp-2006-01.pdf](https://lehd.ces.census.gov/doc/technical_paper/tp-2006-01.pdf)).

### **The Annual Survey of Entrepreneurs (ASE)**

The ASE is a survey of approximately 290,000 firms in the non-farm private sector. The survey is a representative sample of firms in the US with employees. Starting in 2014, the ASE was conducted on an annual basis up to 2016. The ASE will be replaced with the Annual Business Survey starting in 2017. The ASE collects information on up to 4 owners of US businesses including age, gender, race, ethnicity and veteran status. Additional information includes the business owners' education, experience, and ownership role. The ASE is the source of core demographic statistics of US business owners and includes information such as number of firms, sales and receipts, annual payroll, and employment by gender, ethnicity, race, and veteran status. The survey includes modules to collect information on specific business activities. In 2014 the ASE collected additional information on R&D and innovation and the 2015 survey asked questions about management practices. For additional information see Foster and Norman (2017).

### **The Census Numerical Identification System File (Numident)**

The Census Numident is sourced from the Social Security Administration (SSA) applications for Social Security Numbers (Form SS-5). This is the SSA's master list of social security numbers (SSN) and includes all individuals in the US that have been issued a social security number. The Numident file is updated annually with years currently through 2016. Key variables include a protected identification key (PIK, which replaces the individual's SSN so as to protect their identity), date of birth, country of origin, gender, race and ethnicity. Starting in 1980 the SSA changed its collection of race and ethnicity so these data became non-mandatory items. The Census enhances these files with demographics data from its own data holdings including the decennial census and the American Community Survey to improve its quality.

### **The Patent Longitudinal Business Database Crosswalk (LPBD)**

The LPBD is a crosswalk file linking individual firms to specific patents in the US Patent and Trademark Office patent grants database. Starting in 2000 the LPBD links all inventors and firms identified in patent grant documents to firms in the LBD. The LPBD uses a triangulation strategy where the best possible match is identified by comparing matches to two alternative data sources: inventors are matched to the LEHD jobs file for workers, and patent assignees are matched to the Business Register file for firms. The file starts in 2000 and is updated annually. The LPBD currently cover years through 2015. Key variables in the LPBD include firm id, patent id, application year, assignee country, assignee state, and assignee type. The LPBD is able to match in excess of 75% of all patent-assignee combinations in the USPTO and 91% of patents with US firm assignees. For additional information see Graham et al. (forthcoming).

### **The Private Capital Research Institute-LBD Bridge (PCRI)**

The Private Capital Research Institute (PCRI) is a database of private capital data assembled by PCRI directly from several dozen private capital firms as well as from four major data vendors and private capital associations, including the Emerging Markets Private Equity Association (EMPEA), NYPPEX FUNDSIQ (“NYPPEX”) Thomson Reuters, and Unquote. PCRI were matched to the Business Register using name and address linking techniques. Key variables in the PCRI database include a company id, business name, street address, zip code, state, country, day of investment, and investment category. The PCRI bridge provides a link between the LBD and the PCRI database. Match rates of US headquartered firms to the LBD are in excess of 90%. For additional information about the matching methodology see Brown and Tello-Trillo (2017). External researchers wishing to use the linked PCRI and LBD data and both internal and external Census researchers wishing to use additional PCRI variables need to submit a proposal to Leslie Jeng, Director of Research at PCRI ([leslie.jeng@gmail.com](mailto:leslie.jeng@gmail.com)). The proposal guidelines can be found at [http://www.privatecapitalresearchinstitute.org/images/news/call\\_f\\_proposals.pdf](http://www.privatecapitalresearchinstitute.org/images/news/call_f_proposals.pdf).

### **VentureXpert**

VentureXpert is a commercial database for information covering venture capital and private equity investments. The data are linked to the LBD using name and address matching techniques. Key variables include firm name and address, funding type, funding round, amounts, date of funding, and names of the VC firms. Years covered include 1980-2005.

## **Compustat Bridge & Compustat**

The Compustat Bridge provides a link between the COMPUSTAT data and the LBD. Compustat provides financial, statistical and market information for publicly traded companies.

## **Prior Wage Analysis**

To examine the relationship between entrepreneurial entry, success, and wages we first constructed the prior wage history of each wage earner using each individual's W-2 records. For the analysis, we defined the "prior wage" as the maximum of the annual wage payments to that individual over the prior two years (a two-year window is used to help address the timing of entrepreneurial entry, which could come mid-year).

