Title

Of Experience and Enterprise: Careers, Organizations and Entrepreneurship

Author

Ng, Weiyi

Of Experience and Enterprise:
Careers, Organizations and Entrepreneurship

By

Weiyi Ng

A dissertation submitted in partial satisfaction

of the requirements for the degree of

Doctor of Philosophy

in

Business Administration

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:
Professor Toby E. Stuart, Chair
Professor Ming D. Leung
Professor Noam Yuchtman
Professor Morten T. Hansen

Spring 2017

Abstract

Of Experience and Enterprise:
Careers, Organizations and Entrepreneurship

By

Weiyi Ng

Doctor of Philosophy in Business Administration
University of California, Berkeley
Professor Toby E. Stuart, Chair

This dissertation examines the antecedents of entrepreneurship through the empirical analysis of over 2 million resumes that constitutes a sample of the high technology start-up ecology in the United States. The first chapter characterizes the latent issues surrounding the study of entrepreneurial entry (Chapter 1). I then resolve these issues through the development of a sociological career framework of entrepreneurship in two parts. The first establishes the framework and distinguishes two types of entrepreneurial activity: high potential ventures and common self-employment (Chapter 2). I show that machine learning models applied to the identity claims of hundreds of thousands of entrepreneurs can successfully classify, characterize and distinguish these types in the tech sector. The two entrepreneur types exhibit diametrically opposing human capital and career based antecedents. In doing so, I demonstrate a necessary de-conflation of entrepreneurial events; the career framework provides a crucial precision in the definition, observation and measurement of the entrepreneurial outcome variable. The second part exemplifies an application of the framework to demonstrate an efficacy in the identification and study of specific sociological mechanisms. Through the introduced apparatus and a prospective sample of the data that represents the graduates of the top 23 science, technology, engineering and mathematics (STEM) colleges in the United States, I study the effect of status gain on entrepreneurial entry and success by examining different forms of entrepreneurial activity of the alumni of companies that experience liquidity events: initial public offerings (IPOs) and large scale acquisitions (Chapter 3). I find that upon vicariously experiencing these liquidity events, the alumni are on average 23% more likely to enter into high potential entrepreneurship and 17% less likely to enter into contract self-employment. However, such forms of status gain confer no significant funding advantages to the nascent venture. I conclude by discussing future directions: this dissertation serves as but an introduction to and an advocate for a larger program of research that seeks to clarify and advance the study of entrepreneurship through sociological career theory.

ACKNOWLEDGEMENTS

It has come to this: the amount of help, support and attention that I have received in the pursuit of this dissertation will be merely given a brief nod by these few words, stylized in a ubiquitous typeface and prefaced to a document which will probably receive scant attention. Regardless, I do hope that this attempt at eloquence will convey the immense gratitude and sense of debt I owe to the people mentioned here; a dedication to the social structure, if you will, that supported and facilitated not only my own intellectual growth and career, but also gave meaning to these years of my life.

I will start with the chair of my dissertation committee and my advisor, Toby Stuart. The effect of Toby's influence and instruction cannot be understated. To be absolutely sure, this is an obvious (and rather unnecessary) statement. But beyond the duties of overseeing my education and development, I have grown to really appreciate Toby, not just as a teacher and advisor, but also and more significantly as a mentor and friend. He genuinely cares and for some reason, believed in my work from the very beginning with a sort of saintly patience. And perhaps more than anything, it is not just the directions and discourse, but also the pep-talks and pick-me-ups that have seen me through some rather dark times, sustained the whole effort and made all this possible. It is quite amazing, and slightly unnerving, to find an advisor who sees more in you than you see yourself. What I am definitely going to miss are the conversations over drinks, cheese and nuts. These interlocutions over a strangely specific diet have repeatedly reinforced a personal mantra: a person whom you can have honest, memorable and non-judgmental discussions over a beer is someone you can call a lifelong friend.

If meetings with Toby are marked by a lack of food, with Ming Leung it is by an abundance of. With open minds and filled stomachs, our work and mentoring relationships often feature frank, productive and stimulating discussions over caloric breakfasts. This abundance mirrored his mentorship: Ming has always been generous with his time, teachings and advice. This generosity extended beyond the supervision of my own research: he went out of his way to include me in conferences, thus navigating and demystifying the ins and outs of the field. All this is done with an enviable sunny disposition and optimism.

I first met Noam Yuchtman as a student in his research methods class. There, in awe, I watched him dart around the classroom and whiteboard, energetically discussing the papers he loved and the ideas that sprung forth. This energy has translated to his advising. His insight into the research process and empirics was uncanny, his enthusiasm and inquisitiveness was inspiring. Conversations with Noam were characterized by an effervescent font of ideas, his mentorship by a focus on and concern for my own curiosity and fascinations. From Noam, I have come to appreciate the authenticity inherent in our profession, and I will be happy to inherit even a fraction of his passion for his work.

I like to thank Morten Hansen for being supportive towards my dissertation, and for supervising the process since my qualification exams. I will also like to express my gratitude for the following professors for their moral support and advice at various times throughout this process: Jennifer Chatman, Gina Dokko, Andreea Gorbatai, Heather Haveman, Don Moore and Sameer Srivastava.

This effort, and make no mistake: it was quite an effort, was significantly lightened and made bearable by the student culture at Haas. Beyond the powerhouse of intellect and ideas, the personal and professional support provided by this family of students made the entire journey that much more enjoyable. The laughter that was shared and tears that were shed… I missed and will continue to miss this group and those times deeply.

Brian Reschke and Eliot Sherman were the pillars of this community. Brian brought about a calm that was an Apollinian veil around what would otherwise be a Dionysian chaos of graduate school. Beyond the care and meticulousness that he brought to his work, Brian was selfless in his support and generous in his contributions to our student culture. How he achieves this on top of his dedication and devotion to his family is beyond me. With Eliot Sherman, there was a vibe, a feeling that everything will be alright. Not only was he a trove of ideas and a beautiful writer, he dispensed opinions and advice freely with a certain wit and whimsy that made coming to work fun. The group was never the same again with their absence.

My cohort mate: Jennifer Logg. How far we have come! I still remember the first meeting at Jupiter where we discussed, probably naively, our interests and wants and hopes and dreams. Jenn was all that I wished for in a cohort-mate, office neighbor and a friend. Beyond the largesse of her snack drawer, I will cherish the mutual amateur counseling sessions, the sushi dinners, her attempts at making dinner and generally being there for all the joys and sorrows of the past 6 years.

I will like to thank Sanaz Mobasseri, whose empathic attunement to everyone’s emotions and troubles provided an emotional safety net that was always there. I will also like to thank Nicole Zeng for her positive outlook and confidence. Nicole had an enthusiasm for my work and my career that I do not deserve. And of course, Laura Howland, for her frequent attempts at feeding everyone around her simple carbohydrates and for those days of late evening office companionship as we prepared for the competency and qualifying exams.

And the junior students. Katherine Ullman has been a wonderful friend and collaborator; I greatly enjoyed our conversations on research and life. Daron Sharps was always there when I needed help in any way. And last but not least, Richard Lu, whose quiet company in the office was always appreciated.

Special mentions go to my friends Minah Jung, Sheng Li and Donatella Taurusi. In Minah I have made a lifelong best friend; not only is her insight into the dynamics of academic life uncannily accurate, she is someone that has my back and was always there for whatever my troubles or celebrations. Sheng is one of the kindest people I know, and his frequent dinner parties, events characterized by amazing food and rich conversations amongst friends punctuated an otherwise Sisyphean enterprise. Donatella had a love for all things life which she freely shared and I greatly admired. Her sunshine disposition, openness and generosity is something that I treasured and will continue to treasure in our friendship.

I will also like to thank those that happened to be in my vicinity during this journey. They include but are not exclusive to the following: Moshe Barach, Jon Cowan, Angus Hildreth, Yoonha Kim, Jeffrey Kuhn, Sibo Lu, Mike o’Donnell, Jessica Pan, Elizabeth Tenney, Santiago Truffa and Alex Van Zant. In all my friends are the shared characteristics of generosity and again, a belief in my person and a confidence in my abilities that is greatly humbling. So, to all my friends, thank you.

It must be mentioned that this work is a result of the work of my wonderful undergraduate research assistants: Kangting Yu, Cindy Mo, Phoebe Wong, Jong Ha Lee, Gao Xian Peh and Vladyslav Rudoy. I consider myself very lucky to have been able to work with and know better this group of brilliant students. The talent and enterprise of Berkeley’s undergraduate students constantly surpasses all my expectations; their energy and enthusiasm remain deeply inspiring. I hope that they have learned as much from me as I have from them and I want to say a big “thank you” for all their work and dedication.

I like to thank Kimberly Guilfoyle and Melissa Hacker for always being there for me and for helping to sort out otherwise Kafkaesque experiences of Berkeley bureaucracy.

Sean Goh has been my key partner in crime. How far we have come! We have ventured and questioned and survived and celebrated. Sean has been both an unmovable emotional pillar and a source of reflection and perspective. With Sean, I have an escape; with Sean, things seem so much simpler. I am immensely grateful to have found such a patient partner and a best friend who believes in everything I do without any bounds and conditions. For waiting across the Pacific for all these years… regardless, our laughter and tears have imbued my life with meaning and made me a better person. And for that, a very big thank you.

My parents. Throughout this process they have invested a trust in my decisions and abilities and have been very supportive of all my ambitions and needs. I know they have missed me very much, and I cherish their pure, unconditional love. Our relationship has definitely developed through this experience, and I look forward to spending more quality time with them.

Finally, I want to thank the Kauffman Foundation for their generous financial and academic support towards the completion of this dissertation. I also thank Bloomberg Beta (a shout out to Roy Bahat, Dan Strickland and Morgan Polotan) and People.co (Rob Dennis and Greg Hudson) for assisting in the acquisition of various datasets critical to these studies.

The contents of this publication are solely the responsibility of my own.

TABLE OF CONTENTS

CHAPTER 1
ON ENTREPRENEURIAL ENTRY

Who becomes an entrepreneur? A systematic attempt at an inquiry into the antecedents of entrepreneurial entry immediately raises several issues that demand attention.

There is the definitionally elusive nature of the phenomenon of study: this simply is the question of "what *exactly* is entrepreneurship?" (e.g. Sorenson and Stuart, 2008). Indeed, in both scholarly literature and popular conception, the term is invoked to encompass not any particular and singular phenomenon but rather a loosely bounded collection of myriad phenomena. For instance, classic work in sociology conceptualizes entrepreneurship as a response to blocked economic mobility and restricted access to the primary sector of the labor market (e.g. Portes et al., 1989). Conversely, others have instead attended to the creation and evolution of venture financed high potential, high-growth science and technology based companies (e.g. Stuart and Ding, 2006). In the practicing industry, the term "entrepreneur" has developed a buzz status, adorning the covers of airport book stores and paraded on resumes and *curriculum vitae*. On the one hand, lay folk often think of entrepreneurs as movers of the contemporary world, a term reserved for the capitalist legends and icons of our day. On the other, they are eager to be associated with all the supposed status and positive connotations of the label, freely associating any instance of their own enterprise, however large or minute, as entrepreneurial. In both realms, the use and abuse of the "entrepreneurship" term threatens to dilute its semantic value to non-existence.

Then is the issue of "when does entrepreneurship *start?*" Largely, the academic field agrees that entrepreneurial entry broadly involves a "founding event": a primary process initiated at a point in time by individuals in an attempt to start new organizations. Concordantly, the success or failure of the founding attempt is aside to the founding event; they constitute a secondary process. Yet most archival documentations of founding events are "culmination of the operation of both processes." (Carroll and Khessina, 2005). Be it incorporation databases or venture funding data, records of organizational founding do not capture individual and organizational attrition at the early stages: a failure to account for failure. These archival datasets implicitly impose an observation criterion that generates observations that are susceptible to survivorship bias at best, and samples on the dependent variable at worst. While an issue that is often whispered at the sidelines of footnotes or relegated to discussion sections, it remains non-trivial and controversial. Affixing a $t=0$ to the instantiation of entrepreneurial activity forms an observational puzzle that is dealt with usually via academic hand waving of various subtleties. Consider that the beginnings of entrepreneurship simply involve an idea and its corresponding claims: pinpointing the exact beginnings and time of entry into entrepreneurial activity proves to be challenging and has thus far eluded archival research.

On top of these issues lay the overall rarity of the phenomenon and complexity of its causes. While reportedly, the (success agnostic) incidence of entrepreneurship amongst the population is rare and unusual (although in its pluralistic entirety, increasing), estimating a base rate for the various forms of entrepreneurial activity forms yet another problematic. The aforementioned issues suggest a non-trivial task of numerating these rates; in addition, the difficulty of observing and defining a proper risk-set hampers attempts the accurate denomination of the phenomenon. The problem is not merely descriptive; there has been little consensus in theoretical frameworks that examines entrepreneurial entry. Entrepreneurship reportedly constitute the "multitude of necessary, but not sufficient factors." (Shane and

Venkataram, 2000) While significant scholarly effort has been applied to discerning micro-mechanisms in various disciplines that explains some portion of the phenomenon, these multi-disciplinary factors individually have demonstrated little to no predictive power.