We ran regressions to capture the entry frequency with age, both with and without controls for prior wage. The regressions take the form

$$y_{it} = age_i + \mu_t + \mu_s + \theta logwage_{it-s} + \varepsilon_{it} \quad (\text{A1})$$

where  $y_{it}$  is equal to 1 if individual  $i$  founded a firm in year  $t$  and is 0 otherwise,  $age_i$  are age fixed effects from age 20 to 65,  $\mu_t$  are founding year fixed effects,  $\mu_s$  are the prior job's 4-digit industry fixed effects, and  $logwage_{it-s}$  is the individual's prior period log wage. The sample consists of a randomly selected 1% of the US population from each cohort between 2007 and 2014.

Figure A5 presents the age fixed effects, for both the regression above and for the same regression without the wage control. In explaining entrepreneurial entry, we see that the peak in middle age prevails regardless of whether we control for prior wages.

To further explore the relationship between wages and entry, and any differences for highly successful entrepreneurs, we then considered the distribution of prior wages, comparing founders with other workers. Specifically, we consider the percentile ranks of founders' wages (prior to starting their firm) in the wage distribution of the workforce.

Figure A6 shows these wage distributions. By construction, the percentile ranks for the broad workforce are uniformly distributed. Looking at all founders, we see a non-monotonicity. Founders appear disproportionately common among lower-wage workers and disproportionately

common among very high-wage workers. By contrast, founders of the highest-growth firms are far more likely to come from the upper end of the wage distribution.

While descriptive, these wage results can provide further facts to discipline conceptualizations of entrepreneurship. First, individuals with quite modest outside options start lots of ordinary firms, while those with unusually strong outside options tend to start growth-oriented firms. Second, the prior wages of high-growth founders suggests these individuals have outsize success both in the labor market and in founding firms. This finding is consistent with high-growth founders being skilled in multiple domains; it is also consistent with screening, where high-growth founders set a high bar for entry into entrepreneurship, given a high opportunity cost of leaving the ordinary labor market behind.

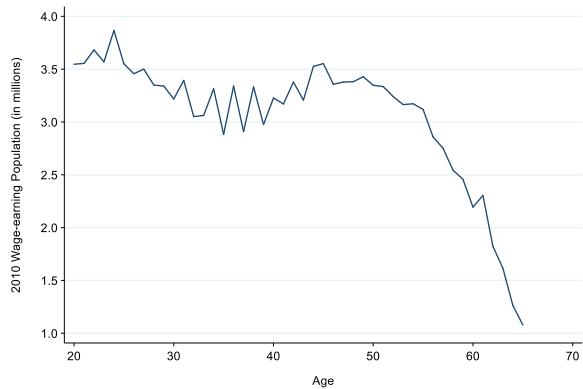
### **Analysis by Calendar Cohort**

While the main text pools the founding years, we can also provide additional analysis for each individual cohort year. This analysis provides a further way to generalize the findings while also demonstrating that the findings are robust outside the years 2007-2009, which overlap with the Great Recession. In particular, we provide a cohort-by-cohort analysis as far forward in time as our datasets allow. First, we extend analysis for each founding year through 2014 for our overall startup data. Second, we similarly extend the analysis for each founding year through 2014 using our “ex-ante” growth-orientation measure based on high-tech employment. Third, we extend the individual year analysis through 2011 for VC-backed startups and patenting startups, which is the limit these data allow. Finally, for “ex-post” growth outcome measures, and shortening the post-founding window to three years, we can look at individual cohorts through 2011. Table A4 presents the average founder ages for these separate cohorts. We see that the middle-age tendency is highly robust.

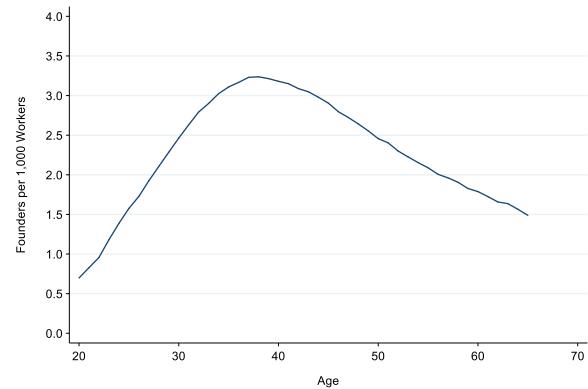
## Appendix II: Additional Figures and Tables

**Figure A1: Founder Rates by Age**

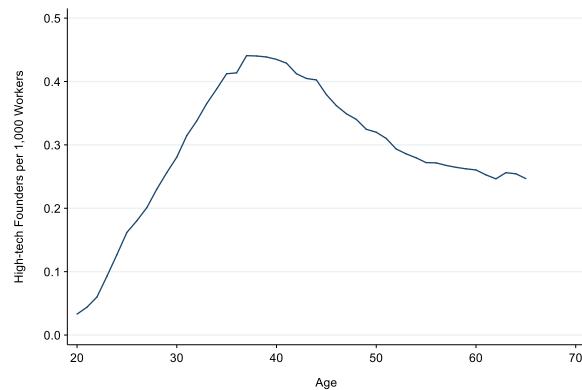
*Fig. A1-A: Size of Workforce by Age*



*Fig. A1-B: Founders per Worker, by Age*



*Fig. A1-C: Tech Founders per Worker, by Age*

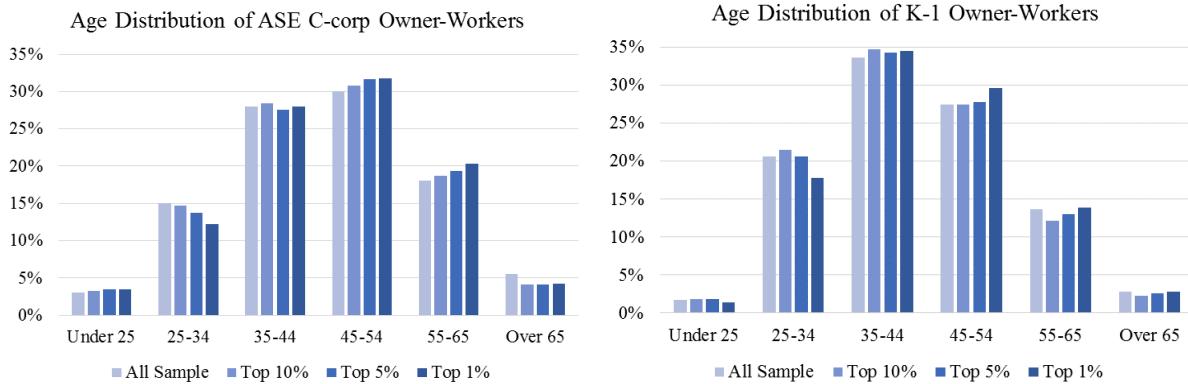


*Source:* Authors calculations based on W-2 earnings records, form K-1, and Longitudinal Business Database between 2007 and 2014.

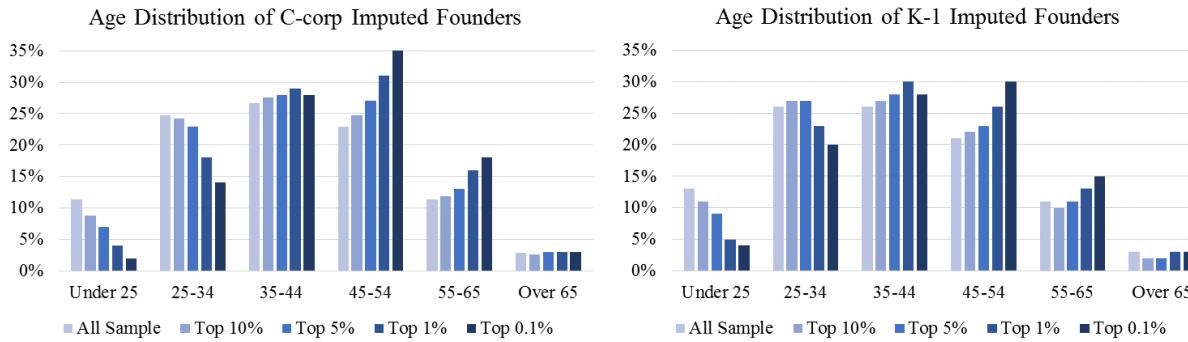
*Notes:* These figures show the number of wage earners, founders, and high-tech founders in the US. Each bin represents an age cohort. Ages between 20 and 65 are incorporated in the plots. Figure A1-A uses the 2010 W-2 file. Figures A1-B and A1-C use data over 2007-2014.