At the same time, the economic and social implications of entrepreneurship and enterprise are reportedly large for such an academically slippery and elusive construct. This (the implications and elusiveness) is especially so in the realm of technology entrepreneurship. As of writing, we are witnessing (have witnessed) a shift in the economic focus of the United States from the East to the West coast as the technology sector slowly becomes the dominant driver of technical innovation, economic growth and social change. In a study published by the Kauffman foundation, the high-technology sector, driven by the robust ecosystem of private equity and venture capital, exhibit an impressive growth in the rate of venture founding: the rate of high-tech firm founding is 69% higher in 2010 since 1980, while the rate of private company founding for the same duration is 9% lower. The same study reports that these high-growth tech-startups play a disproportionate role in job creation; the number of new jobs initiated by these high potential firms greatly offsets any job losses attributed to early business failures; evidently, the labor and economic effects of high technology entrepreneurship demands study and examination. Yet the bulk of existing entrepreneurship literature fails to both conceptually and empirically distinguish such high potential ventures from smaller scale (but equally worthy) forms of independent contract-based self-employment. Distinction between these varieties remains fuzzy (e.g. Hsu et al. 2007) but proves to be especially crucial when we consider that entrepreneurs are drawn from the two tail-ends of the wage spectrum (Elfenbein et al. 2010).

This dissertation addresses these issues in two parts. I seek to establish a theoretical framework for understanding and analyzing entrepreneurship (and the incidence of), thereby facilitating the systematic characterization of the individual antecedents of high technology entrepreneurial entry. I begin by drawing upon the sociological conception of the individual *career*: the dual construct of both a person's movement through the fabric of social space-time, and the sense-making that crystallizes these passages into identities (Goffman, 1959; Hughes, 1958). Sociological career theory thus forms a theoretical rubric to systematically address the otherwise murky and vague phenomenon of entrepreneurial entry. This approach brings certain observational clarity by conceptualizing diverse types of entrepreneurs as different clusters of career and identity attributes. Simultaneously, career theory provides a clear temporal point of entrepreneurial entry: the self-declaration of entrepreneurial identity.

To do this, I have assembled a novel large, novel, unstructured dataset of over 2 million resumes. Resumes provide a convenient collection of self-reported employment and educational histories; online resume and *curriculum vitae* repositories provide the infinite shelf-space required to document often fully the extended histories and passages of urban life. The bulk of resume data is constituted by unstructured text: titles, majors, institutions, companies, organizations, roles and responsibilities. The richness of these textual reports enables the observation and measurement of not just career phases, but also self-presented identity. By considering the beginnings of entrepreneurial identity, the data set as collected allows for an examination of entrepreneurial antecedents that is free from survivorship bias and risk-set definition issues.

In the first part, I analyze transition rates from multiple labor market statuses (employed, student, unemployed) to two entrepreneurship-related states: self-employment and investor-backed (primarily, technology-based) company founder. By considering the considerable number of textual descriptions of founding job titles, I model the differences in the identity claims of the self-employed *vis-a-vis* venture funded founders by supervised machine learning. This approach demonstrates not only the differences in the phrases and terms that constitute the two founding identities, but also develops a predictive text classifier to identify otherwise ambiguous founding claims ("founders" who have not received venture-funding nor identify as self-employed). Intrinsically, conceptualizing the founding moment as an identity shift captures much of the early founding activity that addresses the issue of "*when*". I similarly exploit this granularity of resume data to structure employment and education histories, applying unsupervised clustering and rating algorithms to job descriptions and skill tags to classify and rank the overwhelming number of employment role and education qualifications. I propose that the two classes of founders not only invoke different identity claims, these claims are a product of starkly different social positions. This essay thus addresses the dual issues of *what* and *when* by unpacking the objective and subjective faces of the careers leading up to entrepreneurship (Barley, 1989). In so doing, I demonstrate that entrepreneurs across the spectrum represent fundamentally different people with different goals and aspirations, establishing the need for both contextual and empirical precision in the declaration of the entrepreneurship dependent variable.

The second part revisits the effect of status via the careers framework by exploiting positive career punctuations. Research into the organizational spawning of entrepreneurs have examined the advantages accorded by affiliation with prominent organizations (Burton et al., 2002). Here, I propose to refine this finding through the study of entrepreneurial entry in the advent of large liquidity events: initial public offerings (IPO) and large scale acquisitions (e.g. Stuart and Sorenson, 2003). Conceptualizing these events as status boosts in a person's career that are localized in both social space and time, I identify the effect of status affiliation from resource gain by considering the entrepreneurial propensity of the alumni of these organizations that have recently experienced liquidity. This is done through a comparison with multiple control groups defined by various matching strategies. I will consider not only entrepreneurial propensity, but also subsequent likelihood of success as defined by the reception of venture funding.

Finally, I conclude by considering the implications and future directions of this research. The potency of the careers framework and richness of the data at hand suggests a research program that harbors several fruitful directions which I will discuss in this final chapter.

CHAPTER 2
OF HOBOS AND HIGHFLIERS: DISENTANGLING THE CLASSES AND CAREERS OF TECHNOLOGY-BASED ENTREPRENEURS

Abstract

Adopting a careers perspective of entrepreneurship, we theorize new venture creation as a phase transition in the course of a career. We then analyze individuals' transitions to founding high-potential startups or entering self-employment in the high-technology ecology in the United States. We first show that machine learning models applied to the identity claims of hundreds of thousands of entrepreneurs can successfully classify types of entrepreneurial activity in the tech sector. Next, in an extensive risk set comprising two million career histories of could-be entrepreneurs, we show that the human capital and career-based antecedents of these two types of entrepreneurship are not just distinct—they typically are diametrically opposed. Results show that not only do these different groups of entrepreneurs, so-called "hobos and highfliers", exhibit stark differences in identity claims, but the individuals who create these ventures depart from fundamentally different social positions and career pathways. We conclude that an overly broad definition of entrepreneurship hampers the accumulation of systematic knowledge, and we suggest that future studies of entrepreneurship must adopt precision in the definition and measurement of the outcome variable.

The literature on entrepreneurship spans a wide gamut. In sociology, for instance, much of the classic work on entrepreneurship considered the act to be a response to blocked economic mobility and restricted access to the primary sector of the labor market. These insights spawned research on small-scale episodes of entrepreneurship in ethnic enclaves (Aldrich and Waldigner, 1990; Portes and Jensen, 1989). In this view, entrepreneurship is a byproduct of economic exclusion. Conversely, others have applied a sociological lens to the creation and evolution of the highest potential, science- and technology-based, venture capital financed, high-growth companies (e.g., Sorenson and Stuart, 2001; Stuart and Ding, 2006; Baron, Hannan and Burton, 1999; Burton et al., 2002). And between these two extremes, scholars have studied many variants of self-employment, small company creation, and the transition from paid to non-wage employment in entire economies (Ruef et al., 2003; Dahl and Sorenson, 2012). Contemplating the literature, an adage comes to mind: if one chases two rabbits, he is unlikely to catch either. Because to the breadth of the literature and the diversity of its tributaries, we believe that the accumulation of systematic knowledge, even with respect to some of the most basic, descriptive facts of the entrepreneurial endeavor, have eluded researchers. In fact, Sorensen and Fassioto (2011) note that the entrepreneurship literature even has failed to reach any consensus on the definition of the term itself. Moreover, these authors express skepticism that consensus is possible, given the diversity of acts of entrepreneurship and the stage of development of the literature.

While recognizing that a straightforward definition of entrepreneurship may be infeasible, the question remains: how do we develop a coherent literature to investigate a phenomenon that dodges our best efforts to define it? A common understanding of "entrepreneurship" truly has resisted pinpointing (e.g. Sorenson and Stuart, 2008). And so we find ourselves at an awkward intersection, in which the apparent magnitude of entrepreneurial activity and its social and economic implications seem never to have been clearer (Sorensen and Sharkey, 2014), but at the same time, the opacity of our theoretical and empirical conceptions of the phenomenon arguably has stalled the accretion of knowledge in the scholarly field.

In this chapter, we develop a theoretical umbrella for understanding entrepreneurship that provides leeway for a heterogeneous set of empirical manifestations. Specifically, we conceptualize types of entrepreneurship as clusters of attributes of individuals' *careers*. A career comprises a person's chronological movement through the fabric of social space-time *and* the sense-making that converts such passages into identities (Goffman, 1959; Hughes, 1958). Sociologists have aptly labeled these two career components, "phases" and "phrases" (Rock, 1979). We attempt to clarify entrepreneurship by jointly considering *ab initio* the entrepreneur's phases: her temporal status passages through social positions; and her phrases: the identity claims she make *vis a vis* an intended audience.

Framing entrepreneurship as heterogeneous but distinctively clustered phases and phrases offers a theoretical unification that does not preclude a coherent empirical analysis. The framework is flexible because each, broad, phase-phrase cluster can be construed to be a category of entrepreneurship that may have very different empirical manifestations and determinants, but ultimately can be understood simply as a one type of entrepreneurial career. This means that the same set of phases and phrases can indicate one type of entrepreneurship and contraindicate another. We feel this is one, and perhaps the only, pathway forward in the literature: a common theory that offers empirical flexibility.

To conduct the analysis, we have assembled a large dataset with a few million resumes of individuals that (broadly speaking) are at risk of participating in entrepreneurial acts in the high technology ecosystem. The data are rich; they offer detailed educational and career histories for more than two million people, which are merged with multiple other data sources to incorporate hundreds of thousands of instances of the transition to entrepreneurship. The analysis makes extensive use of machine learning to parse entrepreneurial acts by types and to classify many ambiguous data elements. As we describe, these tools are essential to codify large quantities of unstructured resume data.

Our findings highlight a fundamental distinction in forms of entrepreneurship in the data that is reminiscent of the colorful intuition that entrepreneurs can be classified as "hobos and highflyers". Specifically, this vibrant nomenclature stems from an empirical hypothesis that entrepreneurs often hail from the two, opposing, tail ends of the wage distribution (Elfenbein et al., 2010). More concretely, hobos are self-employed entrepreneurs who often depart relatively low-wage jobs and may further sacrifice income for the autonomy of self-employment. Conversely, high flyers exit high-wage, high-advancement careers to launch high potential companies (e.g. Hsu et al., 2007). We illustrate that a machine-learned algorithm can distinguish between hobos and highfliers based on a large dataset of the identify claims of entrepreneurs. In regressions of the hazard rate of transitioning to these two different types of entrepreneurship, we then show that the machine-assigned types of entrepreneurship have almost diametrically opposed antecedents. The resounding implication of the empirical analysis is that failure to distinguish by type of entrepreneurial career will produce very misleading findings regarding the underpinnings of the transition to entrepreneurship. Therefore, we conclude that the accretion of empirical evidence in this field of research vitally depends on finer-grained categorizations of acts of entrepreneurship.

Theory: of Phases and Phrases

Phases. The concept of a career has held such sway in sociology in part because it harnesses one of the discipline's foundational assertions: there is an intrinsic duality between positions and their occupants. Social structures are analytic abstractions created through linkages that define positions as recurrent relational patterns in social space. The cornerstone of an enormous amount of research in the field is a description of how characteristics of these structural abstractions are arbiters of the distribution of opportunities and constraints in any arena in which social mobility occurs. Careers, in other words, are one of the most important forms of social structure, and there is every reason to believe that their generalizable characteristics will associate (or disassociate) with some set of entrepreneurial tendencies.

In formulating our theory, we rely on Hughes's (1958) evocative characterization of a career as an intricately twined series of "phasings" and "phrasings" (Rock, 1979). The former refers to the more literal statuses and state transitions that constitute the workplaces and job roles in a career, and the latter, the verbalization of the identity implications of these mobility sequences. Although there are many distinct conceptual formulations of the career (cf. Barley, 1989), all share a core emphasis on a set of positions or statuses that are woven together through well-trodden mobility patterns. In the case of the professions, these may be age-graded, structured pathways into and through occupational certifications, or they may occur in the form of ascending the rungs in intra-organizational career ladders. The central idea is that we can comprehend careers—sequences of positions or statuses and the transitions between them—to be supra to any individual actor.

Scholars of work have richly described the prototypical career patterns according to which (some) individuals advance in organizations (e.g., Spilerman, 1977; Abbott and Hrycak, 1990; Barnett, Baron, and Stuart, 2001). As Zuckerman et al. (2004) observe, however, the extensive research on the structure of internal markets (e.g., White, 1970; Stovel, Savage, and Bearman, 1996) belies a paucity of explorations of the pathways of mobility through the external labor market. This is problematic for a few reasons. First, as a general matter, there is a trend toward increased inter-organizational mobility (we present corroborating evidence in the descriptive statistics that follow). The metaphor of an internal job "ladder" seems to have become less accurate over recent time: modern work life increasingly is characterized by mobility across organizational boundaries and even occupational jurisdictions. The modern career often comprises not just movement up an organizational ladder, but it contains multiple passages between the precincts of organizations, professions and institutions. The fluidity of these transitions has diffused the newer metaphor of the "boundaryless" career (Arthur and Rousseau, 1996).

A second issue is the extensive incidence of entrepreneurship itself. As new, comprehensive datasets have become available, scholars have realized that entrepreneurship (in its heterogeneous forms) is in fact a very common form of career transition. Ferber and Waldfogel (1998), for example, estimate that as many as a quarter of the men in the US workforce undertake some form of entrepreneurship prior to their mid-30s. As Freeman (1986) and many since have noted, entrepreneurship and inter-organizational mobility generally are two sides of the same coin. Because the great majority of new ventures are spawned by actors who depart from an incumbent organization (e.g., Burton, Sorensen and Beckman, 2002; Sorensen and Fassiotto, 2011), a high incidence of entrepreneurship in the economy is tantamount to frequent episodes of inter-organizational transitions. Of course, entrepreneurs are the initial links that connect existing organizations to newly created ones (e.g., Phillips, 2002, 2005).

If a career is a set of linked phases, with each one characterized as a nexus of positions in distributions of occupation, specific job role, type of employer, and so on, then it is easy to see that modern careers will exhibit highly variegated patterns. In fact, we have good reason to expect particularly significant variability in the careers of entrepreneurs. First, as Burton, Sorensen and Dobrev (2016) observe, careers in traditional professions often follow prototypical sequences. The pathway to becoming a doctor, for example, entails a timed, sequenced, and institutionalized set of positions that are required to obtain certification and to progress through the career. Conversely there are no specific prerequisites or life stages that necessarily predate the transition to entrepreneurship.

Returning to the introductory section, a second reason to expect heterogeneity in the careers of entrepreneurs is that there are vast differences in types of entrepreneurship. Just as we understand that the phases leading into and through the life-course of the career of an attorney will differ from the statuses and transitions characteristic of a physician, we also anticipate differences in the prior careers of the self-employed relative to founders of, for instance, biotechnology companies. In fact, just this type of distinction is made in a number of papers that highlight a distinction between necessity-based (e.g. Borjas and Bronars, 1989) and opportunity-driven entrepreneurship (e.g. Burton, Sorensen and Beckman, 2002; Shane and Stuart, 2002; Sorenson and Stuart, 2003; Stuart and Ding, 2006). Likewise, a difference in career antecedents is directly implied in the empirical postulate that entrepreneurs are more likely to be "hobos or highfliers". In other words, the transition rates to entrepreneurship are higher at the tails of the income distribution than in its center. More recently, scholars have presented a variety of frameworks that are intended to categorize "types" of entrepreneurship (Sorensen and Fassiotto, 2011). The heterogeneity of the entrepreneurship phenomenon/phenomena lead us to postulate:

Proposition 1: Entrepreneurs of different types will exhibit significantly different career antecedents. They will transition to entrepreneurship from systematically different points of departure, including specific job roles, educational and professional histories, and life phases.

In short, distinct types of entrepreneurship will correlate with different types of predecessor careers. As a general matter, career passages of certain kinds presage different types of entrepreneurial transitions.

Phrases. Self- and social perceptions of identity change as individuals transition along the different corridors of a career. Many of the major bodies of theory in sociology touch on the identity shifts that are concomitants, precursors, or consequents to life's status passages. Indeed, a core premise of symbolic interactionist perspectives is that there is reciprocality between self and society; the self mirrors interactions with a structurally differentiated society, which provides the shared understandings and vocabularies that constitute the ecology of social roles and identities that exist in a given time-place (Mead, 1934; Stryker, 1980). Actors develop multiple identities for each of their distinct positional and role designations in life, such as "mother", "teacher", and so on.

Theories of identity draw on the fact that the social world comprises classificatory systems, and the labels attached to classifications convey meaning in the form of shared understandings and expectations for behavior. These labels are both the means by which we recognize one another as occupants of particular status positions and they are the basis on which we form behavioral expectations of others. Though time-stationary ascriptive characteristics—predominantly gender and race—do greatly influence self- and social perceptions of identities, there also has been much thought on what causes identities to change. As status transitions occur, individuals adopt new roles and then experience a change in their conceptions of self, which turns on the process of labeling the attributes of one's new status. In Hughes's work, the identities tied to phase transitions are described as "phrases"; these are the language shifts that align and reconcile changes in roles to shifts in identities.

But phrases are not simply conceptions of self-identity; they too are used by external audience members to classify and stratify the actors they evaluate. In the economic sociology and entrepreneurship literatures, much of the work on identity concerns how and why entrepreneurs proffer specific identity claims. Throughout broad literatures in institutional theory, organizational ecology, categorization processes in markets, and cultural sociology, there is a view that established categories and cognitive schemas provide the building blocks of a "cultural toolkit" (Swidler, 1986) that actors can invoke to erect identities. In the two-stage models of audience choice (Zuckerman, 1999), choosers begin by selecting the members of a consideration set and then make a final selection from within it. Construed in this way, one of the critical, early tasks of an entrepreneur is to construct a social identity that functions to admit her to the consideration sets of an appropriate group of resource holders.

Why? The argument boils down to the fact that by definition, all acts of entrepreneurship involve the new (Stinchcombe, 1965). Indeed, in the earliest days, an entrepreneurial venture often is little more than a list of claims. In its formative days, a new entity has yet to act or to do; it begins as a statement of intention (Lounsbury and Glynn, 2001). The uncertainty engendered by novelty causes critical resource holders and would-be customers to be skeptical of the claims of new organizations. This is where the social identity literature comes into play: entrepreneurs aspire to construct identities that resonate with resource-holders (Rao, 1994; Lounsbury and Glynn, 2001; Navis and Glynn, 2010). The cultural language and category systems of a market provide the legitimated domains of activity that can be deployed by entrepreneurs for strategic ends (Rao, 1998; Weber, Heinze, & Desoucey, 2008; Patterson, 2014).

Because entrepreneurs are not beyond the demands of legitimacy and more tangibly, because they must communicate their product or service offerings to the market, we posit that the labels that entrepreneurs invoke to describe their ventures will significantly vary by type of venture. Constrained by the nature of the opportunities they pursue, entrepreneurs must choose language that conforms to archetypes and market categories that pre-exist in audience members' mental models. Freelancers, for instance, must gain entry to the consideration sets of would-be clients of small-scale services. Conversely, venture founders aim to appeal to would-be angel and institutional investors and potential, early hires. As such, we anticipate that founders of these two types of organizations will choose to present themselves with very different identity claims. Not only will different types of entrepreneurs have travelled through different career phases that engender distinctive constellations of self-identities; they will also be in pursuit of strategically distinct, public identities. This leads to our second proposition:

Proposition 2: Entrepreneurs' identity claims can be used to categorize entrepreneurial activity into distinct types.

In terms of the subsequent analysis, we propose that Freelances and Venture Founders, the focus of this empirical analysis, will exhibit very different identity claims. They will describe themselves with different language, and the linguistic choices will be sufficient for a machine to learn to assign entrepreneurs to specific types.

Data and Methods

The classification of entrepreneurial careers begins with the identities of the entrepreneurs themselves. We have proposed that entrepreneurial identities are not *a priori* injections but rather, they coalesce as careers evolve and intentions form. Concordant with this view, we model entrepreneurship as a career transition. As such, we must conduct the empirical analysis in reverse order of the development of the propositions: we begin by examining the embodied social classificatory systems in entrepreneurial claims, which allows us to test Proposition 2. After showing that we can exploit entrepreneurs' identity claims to create a finer-grained classification of entrepreneurship by type, we then set up a set of hazard rate regression models to demonstrate the heterogeneity in career phases that underpin transitions to the two types of entrepreneurship we study. Thus, the logical flow of the empirical analysis reverses the order of the propositions, because the classification of identity claims establishes the state space for entrepreneurial transitions.

For an empirical context, we have chosen to examine the ecosystem of technology-based entrepreneurship. It is of general interest because the creation of financial value and employment opportunities in the sector has been so remarkable. In addition, entrepreneurial activity in technology is quite well documented. Episodes of entrepreneurship in technology also are plenty divers. On one hand, there have been millions of attempts to create very high-potential, outside-investor-backed, high-growth companies. In parallel, there are even more instances of small-scale entrepreneurship, in which individual service provides transition from educational institutions or paid employment to create sole proprietorships that sell into the tech sector. Using machine learning, we will first distinguish these two, broadest classes of entrepreneurial activity based on entrepreneurs' identity claims. We will then show that combining them in a single analysis leads to a nearly uninformative picture of the career phases that correlate with the transition to entrepreneurship; it results in an averaging of opposing effects.

The data requirements to conduct the analysis we propose are extensive. Specifically, to avoid sampling on the dependent variable, we must gather a large sample of individuals that constitute a viable risk set for transitioning to entrepreneurship (cf. Carroll and Mosacowski, 1987; Stuart and Ding, 2006). For valid inference, we must observe attempts at founding, in addition to just successful founding events (Aldrich and Reuf, 2006). For purposes of estimation, we require full career histories that are not left-censored, with rich detail on educational and work histories. Finally, we must observe entrepreneurs' identity claims.

We have undertaken a very extensive data collection and processing effort to meet these stringent requirements. The bulk of the data come from three sources. The first is CrunchBase, which chronicles the (mostly technology) startup ecosystem. CrunchBase acquires information from TechCrunch news and a crowd-sourced community with approximately 50,000 participants. To date, CrunchBase lists 320,337 distinct founding events. The second source is AngelList, which has become a very influential online community in technology. A great many individuals who launch technology-related companies create their own AngelList profiles. AngelList has become a broader network of actors in the tech ecosystem, but because it is primarily a market for seed-stage funding, many entrepreneurs create AngelList profiles before or near to the time of inception of their ventures. In addition, the site also retrospectively aggregates data on startups from multiple news sources, creating a “LinkedIn for startup and startup investors.” The AngelList data date back to 1990. It comprises 437,289 founders, investors and employees in the startup social network.

CrunchBase and AngelList provide information about attempts at entrepreneurship. However, they only offer snapshots of founders’ career histories. Furthermore, using only these data providers would amount to sampling on the dependent variable—we would be selecting only the employees, entrepreneurs and investors who self-select into the community. Though both data sources contain information on many individuals who are not aspiring founders, they do not constitute a representative sample of at-risk individuals. To rectify these shortcomings, we obtained public LinkedIn profiles for all individuals in the CrunchBase and AngelList databases, which we then augmented with the profiles of several million additional individuals.

Many features of LinkedIn are attractive for this purpose. First, the public, networked nature of the online resume site ensures a high level of data integrity; LinkedIn members are unlikely to post fallacious career histories, given that the site is public and that individual members are connected to professional associates. Second, because individuals generally post complete career histories on LinkedIn, the database contains full resumes for most members, which means that sampling in the present provides detailed information on members’ previous employers, job titles, and so on. For instance, the average 40-year-old member lists 4.48 distinct employment episodes at 3.99 distinct employers. Third, although the data are unstructured and are completely unusable without very extensive cleaning and disambiguation, public LinkedIn profiles generally include job descriptions and skill tags. These data elements are crucial for the use of unsupervised machine-learned classifiers to disambiguate and systematize employers, educational institutions, job titles and undergraduate majors.

Fourth, for all LinkedIn users, we are able to obtain a list of similar alters. For each individual on the site, LinkedIn provides a list of "People Also Viewed" (PAV). This is literally a structural equivalence network that is constantly (re)created through the search and click patterns of all LinkedIn users. A given alter appears as a "person also viewed" alongside ego insofar as the same third parties view both ego's and alter's profiles. The view network therefore enables us to create a snowball sample of individuals at various degrees of proximity to the at-risk subpopulations of each type of entrepreneur. In essence, the PAV is a means to start with a target sample of entrepreneurs and then to snowball out to the broader LinkedIn membership. Crucially, we can use the PAV at successive distances from a focal individual to achieve a near-random sample of the entire LinkedIn database (e.g., to move two steps from a focal person, we sample the PAV of ego's PAV. In other words, ego(i)-->People Also Viewed(j) alongside ego(i)-->People Also Viewed(k) alongside PAV(j) of ego(i)).

Finally, individuals on LinkedIn report and describe founding events and career transitions that can be cross-checked against other data sources.

Sample

To construct the control cohort that pairs to the cases, we first identified and collected career histories for all individuals in the CrunchBase and AngelList data that we could match to public LinkedIn profiles. We then collected a 2nd degree proximity sample comprising 2,038,064 individuals. By 2nd degree, we mean the two million plus individuals who were the "People Also Viewed" of the "People Also Viewed" of the CrunchBase and AngelList entrepreneurs. We believe that two degrees from an entrepreneur results in an approximately random sample of the LinkedIn community.

The data collection strategy yields a case-cohort structure that forms a (hopefully) representative sample of the technology startup ecology of the United States. One shortcoming we must acknowledge is that there is no feasible way to generate a truly random sample of control career histories and there are no available summary statistics about the true, full, at-risk population. Our assumption is that the twice-removed PAV of entrepreneurs represents an appropriate, random sample of individuals who are likely to have the educational and professional backgrounds that they may feasibly be at risk of new venture creation.

Dependent Variable: Venture Founding vs. Self-Employment

Our empirical strategy is to capture entrepreneurs' self-characterizations to categorize types of entrepreneurial transitions. To create a data set of entrepreneurial identity claims, we searched all LinkedIn job titles for each instance of the following *strings*: "owner", "found", "freelance", "self-employed", "independent", "contractor". This yielded a pool of job titles and accompanying, member-generated, free-text descriptions that characterize the identity claims of each of these probable episodes of entrepreneurial activity. We construct this set of job titles and companies, and then we then use fuzzy merge algorithms to bring in funding data from CrunchBase.