**Figure A2: Results by Founder Definition and Legal Form**

*Panel A: Owner-Worker, C-Corporation and K-1 Firms*



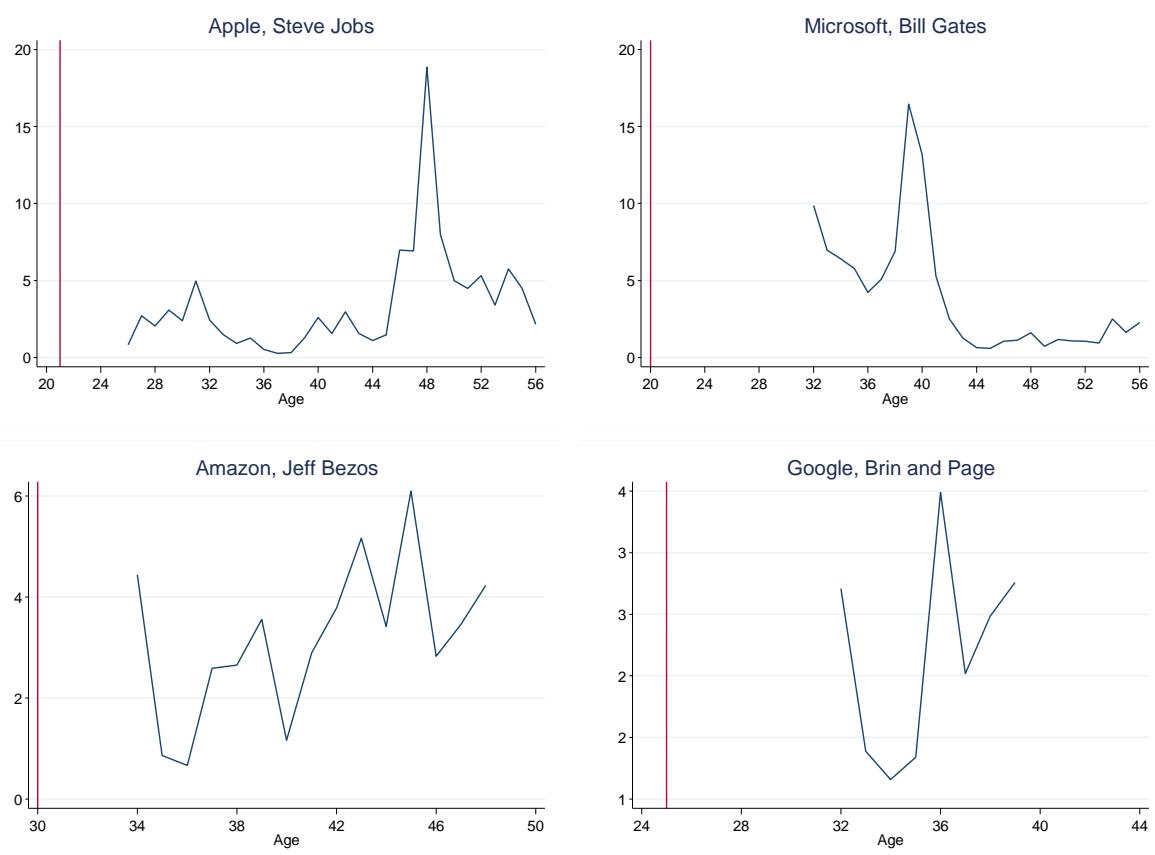
*Panel B: Initial Top 3 Earners, C-Corporation and K-1 Firms*



*Source:* Authors calculations based on Longitudinal Business Database, W-2 earnings records, form K-1, and Annual Survey of Entrepreneurs.

*Notes:* Startup firms born between 2007 and 2012 in the Annual Survey of Entrepreneurs (ASE) are included for the left side of Panel A. Growth outcomes are calculated over a three-year window for each cohort and the top 1%, 5% and 10% is identified from the distribution. The rest of the figures include all new C-corporations, S-corporations, and Partnerships in the Longitudinal Business Database (LBD) born between 2007 and 2011. Growth outcomes are calculated over a three-year window for each cohort and the top 0.1%, 1%, 5% and 10% is identified from the distribution. The left side of Panel A is based on founders of C-corporation firms in the Annual Survey of Entrepreneurs. The right side of Panel A is based on founders of S-corporations and partnerships in the K-1 database. Panel B is based on imputed founders (first-year joiners who are among the top three earners) using W-2 wage-records.