These search strings yield 546,785 "entrepreneurship" job titles and job descriptions among the 2,038,064 resumes in the dataset. Two groups of entrepreneurs are well-defined within these data. First is the set of entrepreneurs that founded a company that we know eventually received venture capital financing or angel investor funding, as documented in

CrunchBase. Henceforth we will call this the sample of “Venture Founders”. Second is a group of individuals who are self-declared, self-employed freelancers. The job titles these individuals use to describe their roles leave no ambiguity about their entrepreneurial intentions. We will label this second group the sample of “Freelancers.” Within the broader pool of 546,745 episodes of entrepreneurship in the data, we observe 33,495 job descriptions of known venture founders and another 133,892 job descriptions of known freelancers. The remaining 379,358 founding events are unclassified; the job titles and supplemental datasets do not provide enough information to code these employment transitions as either “venture founder” or “freelancer”.

It stands to good reason that venture founders and freelancers will employ different lexicons in public self-characterizations of their endeavors. These two groups claim identities to different audiences with heterogeneous concerns: venture founders often wish to interest investors and prospective employees, while potential clients will be foremost on the minds of freelancers. To formally examine proposition 2, that entrepreneurs of the two types will present systematically distinguishable identity claims, we analyze the self-presented claims in the LinkedIn job descriptions of venture founders and freelancers as a text classification problem. If proposition 2 is supported, content analysis of job descriptions should establish a well-defined machine (described below) that will succeed at classifying founders by category from text analysis. In addition, manual examination of the statistically significant text weights that define the classifier should exhibit face validity. As such, a well-performing, interpretable classifier will verify *Proposition 2*.

We proceed with the analysis as follows. First, we create a text corpus based on the identity claims of the two well-defined groups of entrepreneurs. This group of 167,387 unique founders defines the "ground truth"; it is the TechCrunch-verified venture founders, and the neatly self-declared freelancers. We consider these entrepreneurs to be *a priori* classified by type, which allows us to employ a supervised machine learning approach. This group of entrepreneurs form the training data we use to build a machine-learned classifier that then assigns the remaining 379,348 founders of unknown type to one or the other entrepreneurial groups. From the documents of the identity claims of these 167,387 unique founders, we purge common stop-words (“if”, “and”, “the”, “a”, etc.) and then stem all remaining words (“consulting”, “consultant”, “consultation” → consult). This text corpus features 478,321 unique stems and a total of 16,752,285 stem-tokens.

Each document is then reduced to stem occurrences. We do not retain the order of words, which is often called a “bag-of-words” model of documents. Following convention, the stem-counts are then normalized by the total number of words in each document to yield an input dataset with the proportional use of each word stem.

The Lasso Regression Model

Generalized Linear Models (glm) are the benchmark for supervised Machine Learning (ML). Naively, a basic glm classifier runs a logistic regression of outcome (venture founder/freelancer) against the text feature regressors (478,321 unique word stems). This represents the familiar, classic linear regression model, which predicts a response variable $\boldsymbol{y}$ from a matrix of predictors $\boldsymbol{X}$ by estimating the vector of coefficients β:

$$\boldsymbol{y} = \boldsymbol{X}^T\beta \qquad (1)$$

The coefficients can be obtained by solving for the global minimum of the Residual Sum of Squares (RSS) of β for N points, as given by the quadratic function:

$$RSS(\beta) = (\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta) \tag{2}$$

This is also known as the loss function, and has a derivative:

$$\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\beta) \tag{3}$$

Under the standard regression assumptions, solving (3) yields the coefficients β.

Text data, however, pose issues that preclude this specification. First, the data are "short and fat": they contain many more text features than observations (p >>> n). As such, dimensionality reduction through feature selection is necessary. To accomplish this, we exploit the sparsity of text features, which are approximately power law distributed (Newman, 2005). Selecting only stem words that occur more than 10 times in the corpus reduces the number of unique stems by more than an order of magnitude, from 478,321 to 37,271 features. Despite the order of magnitude reduction, the remaining 37,321 features still account for 95% of all stem tokens in the corpus.

Second, text data introduces multicollinearity. The appearance of certain text features will heavily depend on others (for instance, in the setting we study, the stems "hi" and "tech" often will appear jointly). In addition, it is possible that rare features at the tail end of the text distribution might be randomly linearly dependent due to specific idiosyncrasies of the data. In this case, Equation (2) will not have a global minimum but instead a linear space of minimums.

To ensure that the loss function has a generic global minimum, we use a "regularization" technique. Regularization entails adding an additional term to the loss function to constrain over-fitting. Specifically, we introduce a regularization term $R(\beta)$ to the loss function:

$$(\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta) + R(\beta) \tag{4}$$

The choice of the regularization term characterizes the Machine Learning regression model. Here, we employ the LASSO (Least Absolute Shrinkage and Selection Operator) regression technique, which minimizes the L_1-norm of β (Hastie et al., 2009). The LASSO regression adds a L_1 penalty to the loss function with an arbitrarily small tuning parameter λ. The loss function to minimize becomes:

$$(\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta) + \lambda|\beta| \tag{5}$$

The LASSO logistic regression is frequently used because of the efficacy and parsimony of model results. In particular, the inclusion of the L_1 penalty term in eq. 5 will drive certain coefficients to exactly 0. De facto, this represents an added layer of feature selection. The LASSO model solution is thus sparse and serves to highlight the text features that determine differences in the two groups while suppressing statistical noise. This produces parsimonious, interpretable models (Tibshirani, 1996), which is necessary for qualitative assessments of face validity. In addition, the LASSO technique has had success in many Machine Learning competitions[1] and the consistency of its estimates have been rigorously demonstrated in the fields of statistics and machine learning (e.g. Zhao and Yu, 2006).

We construct the dependent variable $\boldsymbol{y}$ such that venture-founding is coded =1 while freelancers are scored 0. The 37,321 text-stem features from entrepreneurs' identity claims form our predictor matrix $\boldsymbol{X}$. We use the LIBLINEAR package in R to select the tuning parameter λ and estimate the model (Helleputte, 2015).

Model Assessment

We assess both the validity of the model coefficients and the performance of the classifier as a prediction algorithm. We detail them in turn.

As discussed, the LASSO drives model coefficients toward zero; significant word features that remain represent conservative estimates of the model. Given the number of repressors, we reject all coefficients with p-values > *0.001*. Despite the stringent threshold, there are still far too many statistically significant stems to report in a table. Instead, we display visualizations of the statistically significant model coefficients in two word clouds. The first cloud illustrates text features that are positive and significant in the model; these word stems predict venture founding. The second presents features that are negative and significant; the word stems in the second cloud identity freelancers. The size of the font in the figures corresponds to the estimated parameter weights. In other words, large-font words in the clouds are most strongly associated with the respective types of entrepreneurship.

The true value of a machine learning model lies in its predictive performance. We adopt 10-fold cross validation to ascertain the performance and validity of the prediction. In other words, we partition the text corpus into ten random subsets. One subset is retained as a test-set and the remaining nine are used to train the classifier. The model is then used to predict venture-founders in the test-set and the results of the prediction are assessed through three metrics.

The first metric is precision. In our case, this represents the percentage of venture founder titles that are correctly identified. Precision drops when actual founders are misclassified as otherwise. If A is the set of all venture founders in the test-set and A' is the set of all venture founders predicted by the classifier, the precision of the prediction is calculated:

$$precision = \frac{|A \cap A'|}{|A'|} \tag{6}$$

Precision measures the fraction of individuals that are identified as venture founders, who in fact are. The second metric is recall, which indicates the faction of relevant cases that are retrieved. Recall drops when actual founders are missed. Therefore, in our case, recall is the fraction of all venture founders in the test set that were successfully identified as such. It is calculated:

$$recall = \frac{|A \cap A'|}{|A|} \tag{7}$$

[1] For instance, the Kaggle-Yelp competition of 2013: “Exploring the Yelp Data Set: Extracting Useful Features with Text Mining and Exploring Regression Techniques for Count Data.” Anonymous, http://www.cs.ubc.ca/~nando/540-2013/projects/p9.pdf

There is a trade-off between precision and recall; improving the accuracy of one comes at the expense of the other. The F1 score aggregates these two metrics to assess the overall performance of the classifier:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{8}$$

The F1 score is then compared with the base rate, which is derived from a "random classifier". A random model assigns venture founder and freelancer status according to the base frequencies of these categories in the training sample[2]. An effective classifier will have an F1 score that greatly exceeds that of the random benchmark.

Model Results

We begin with an interpretation of the model coefficients, which are graphically presented as word clouds. The word weights are illustrated in Figure 1.

Examining the figure, the content of the word clouds very much support the hypothesis that entrepreneurs' identity claims can be used to subgroup the meta-category of "entrepreneur" into more narrowly defined subtypes. First, the very highest weighted features of the text corpus for the venture founder class represent outright declarations of identity: *"found" / "cofound".* Conversely, the highest weights in the freelance class include, *"freelanc" / "independ" / "contract".* Following self-characterization of type, the next set of features that determine membership in the venture founding class are signals of innovation *("incub", "acceler", "disrupt", "enable", "empow", "vision", "pioneer"*) or claims related to intellectual property or technical discovery *("patent", "proprietary", "acquir", "discov"*). In contrast, weights that determine the freelance class are statements of services offered *("inhous", "translat", "redesign", "advis", "write", "shoot", "repair", "assist", "consult"*). The target audiences also feature strongly in the model *(*for venture founders, *"investor" / "round" / "partnership"*; for freelancers, *"client"*). Finally, it is interesting to note that venture founders identify as groups through the use of plural pronouns (*"weve"* for "we've") and businesses (*"marketplace", "platform")*. By contrast, freelancers self-identify as individuals (*"ive"* for "I've") and roles (*"adviser", "writer"*).

A qualitative interpretation of the weights concords with proposition 1: types of entrepreneurs offer empirically separable identity claims. The word-weights indicate that the entrepreneurial claims in the data position along an innovation continuum (Sorenson and Fassiotto 2011). In addition, it is quite apparent that the two groups of entrepreneurs craft identity claims to appeal to different types of resource holders: venture founders seek investors, while the freelancer seeks clients. Not only do the significant features exhibit face validity; they also align with extant field and theoretical insights regarding the entrepreneurial high-low innovation spectrum.

[2] The random classifier will have the property: $precision = recall = F_1$, where the precision rate is the base rate of occurrence of the class in question.

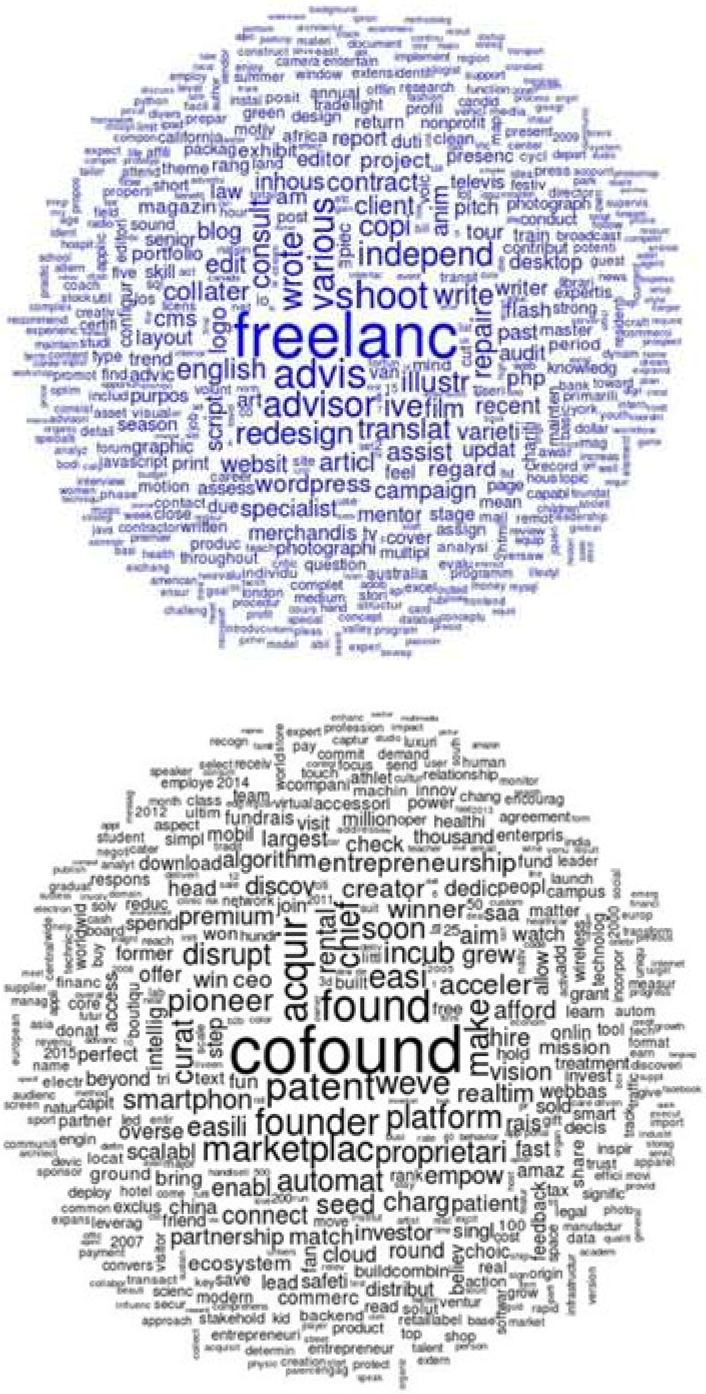


Figure 1: Word clouds representing feature weights for LASSO regression model. Negative weights imply self-employment (above); positive weights imply venture founding (below). Weights shown statistically significant at *p > 0.001* level.

Prediction Results

The LASSO logistic model scores the test set in the interval [0,1]. The distribution of the assigned scores is depicted in Figure 2. This provides a second sanity check: binary classification predictions should be distinctively bimodal at 0, 1, with the modal frequencies reflecting base-rates. This is exactly what we observe in Fig. 2.

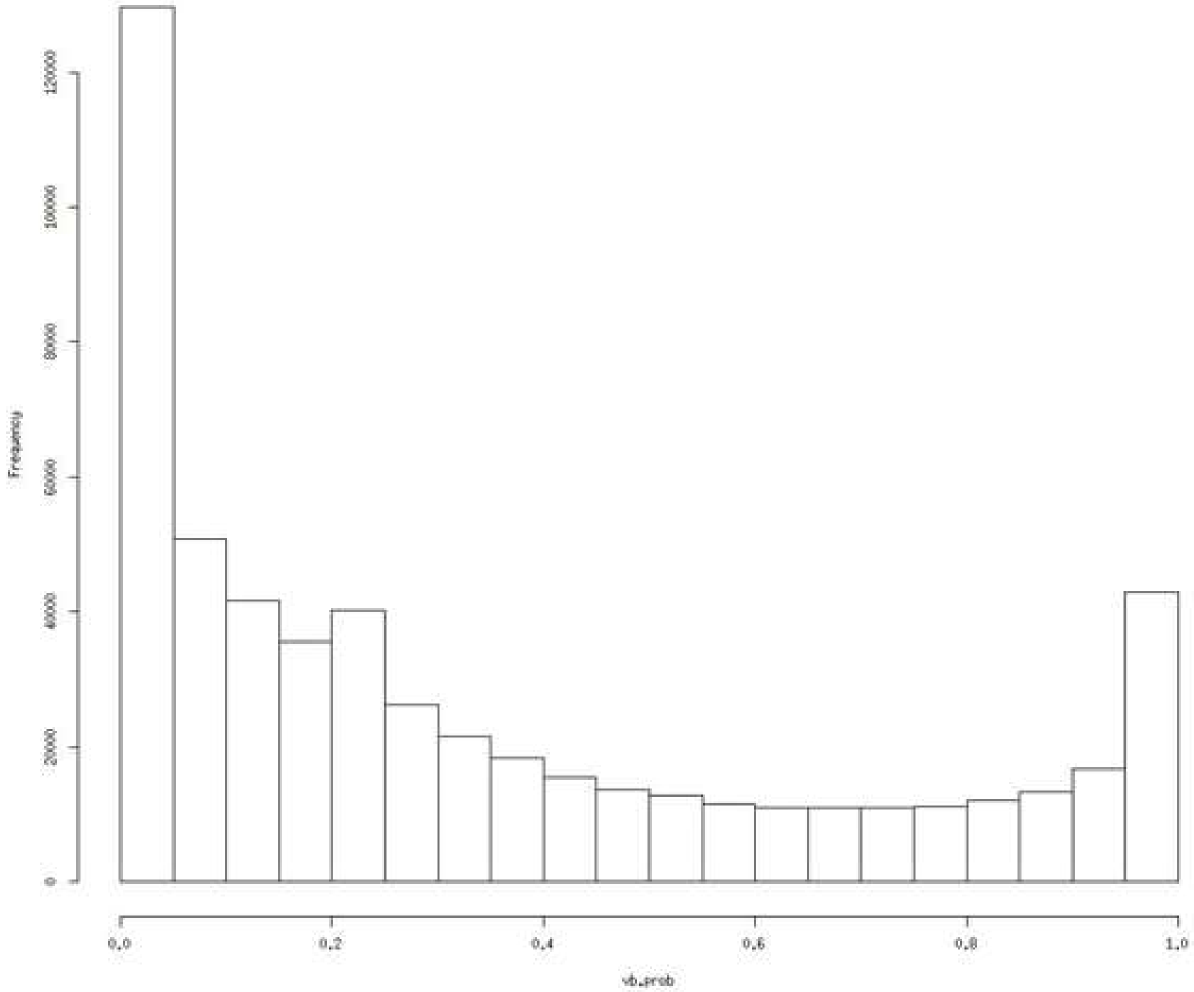


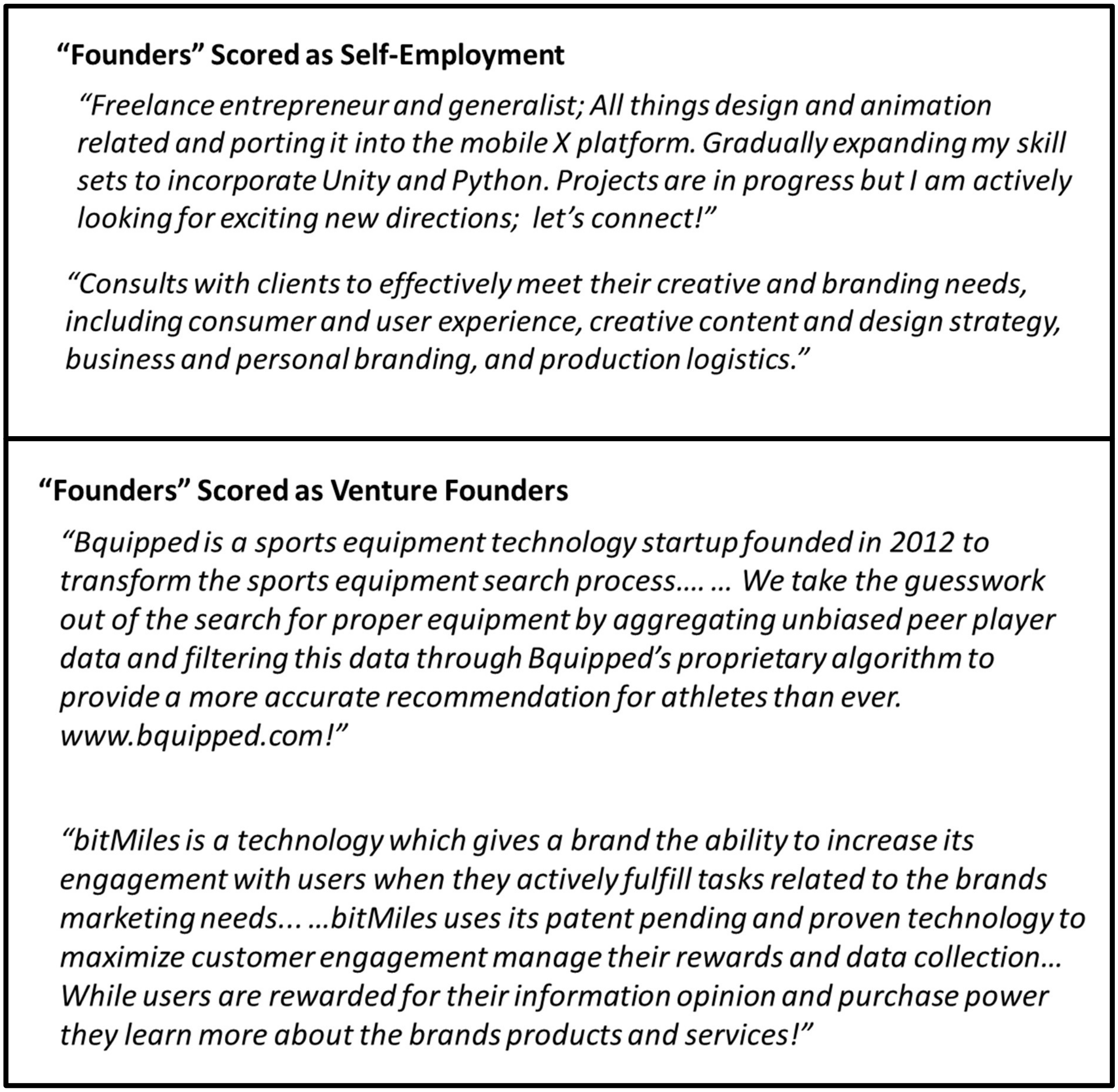
Figure 2: Distribution of predicted scores for ambiguous founding events. Bimodal distribution: scores reflect probability of event being venture founded (as opposed to a null of self-employment).

To assign binary classification, we use the LIBLINEAR default cut-off of 0.5 as the threshold. Founder titles scored above 0.5 are deemed venture founders while those below, freelancers.[3] The results show that the two groups do indeed exhibit different and distinguishable identity claims. 10-fold cross-validation yields a precision score of 80.7% and a recall score of 74.2%. This gives an F1 score for the Lasso Regression classifier of 0.773--much higher than

[3] The choice of scoring cut-off demonstrates the precision-recall trade-off. For instance, higher score cut-offs (e.g. 0.9) will greatly increase the precision of identifying venture founders, but will have considerably lower recall.

that of the random classifier, which is 0.195. This is an especially encouraging result when we consider that a number of the founder descriptions are brief and therefore do not provide much information for assignment to type.

With strong prediction metrics on the test-set, we then run the classifier on the remaining 379,358 founding job descriptions that are not cleanly defined either as venture founding or self-employment. Qualitative verification of a random draw of classification results show further face validity that the machine has sufficiently learned to distinguish between these two groups. Figure 3 illustrates a few examples of the classification results. Overall, we conclude that the LASSO regression classifier provides good support for the first proposition.



"Founders" Scored as Self-Employment

"Freelance entrepreneur and generalist; All things design and animation related and porting it into the mobile X platform. Gradually expanding my skill sets to incorporate Unity and Python. Projects are in progress but I am actively looking for exciting new directions; let's connect!"

"Consults with clients to effectively meet their creative and branding needs, including consumer and user experience, creative content and design strategy, business and personal branding, and production logistics."

"Founders" Scored as Venture Founders

"Bquipped is a sports equipment technology startup founded in 2012 to transform the sports equipment search process.… … We take the guesswork out of the search for proper equipment by aggregating unbiased peer player data and filtering this data through Bquipped's proprietary algorithm to provide a more accurate recommendation for athletes than ever. www.bquipped.com!"

"bitMiles is a technology which gives a brand the ability to increase its engagement with users when they actively fulfill tasks related to the brands marketing needs… …bitMiles uses its patent pending and proven technology to maximize customer engagement manage their rewards and data collection… While users are rewarded for their information opinion and purchase power they learn more about the brands products and services!"

Figure 3. Examples of classified ambiguous "founding" events: Self-employment (top) versus Venture Founding (bottom).

Phases: State Passages through Social Positions

Careers represent sequences of social positions across the lifespan that can be represented as event histories. After defining a machine to classify founding events based on identity claims, we then examine entrepreneurial entry in a competing risks, non-repeated events framework. In each employment spell, individuals in the sample can participate in the labor market in some form other than entrepreneurship, they can experience an interval of unemployment or education, or they can transition to self-employment or venture founding. We analyze the rate of entry into the two types of entrepreneurship as a discrete time hazard rate:

$$P_{ikt} = \Pr[T_i = t, K = k \mid T_i \geq t, \boldsymbol{x}_{it}] \tag{9}$$

where P_{ikt} is the probability that individual i enters entrepreneurial state k at a particular age *t*.

We model this hazard rate as a linear probability (LPM). A primary benefit of the LPM over more traditional logistic regression models is its ease of interpretation. Coefficients are interpreted as straightforward additive increments over the base hazard rate. Specifically, we estimate:

$$P_{ikt} = \alpha_{t\tau} + \boldsymbol{\beta}^T \boldsymbol{x}_{it} \tag{10}$$

Where $\boldsymbol{\beta}$ represents the coefficients to be estimated and $\alpha_{t\tau}$ is a constant given by:

$$\mathrm{\alpha_{t\tau}} = \alpha_0 + \alpha_1 t + \alpha_2 \tau \tag{11}$$

In eq. 11, t is the age of the individual (see below) and τ is the calendar year. In other words, the regression includes a full suite of person-age and calendar-year fixed effects. The person-age dummy variables are tantamount to a non-parametric specification of the baseline hazard.

Because individuals' birth dates generally are not reported in the data, we approximate person age *t* as the number of years from college graduation. This requires us to remove all individuals who do not report a year of college graduation. We set the age clock *t* to 0 at graduation. However, to account for any entrepreneurial activity prior to graduation, we extend the clock backwards by 5. As such, *t* begins with -5 and extends until an event is experienced or the individual is right censored at $\tau = 2014$.