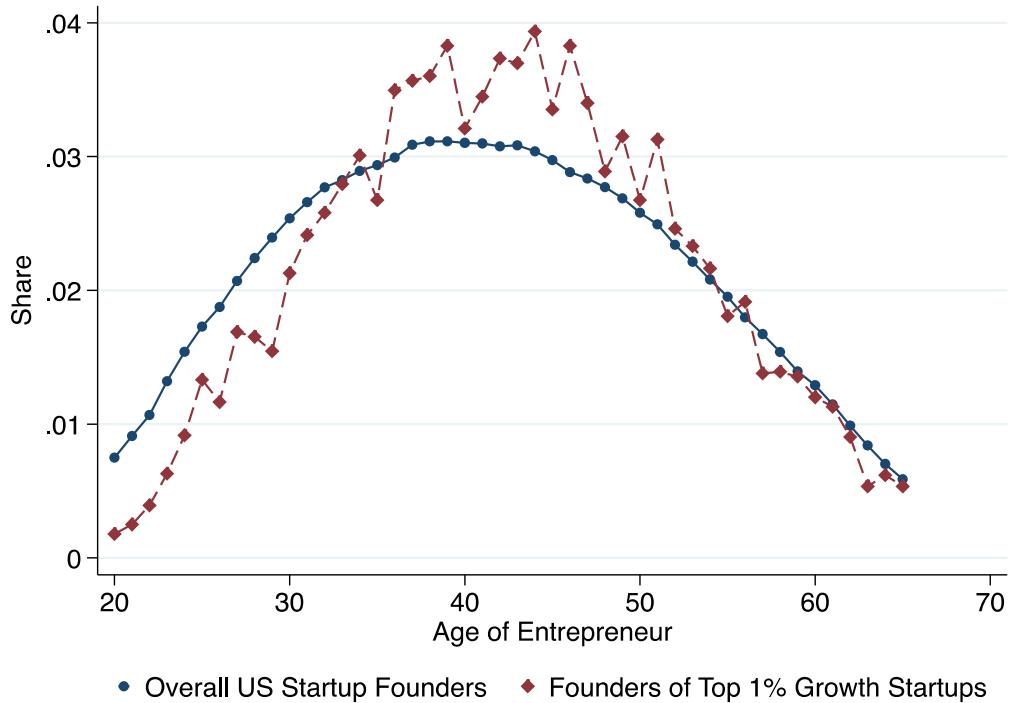
**Figure A3: Forward Stock Multiples as the Founder Ages:  
Apple, Microsoft, Amazon, and Google**



*Source:* Authors calculations based on public data.

*Notes:* The vertical red line indicates the founders' age in the year of the firm's founding, and the x-axis presents the age of the indicated founder as time passes. The forward stock-price series begins in the year of the initial public offering for each firm. For Google, Brin and Page were born in the same year (1973). Historical share prices are sourced from Bloomberg.

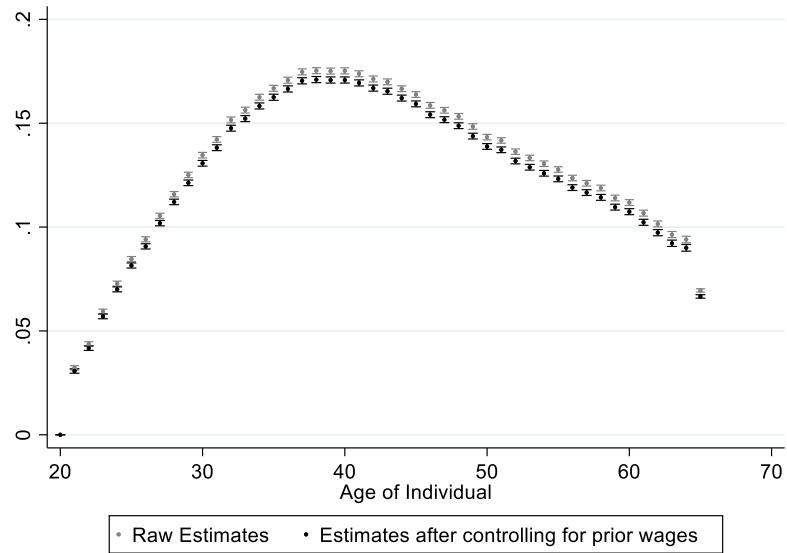
**Figure A4: Founder Age Distribution:  
All Startups and High Performance Startups by Sales (5 Years after Founding)**



*Source:* Authors calculations based on W-2 earnings records, form K-1 and Revenue-Enhanced Longitudinal Business Database.