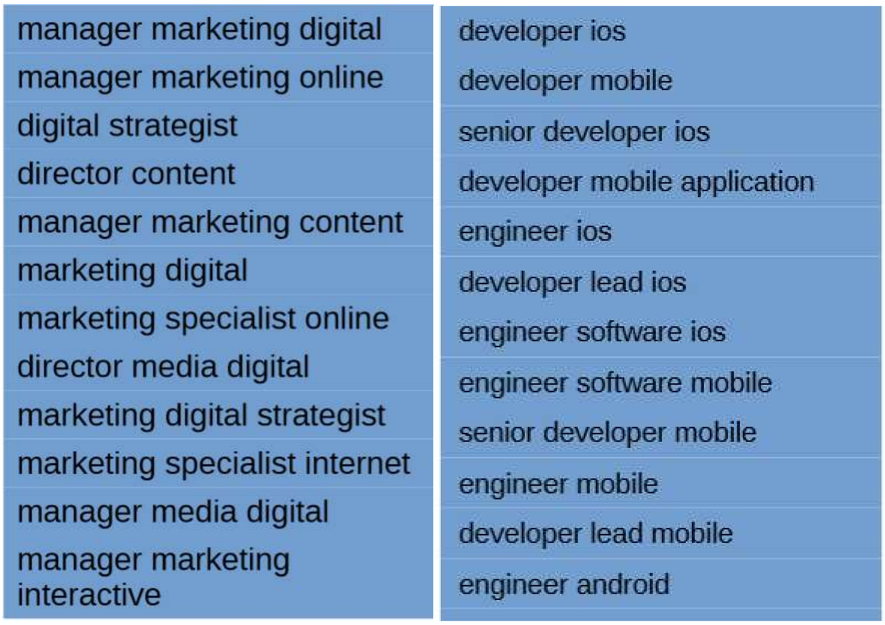
A caveat of public resume data is that not every individual fully lists all education *and* employment phases. We limit the analysis to cases in which we possess full career histories (Klein and Moeschberger, 2005). We also exclude individuals that exhibit an employment and education of gap that exceeds three years between their undergraduate college degree and the beginning of their next, listed career phase.

Finally, the validity of the meta-data on venture founding from AngelList and CrunchBase is most reliable after 1995. Because of this, we subset our sample to consider only cohorts that graduated from college in the years 1990 and later. In effect, this means that we right censor the career histories of non-entrepreneurs at an approximate, maximum age in the mid-40s.

After the imposition of these filters on the data, the initial pool of more than two million resumes shrinks to 881,199 individuals who graduated college and provide complete resume data. Within these 881,119 resumes, we observe 1,235,052 unique job titles; 395,720 unique education majors; and 12,375,284 person-year observations of career states.

Employment Histories: Job Titles

The set of unique job titles manifest the challenges of unstructured data. In the full, pre-filtered dataset, we observe 2.9 million uniquely spelled job titles, which drops to 1.73 million after cleaning. While some of the differences in titles reflect actual differences in job roles, the vast majority result from the multitude of synonyms, acronyms, abbreviations, and spelling errors that are characteristic of unstructured text data. Figure 4 excerpts two examples from the data. The two lists detail processed, unique job titles; a glance suggests that the two lists involve very similar roles and can be in fact grouped together as a single title cluster.



manager marketing digital	developer ios
manager marketing online	developer mobile
digital strategist	senior developer ios
director content	developer mobile application
manager marketing content	engineer ios
marketing digital	developer lead ios
marketing specialist online	engineer software ios
director media digital	engineer software mobile
marketing digital strategist	senior developer mobile
marketing specialist internet	engineer mobile
manager media digital	developer lead mobile
manager marketing interactive	engineer android

Figure 4. Examples of job title synonyms of "digital marketing" (left) and "front-end developer" (right).

The job tittle data therefore need to be aggregated into larger clusters, but how? One option is to impose a top-down schema to categorize titles. However, the plethora of job titles suggest that any *a priori* categorization schema is unlikely to capture much of the variation in the job roles and responsibilities of the data set. Moreover, the heterogeneity in title word usage is so substantial that this would be a very labor-intensive process. Therefore, we choose instead to use a bottom-up, unsupervised machine learning algorithms to cluster titles.

The critical data element for clustering job titles once again is LinkedIn members' self-characterizations of their work roles. Regardless of how individuals choose to portray their job titles, descriptions that employ common language are likely to refer to similar work roles. To cluster job titles, we first perform a basic cleanup of the data. We create a dictionary of common acronyms (e.g. VP, V. President, Vice President; CEO, Chief Executive Officer etc.) through multiple, iterative, qualitative examinations of the most common job titles. Next, we remove all stopwords from the descriptions ("of", "the", "from") and we run a written-language detection algorithm through the R package textcat (Hornik et al., 2013) to remove individuals that post non-English resumes. We again utilize the power law distribution of word frequencies by purging all words that occur less than 500 times in the pool of job titles. Setting the threshold at 500 occurrences retains 93% of all words used in the corpus. Finally, we alphabetize all job title words (e.g. “ios developer expert” and “expert ios developer” both become “developer expert ios”). These steps reduce the number of unique titles from 2.9 million to 1.73 million.

After cleaning job titles, we process the actual descriptions by stemming words and implementing feature selection (> 10 unique occurrences) to create a multinomial bag-of-words. We then employ Principal Component Analysis (PCA) on the text to project the data on a lower number of dimensions and features. This provides two benefits. First, PCA reduces the dimensionality of our feature matrix by looking at the main components of variance. For job descriptions, we find that 12 dimensions accounted for 80% of the variation; as such, we build our clustering algorithm off these 12 dimensions. Second, the PCA rotation loadings should reveal text correlations that underlie the different job roles in this ecology. Qualitative examination of these dimensions should demonstrate face validity.

Finally, we employed Ward hierarchical clustering (Ward, 1963) to group similar job descriptions and skill tags via their Euclidean distances in description-space. An advantage of hierarchical clustering is that it requires no *a priori* selection of the number of clusters. Another benefit is that the number of clusters and associations can be viewed as a tree, which allows for broader or more specific definitions of job title categories depending on where the tree is pruned. From 1.73 million unique job titles, the resultant clustering algorithm generates 54 clusters at the bottom of the hierarchy tree. 80% of job titles are successfully clustered into roles. We operationalize job roles as a categorical variable with 55 categories: the 54 clusters and a category “unclassified.” A full discussion of Job descriptions, PCA statistics and outcomes is discussed in Appendix A. The number of jobs in each cluster and the top three most frequent jobs per cluster are shown in Appendix B.

Employment Histories: Seniority Rankings

To create a seniority order of job titles, we consider individuals' mobility from origin to destination job titles, either within or between companies. Working on the assumption that the majority of sequential employment spells are episodes of upward mobility, we model each job switch as a game in which the destination job wins over the origin job. For instance, if a “software developer” switches jobs to become a “VP of Engineering”, we model this switch as a game in which “VP of Engineering” wins.

With an average of eight employment spells per person in our dataset, we determine the ranking of each job by an Elo rating system (Elo, 1978). These ratings were first used to rank competitive chess players. Elo ratings depend on both the opponent and the outcome of the game. A win causes the ranking of the destination job to increase, and a loss causes it to fall. Wins against an opponent of a higher Elo rating will cause a larger increase compared to wins against equivalently ranked positions. We execute the Elo rating system with an algorithm developed by Stephenson during the Deloitte/FIDE (world chess federation) Chess Rating Challenge hosted by Kaggle. This is implanted in the R Package PlayerRatings (Stephenson and Sonas, 2012).

The algorithm rates job titles with a score from 1000 – 3000. We bin the ratings into 6 quantiles: $[0,10)$, $[10,25)$, $[25,50)$, $[50,75)$, $[75,90)$, $[90,100]$. 20% of the unique job titles do not occur enough for robust ratings. These titles form a comparison, “unrated” category.

Education Qualifications: Majors and Degrees

While the number of unique education majors reported in the dataset is lower than that of employment titles, the diversity of educational backgrounds remains considerable, reflecting both the range of schooling options and the unstructured nature of resumes. Remarkably, there are 717,120 distinctive education majors in the full dataset.

The strategy for clustering employment titles fails for the classification of education majors. Norms that govern the reporting of educational credentials limit the listing of a person's degree and major. Unlike employment records, which prompt the individual to describe their job responsibilities in a blurb, it is much less common for individuals to describe their educational experiences in their resume in any detail. An alternative source of data for content classification is required.

For this, we turn to skill tags. LinkedIn routinely prompts users and the members of their professional networks to skill tag the actors in the dataset. We record these skill tags and use them as indicators of human capital, which should correlate with major fields of study. A working hypothesis is therefore that a person's education develops her human capital and is thus highly correlated with her demonstrated skills.

Using these skill tags, we preprocess, correlate and cluster education majors in a similar manner as employment titles. In comparison with free-form text, skills are structured and organized. Regardless, skill tags share several similar characteristics with text data. The popularity and frequency of skill tags resembles that of text tokens as they too are power-law distributed. As such, we employ a similar feature selection strategy to exclude the infrequently used word tags from the training dataset.

Preprocessing reduces the number of unique majors from 717,120 to 395,720. The benefits of unsupervised learning are again evident. PCA reveals main variance dimensions of the human capital in our sample. As evidenced by the word clouds in Figure 5 we see that the main principal component describes technical, code-related skills in the positive direction, and management- and business-related skills in the negative direction. The top 12 principal components accounts for 85% of the skill variance in the sample. Once again, we employ Ward hierarchical clustering to produce 24 clusters of majors at the bottom of the clustering tree. Appendix C tables the clusters and their associated majors. Again, we observe strong face validity in this set of results.

The public resumes predominantly report college and post-college degrees. Here, we classify educational degrees into 4 categories: Bachelor's degrees, Master's degrees, Doctoral degrees, and other. As degree information that is not sorted into the first 3 categories exhibit significant heterogeneity, we only consider the effects of bachelor's, master's and doctoral degrees in our model. We treat professional degrees (J.D., MBA. and M.D.) as master's degree with the associated field of study as the major (law, business administration and medicine respectively).

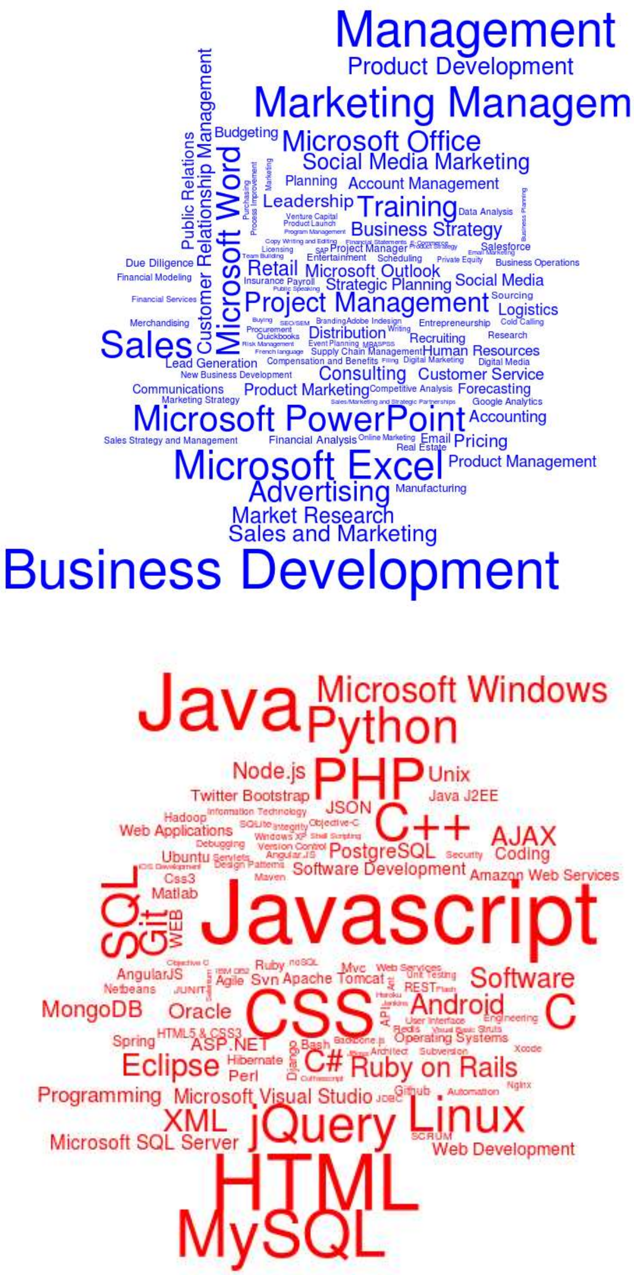


Figure 5: Illustration of first principal skill component: highest 100 positive (top) and lowest 100 negative (bottom) rotation weights.

Results

Base Rate. As expected, the rate of self-employment is almost triple that of attempted venture-foundings. We find that the probability of exiting a current career phase to enter self-employment in a particular year is approximately 1%, while that of exit into venture-founding is

~ 0.3%.[4] To reiterate: each hazard outcome (venture founding and self-employment) is modeled separately. In presenting the results, we note that all the regressions we estimate contain hundreds of dummy variables—we estimate coefficients for every year of person age, every calendar year, every undergraduate major, ever job title, and so on. Therefore, we present results in figures that illustrate critical relationship, rather than tables with too many coefficients to read (complete tables are available on request.)

Person Age. In Figure 7, we observe a stark difference in the hazard rates of the two events we study, venture funding and self-employment, across person-age. In interpreting the age results, recall that we (arbitrarily) set Age=0 to be the year of college graduation. Therefore the low rates of entrepreneurship in the years [-3,-1] reflect the incidence of founding events during the years of undergraduate education. The figure illustrates a marked difference in the effect of age on the hazard rate of the two types of entrepreneurship, both in size and relationship. The founding rate for high-potential companies peaks at approximately 8 years after college (at an assumed age of ~30) and begins to fall off thereafter. In comparison, the peak hazard of self-employment occurs the year of college graduation (versus just a slight uptick in the founding rate for high potential ventures at the time of completion of undergraduate studies). Many individuals in the sample hang their self-employment shingle the year they complete their undergraduate studies. The hazard rate of self-employment then monotonically declines over time, as tenure in the paid-employment sector increases.

We note that while the venture-founding curve replicates similar studies on age and firm founding (e.g. Ruef, 2010), the self-employment curve for this population is different. The findings presented here are inconsistent with a widely circulating myth of the "college-dropout entrepreneur", such as the very well cited case of Mark Zuckerberg and Facebook. In fact, we find that the hazard of venture founding at years prior to college graduation is about a quarter of the average rate, post-college. This result lends credence to genealogical approaches to entrepreneurship (e.g., Freeman, 1986; Philips, 2002; Klepper and Sleeper, 2005) in which founders acquire experience at established organizations before departing to create new, high potential ventures.

[4] Note that these base rate numbers are calculated on a common support for both outcomes of self-employment and venture founding. They are based on the pseudo-random snow-balled sample that characterizes the professional technology ecology on LinkedIn.

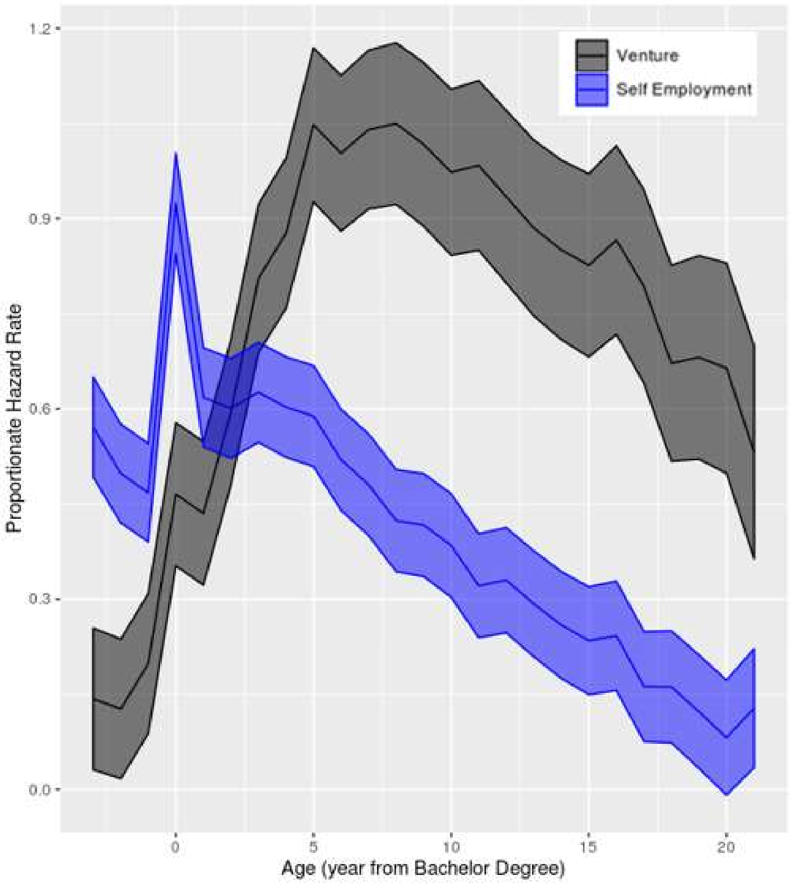


Figure 7. Base-rate normalized age fixed effects across age (time from college graduation) for high potential entrepreneurship vs. self-employment. Fixed effects coefficients are normalized by denominating with the base-rates of high potential entrepreneurship (0.003) and self-employment (0.01) respectively.

Calendar Time. We expect entrepreneurial entry to reflect larger trends of economic and market conditions that significantly vary across calendar time. In particular, the incidence of new entity creation in technology is thought to reflect the booms and busts of the technology sector. Consistent with the theme of differences in the determinants of the two different types of entrepreneurship, we should expect that venture-founding, which is often initiated on spikes of resource munificence during periods of market froth, will reflect these market cycles. Self-employment, which has less-clear intentions and requires many fewer external resources, is likely to be less tethered to broader market conditions.

These trends and differences are in fact reflected in Figure 8. We see the fluctuating incidence of venture-founding during the historical boom periods: the late 90's dot.com bubble and subsequent bubble burst in the early 2000s, and the recent technology start-up boom in the early 2010s.[5] In contrast, the increase in the rate of self-employment across calendar years appear

[5] The contrast in the base founding rates between the first and second tech bubbles may be an artifact of the data. This is because the dataset is filtered to exclude all individuals who graduated from college before 1990. This means that we miss many of the founders in the first tech boom. Because venture funding occurs later in careers than does

to be monotonic: self-employment rates have been steadily increasing throughout the years of the sample. This corroborates extant research that has shown an increase in the proportion of the labor force pursuing contract work and self-employment, with both recessions and bust periods exacerbating the phenomenon (Kalleberg, 2000).

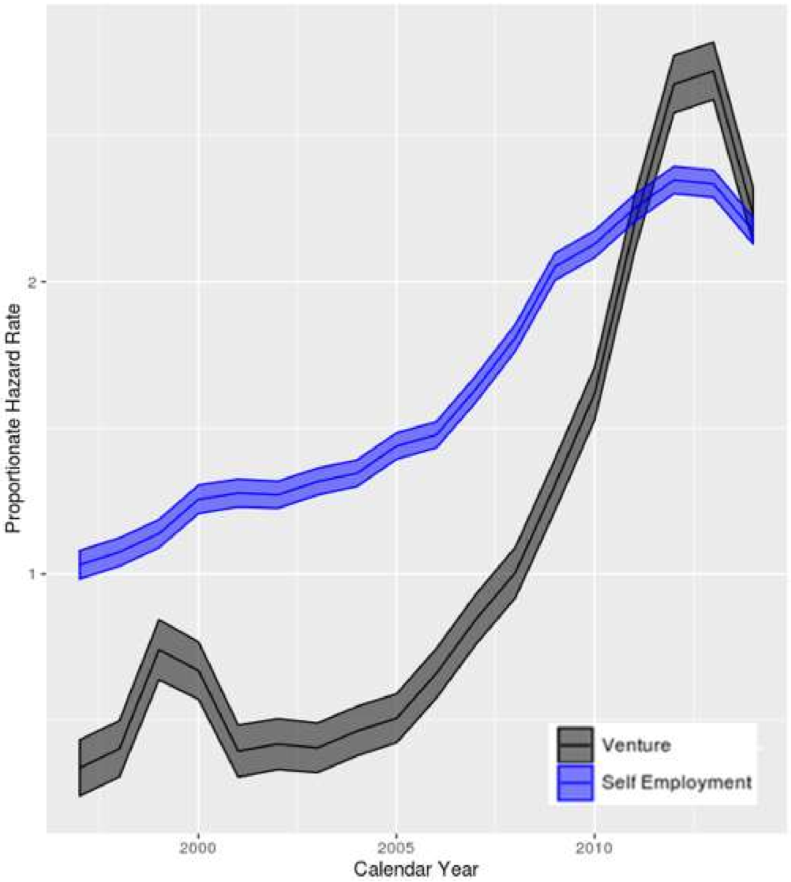


Figure 8. Base-rate normalized calendar year fixed effects across age (time from college graduation) for high potential entrepreneurship vs. self-employment. Fixed effects coefficients are normalized by denominating with the base-rates of high potential entrepreneurship (0.003) and self-employment (0.01) respectively.

Education Effects. Educational level and field of specialization have dramatically different effects on the two types of entrepreneurial entry. In looking at degrees, we compare the possession of a higher degree (master's/doctoral) to the omitted category of having a bachelor's degree. First, we find that the successful completion of higher education has different implications or the likelihood of entry into venture-founding vs. self-employment (Figure 9). The higher the education level, the lower the likelihood of transition to self-employment. Conversely, the possession of master's degree increases the likelihood of venture-founding by about 30%, and that of a PhD by about 20%; versus a falloff in the likelihood of self-employment by 7% and 30%, respectively. These results suggest that venture-founding in general is more likely to require specialized expertise and skills acquired in graduate educational training. Moreover, the

the transition to self-employment, it may be that the results understate the difference in transition rates to the two types of entrepreneurship during the 1990s tech bubble.

negative effect of educational level on self-employment may indicate that investments in higher education create higher opportunity cost trade-offs that deter entry into lower-payoff types of entrepreneurship, relative to higher risk-reward ventures or remaining in paid employment.

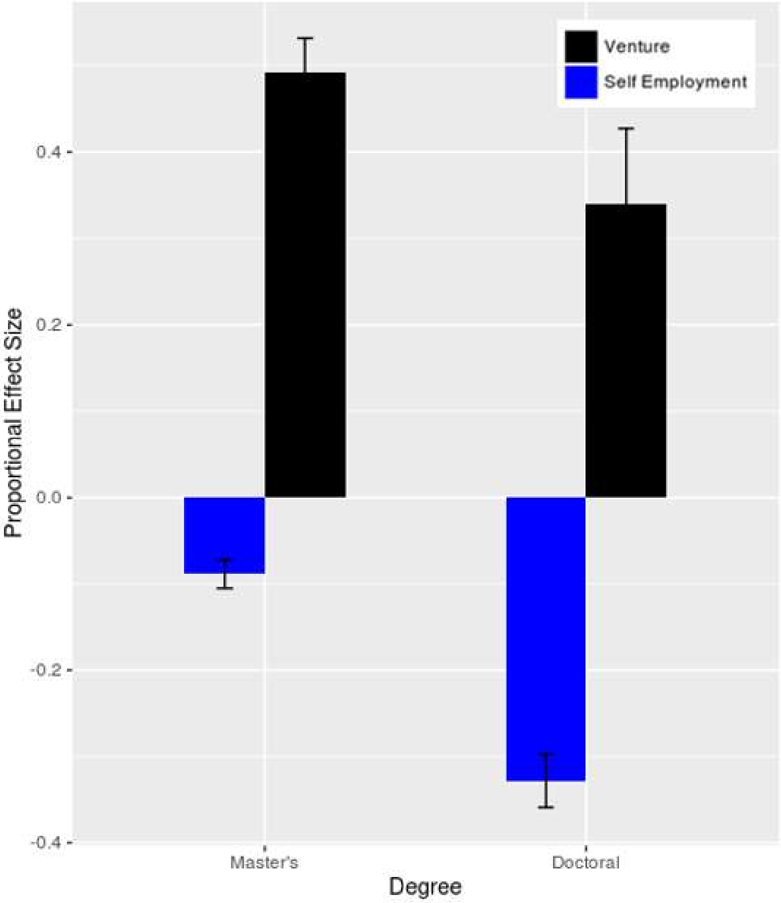


Figure 9. Base-rate normalized effects of higher education on venture founding (dark gray) vs. self-employment (light gray).

Decomposing the education effects to look at clusters of educational majors further illustrates the heterogeneous human capital underpinnings of the two types of entrepreneurship. Figure 10 shows the top and bottom effect sizes on each area of specialization at respective majors for both our outcome variables. All effects here are relative to the omitted category of Economics and Social Science majors. Immediately we see that the specializations of education that inspires venture founding transitions are vastly different from that which drives self-employment. Undergraduate majors that correlate with the highest propensity to transition to new venture creation are directly related to the technical and managerial skills associated with the technology sector. The specialization categories of Computer Science and Engineering, Business Administration and Human Computer Interface/User Experience/Multimedia are high in the likelihood of venture founding transitions. Master's degrees in Business Administration (MBAs) significantly increases the likelihood of transition into venture founding by 148%. Conversely, design and media related majors are more likely to engage in self-employment. The