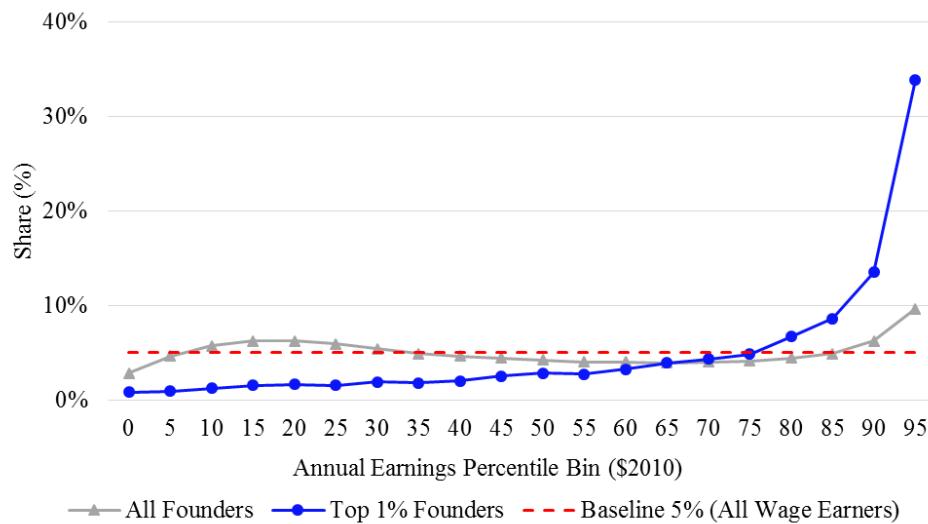
*Notes:* This set of kernel density plots shows the age distribution of startup founders (at year of founding) in the US. Each bin represents an age cohort. Ages between 20 and 65 are incorporated in the plots. The blue (left) plot incorporates all founders of new C-corporations, S-corporations, and Partnerships with employees founded between 2007 and 2014 as identified in the Longitudinal Business Database (LBD). The red (right) plot represents founders of the top 1% growth firms founded over the 2007-2008 period, given that revenues data are available up to 2013. Top 1% revenue growth threshold value is calculated for each yearly cohort based on the real revenue figures from the LBD in the five years after the birth of the firm.

**Figure A5: Founder Age Distribution,  
With and Without Controls for the Founders' Prior Wages**



*Notes:* This figure presents estimates of the age indicator variables in the regression equation (A1), together with their associated 95% confidence intervals, with and without prior wage controls for the individuals in the sample, which consists of a randomly selected 1% of the US population of wage earners in the W-2 from each cohort between 2007 and 2014.. See online appendix text for details of the data construction and regression specification.

**Figure A6: Wage Distributions of Non-Founders, Founders, and Successful Founders**



*Notes:* This figure examines the wage distribution of founders and highly successful founders compared to the background wage distribution of the workforce. Prior wages are calculated from W-2 records and translated into 2010 U.S. dollars. The x-axis represents percentile bins of annual earnings. By construction, the percentile rank for the workforce as a whole is uniformly distributed. Top 1% founders are those whose firms achieve top 1% employment growth within 3 years.

**Table A1: Founder Age – Perceptions from Media & Two Prominent VCs**

	TechCrunch Awards	Inc. and Entrepreneur Magazines	Sequoia	Matrix Partners
Mean	31.0	29.1	33.9	36.5
Median	30	27	33	36
(St. Dev.)	(7.1)	(7.0)	(8.7)	(8.6)
Observations	232	51	415	246
Period	2008-2016	2015	1969-2014	1948-2014
Sectoral Focus (top 5)	Education, Software, Social Media, Consumer Electronics, e-Commerce	Technology, Retail, Media, Consumer Goods, Food Delivery	Semiconductors, Networks, Task Mgmt. Apps, Website Compilers, Cloud	Networks, Applications, Commerce, Platform/ Infrastructure, Semiconductors/ Materials

*Notes:* TechCrunch gives annual awards to the “most compelling startups, internet and technology innovations of the year”. Inc. magazine and Entrepreneur magazine provided “Entrepreneurs to Watch” lists in 2015. The founder ages for new ventures backs by the two venture capital firms (Sequoia Capital and Matrix Partners) were obtained by the authors through researching all the companies listed on their respective websites.

**Table A2: Summary of Data Sets**

Dataset	Units and coverage	Relevant Variables	Period and Frequency	Access
Longitudinal Business Database (LBD)	<ul style="list-style-type: none"> <li>• Establishments and firms</li> <li>• All private non-farm employers in the US and outlying territories</li> </ul>	Firm identifier, establishment identifier, payroll, employment, industry, location, legal form of organization	Annual, 1976-2015	FSRDC/Census approved projects
Schedule K-1 (Form 1065/1120)	<ul style="list-style-type: none"> <li>• Owners</li> <li>• All pass through entities (partnerships and S-corporations)</li> </ul>	Individual identifier, firm identifier, business income, deductions, share of profit/loss	Annual, 2007-2016	Census Bureau employees on approved projects and a need to know
Form W-2	<ul style="list-style-type: none"> <li>• Employees</li> <li>• All workers in the US for whom employers made payments</li> </ul>	Individual identifier, employer identifier, wage income, social security, or Medicare wages.	Annual, 2005-2016	Census Bureau employees on approved projects and a need to know
Longitudinal Employer-Household Dynamics-Employment History File (LEHD-EHF)	<ul style="list-style-type: none"> <li>• Salaried workers by employer</li> <li>• All salaried workers subject to unemployment insurance</li> </ul>	Individual identifier, employer identifier, earnings (quarterly and annualized), industry	Quarterly, 20XX-2015 (Initial year varies by state)	FSRDC/Census approved projects
Annual Survey of Entrepreneurs (ASE)	<ul style="list-style-type: none"> <li>• Businesses</li> <li>• Sample of 290,000 non-farm businesses with paid employees</li> </ul>	Firm identifier, information for up to 4 owners including age, gender, race,	Annual, starting in 2014-2016 to be replaced by	FSRDC/Census approved projects

---

	and receipts of \$1,000 or more	ethnicity, education, experience and type of ownership	the Annual Business Survey in 2017	
Census Numident	<ul style="list-style-type: none"> <li>• Individuals</li> <li>• All individuals with a US social security number</li> </ul>	Individual identifier, date of birth, gender, race, ethnicity, country of origin	Updated annually	FSRDC/Census approved projects
Longitudinal Patent Business Database (LPDB)	<ul style="list-style-type: none"> <li>• Patent-firm links</li> <li>• All patents in the USPTO grants database matched to the LBD</li> </ul>	Firm identifier, Patent identifier, year	Annual, 2000-2014	FSRDC/Census approved projects
Private Capital Research Institute-LBD Bridge (PCRI)	<ul style="list-style-type: none"> <li>• Firms</li> <li>• Private capital deals including buy outs, VC, growth equity, secondary purchases.</li> </ul>	Firm identifier, Category of private capital	1990-2015	FSRDC/Census approved projects prior approval of PCRI
VentureXpert	<ul style="list-style-type: none"> <li>• Firms</li> <li>• VC deals</li> </ul>	Firm identifier, Venture capital funding	1980-2005	Data provided by researcher through a license agreement
Compustat-Bridge	<ul style="list-style-type: none"> <li>• Publicly traded firms</li> </ul>	Firm identifier, financial and market data	1976-2013	FSRDC/Census approved projects prior approval of PCRI

---

**Table A3: Founder Age — Oldest and Youngest Technology Sectors***Panel A: Technology Sectors, Youngest 5*

NAICS Code	Sector	N	Mean
5172	Wireless Telecommunications Carriers (except Satellite)	1,500	38.5
5182	Data Processing, Hosting, and Related Services	6,100	39.7
5112	Software Publishers	3,600	39.8
5415	Computer Systems Design and Related Services	100,000	40.1
8112	Electronic and Precision Equipment Repair and Maintenance	4,900	40.8

*Panel B: Technology Sectors, Oldest 5*

NAICS Code	Sector	N	Mean
4862	Pipeline Transportation of Natural Gas	50	51.4
3251	Basic Chemical Manufacturing	700	47.9
3255	Paint, Coating, and Adhesive Manufacturing	400	47.5
2111	Oil and Gas Extraction	3,100	47.5
3336	Engine, Turbine, and Power Transmission Equipment Manufacturing	400	47.3

*Notes:* Sector is shown in the first column, observation counts of founders in the second column, and mean founder age in the third column. Sectors are defined at the 4-digit NAICS level. Only new firms are included. Counts are rounded to comply with disclosure rules. Sample is all new businesses in the U.S. from 2007-2014 based in the Longitudinal Business Database (LBD).

**Table A4: Mean Founder Age by Calendar Year of Firm's Founding**

Founding Year	All Startups	Ex-Ante Startup Type			Ex-Post Startup Success	
		High-Tech Sectors	VC-backed Firms	Patenting Firms	Top 1% (3-yr)	Successful Exits
2007	41.8	43.2	42.4	44.0	43.8	46.3
2008	41.8	43.2	42.2	44.2	44.2	46.2
2009	41.8	43.3	42.7	45.2	44.6	46.1
2010	41.8	43.4	41.6	45.0	44.1	46.9
2011	41.8	43.4	41.5	45.3	44.9	47.5
2012	41.8	43.1	-	-	-	-
2013	42.0	43.0	-	-	-	-
2014	42.5	43.3	-	-	-	-

*Notes:* This table presents the mean age of founders by year of founding (rows). Mean age is presented subject to data availability of the growth-orientation measure (columns). Data for all new ventures and for new ventures in high-tech sectors are available through 2014. VC-backing and patenting firms are known for firms in the LBD through 2011. For ex-post growth performance, employment growth uses a 3-year window to determine upper tail firms. This growth measure and the successful exit measure are known for new ventures starting through 2011.

**Table A5: Minimum and Maximum Ages within Founder Teams**

*Panel A: Owner-Worker Definition of Founders (K-1)*

	All Startups	High-Tech Startups	VC-backed Startups	Patenting Startups	Top 1%	Top 0.1%	Successful Exit
<b>Within Startup</b>							
Min Founder Age	42.7	44.0	39.8	43.6	40.9	42.3	43.3
Max Founder Age	44.6	45.5	47.8	46.9	45.6	47.8	47.1

*Panel B: Initial Team Definition (K-1 and C-Corporations)*

	All Startups	High-Tech Startups	VC-backed Startups	Patenting Startups	Top 1%	Top 0.1%	Successful Exit
<b>Within Startup</b>							
Min Founder Age	35.1	39.1	36.5	37.8	35.0	37.4	38.5
Max Founder Age	46.0	45.7	47.3	48.4	50.1	51.4	51.4

*Notes:* Panel A incorporates all S-corporations and Partnerships founded over the 2007-2014 period in the Longitudinal Business Database (LBD), except for the Top 1% and Top 0.1% columns, which include those firms founded over the 2007-2009 period for which we can observe 5 years of employment data after founding. Panel B incorporates all S-corporations, Partnerships, and C-corporations founded over the 2007-2014 period, except for the Top 1% and Top 0.1% columns, which include those firms founded over the 2007-2009 period for which we can observe 5 years of performance data after founding.

## **References**

- Brown, David and Cristina Tello-Trillo. "The PCRI/LBD Bridge File." Mimeo, 2017.
- Cooper, Michael, John McClelland, James Pearce, Richard Prisinzano, Joseph Sullivan, Danny Yagan, Owen Zidar, Eric Zwick (2015), "Business In The United States: Who Owns It And How Much Tax Do They Pay?" *Tax Policy and the Economy* 2016 30:1, 91-128
- Goldschlag, Nathan, J. Daniel Kim, and Kristin McCue (2017), "Just Passing Through: Characterizing U.S. Pass-Through Business Owners." *Working Papers 17-69*, U.S. Census Bureau Center for Economic Studies Working Paper No. 17-69.
- Graham, Stuart J. H., Cheryl Grim, Tariqul Islam, Alan C. Marco and Javier Miranda (forthcoming), "Business Dynamics of Innovating Firms: Linking U.S. Patents with Administrative Data on Workers and Firms." *Journal of Economics and Management Strategy*.
- Jarmin, Ron and Javier Miranda (2012), "The Longitudinal Business Database". Center for Economic Studies, CES-WP-02-17.