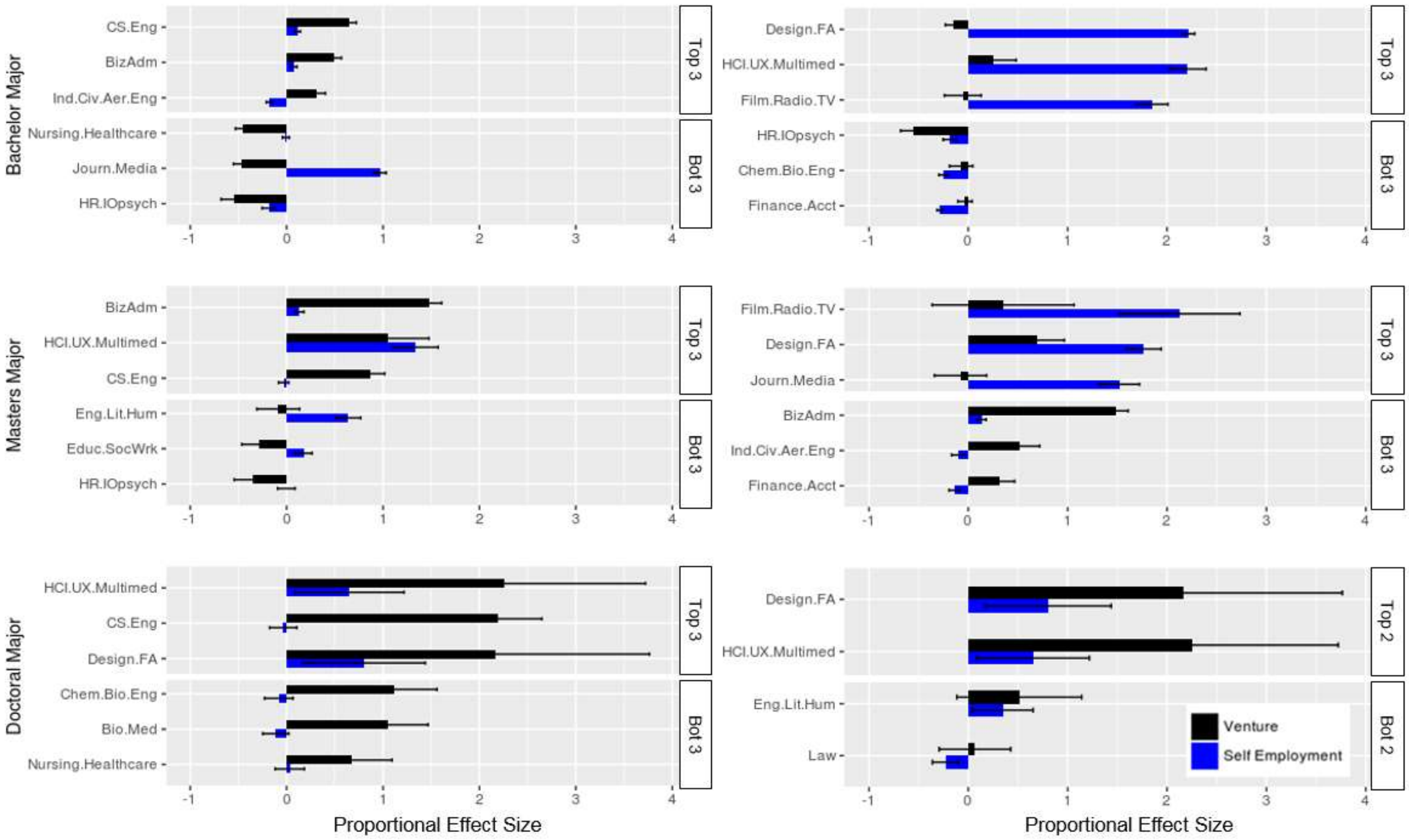


Figure 10. Top and bottom effect sizes of education majors on likelihood of venture founding (left) vs. self-employment (right). Black bars correspond to venture effects; blue bars, self-employment. For clarity, only the top and bottom 3 significant effects for each category are shown *($p < 0.05$)*. Effect sizes here are normalized with a denominator of the base rates of venture founding and self-employment respectively. Comparison (omitted) category is Economics and Social Science. Note that in the case of the effect of Doctoral Majors on self-employment, only 4 majors report significant effects.

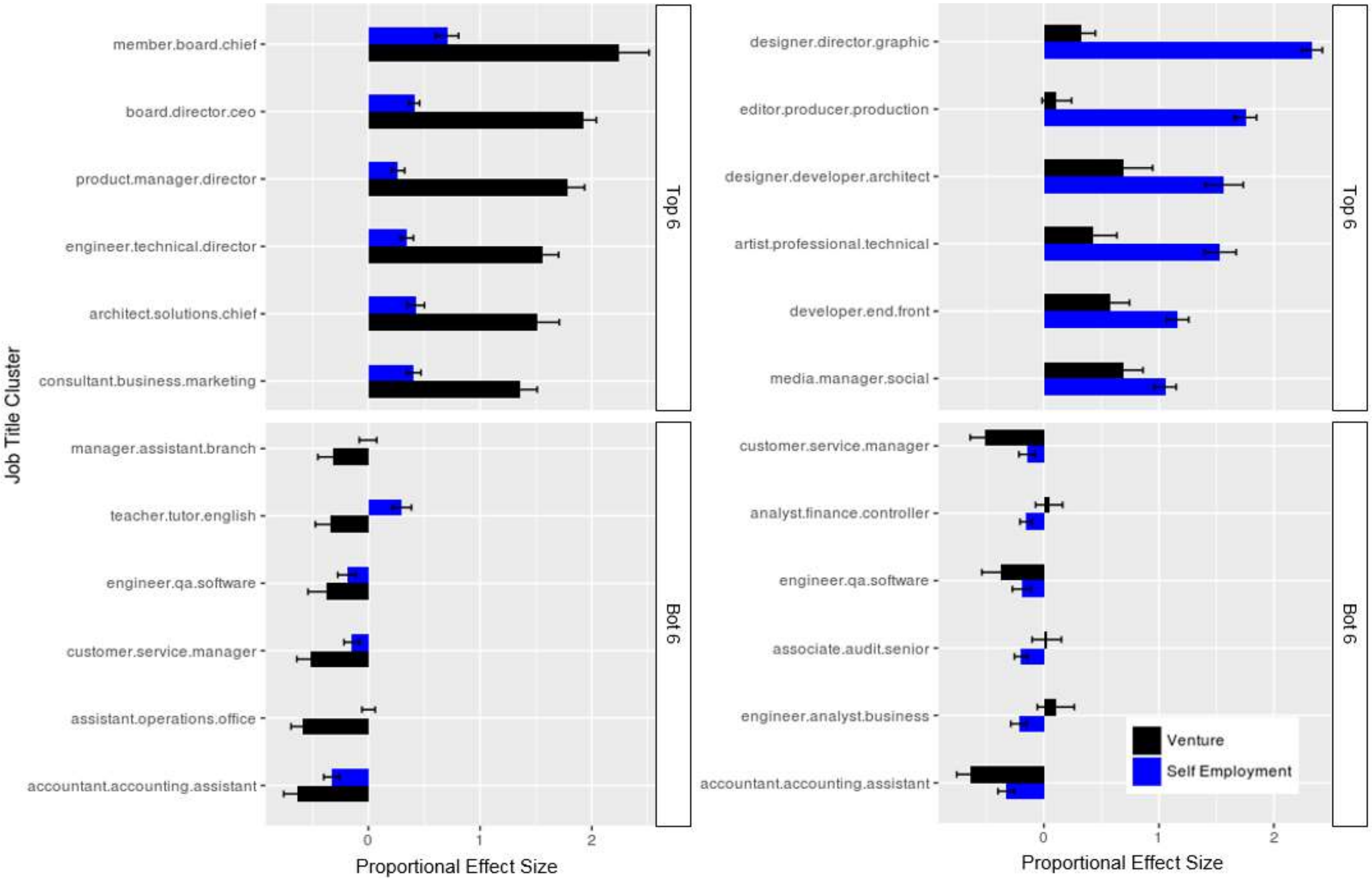


Figure 11. Top 6 and bottom 6 job title categories that lead to venture founding (left) and self-employment (right). Black bars correspond to venture effects; blue bars, self-employment. Comparison category is HR Assistant/Manager. Effect sizes here are normalized with a denominator of the base rates of venture founding and self-employment respectively.

only education major that seems to affect both transitions positively is that of HCI/UX/Multimedia. This interdisciplinary category merges both design, research and software.

Undergraduate majors that correlate with the highest propensity to transition to new venture creation are directly related to the technical and managerial skills associated with the technology sector. At the top of the list, a bachelor's degree in Computer Science and Engineering-related fields increases the rate of venture founding by almost 50%, followed by Business majors at about 40%. Conversely, design and media related majors are more likely to engage in self-employment: a degree in Design and the Fine Arts increases the likelihood of self-employment by over 170%. This is consistent with the fact that a large number of self-employed offer website design, public relations, and related services to clients in the sector.

We note differences between the aggregate effect size trends across the 3 levels of higher education. Even after controlling for the degree of education (Figure 9), we note that the likelihood of transiting into venture founding becomes drastically higher for relevant majors as the education level increases. For instance, a CS Bachelor’s increases the likelihood of venture founding transition by 65% of the venture founding base rate (in comparison to Econ/Social Science majors). This increase rises to 86% and 220% of the base rate at the Master’s and PhD level respectively. In contrast, the trend in effect sizes for the likelihood of self-employment across the degrees is reversed: a Bachelor’s degree holder in Design and Fine Arts gains a 220% relative to base-rate increase for self-employment transitions. This increase drops to 177% of the base rate at the Master’s level, and decreases yet again to 80% of the base rate at the PhD level. In tandem, this suggests that costly investments into human capital results in higher opportunity costs that inspires entrepreneurial activity that promises higher returns, while at the same time drives individuals away from self-employment. (Also note that no significant negative effect of any majors at the PhD level was found for venture founding transitions).

Finally, we note that the results for transitions into self-employment support existing theories of blocked opportunity. Poignantly, investments into Film/Radio/Television at the master’s level and a Humanities or Fine Arts PhD. Both significantly increases the likelihood of self-employment, the latter perhaps a reflection of the paucity of opportunities for PhDs in the academic labor market. In stark contrast, these categories are insignificant for the likelihood of transitions into venture founding.

All in all, the results of the effects of educational backgrounds suggests major human capital differences for the two types of entrepreneurs. They suggest that the varied forms of entrepreneurship likely stem from different skill trainings, opportunity structures and responses to varied opportunity costs gained from training investments.

Employment Effects. We find strong effects of particular job roles and status positions—the phases of a career—on the transition rates to the two types of entrepreneurship. Because current positions define the opportunity cost incurred to leave paid employment for an entrepreneurial pursuit; because prior knowledge acquired in work contexts is a lens for identifying and vetting entrepreneurial opportunities (Shane and Khurana, 2003); because prior job experiences critically contribute to the acquisition of the human capital necessary for entrepreneurship; and because work histories provide many the social networks and social capital that are so vital to resource acquisition in the entrepreneurial process, we expect strong prior employment effects.

Figure 11 show the top and bottom 6 job title categories that lead to both venture-founding and self-employment. The omitted and thus comparison category here is that of HR

Manager. The differences in the likelihood of occupational types leaving paid employment for self-employment or venture-founding corroborates the findings we observe on human capital and education. We find that senior managerial occupations have the largest effect on likelihood of venture founding. After c-suite executives and board members, the next most fecund categories showcase individuals who hold jobs that span both technical and managerial responsibilities (product managers and technical directors). Individuals who hold jobs in the top four categories are on average almost twice as likely to enter into high-potential entrepreneurship. In contrast, career phases that are more likely to be design-, art- and language-based tend to spawn freelancers: graphics and web designers, editorial and production and creative/artistic directors.

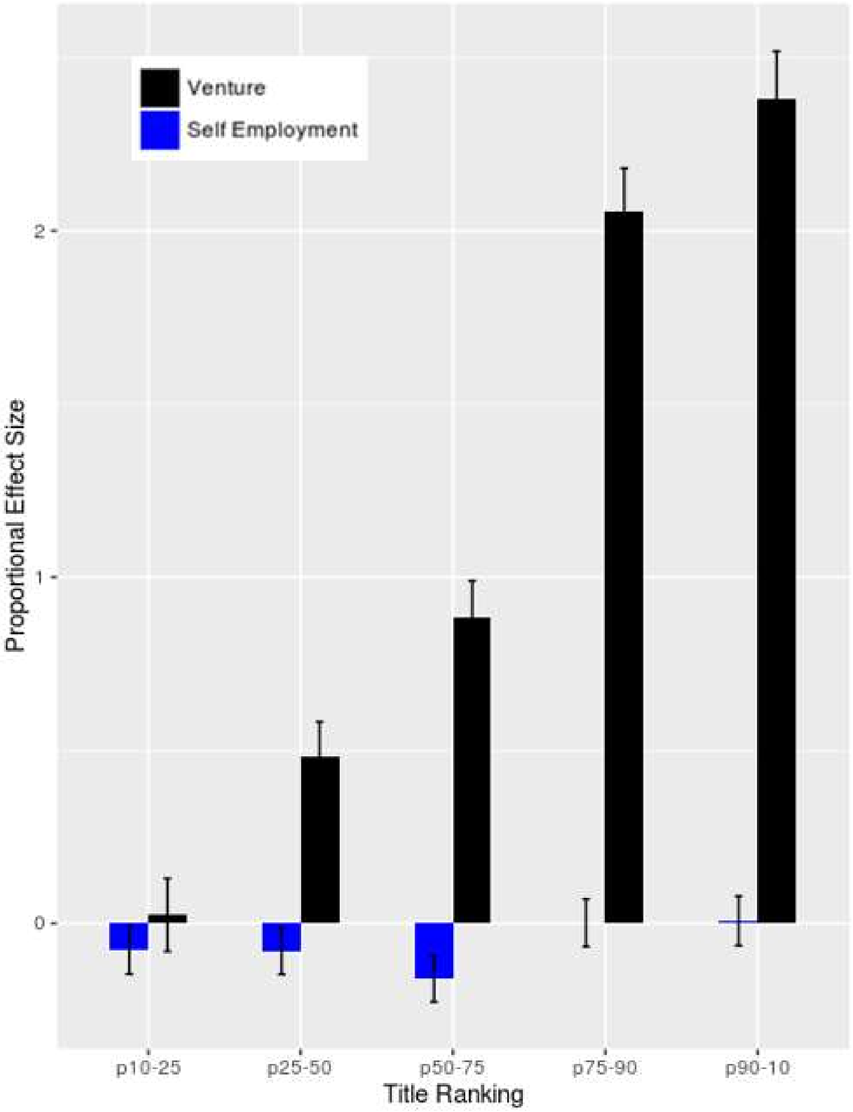


Figure 12. Effect of job title seniority on entrepreneurial transitions across 5 seniority percentile bins. Comparison group is the first decile of job title Elo rankings. Effect sizes here are normalized with a denominator of the base rates of venture founding and self-employment respectively

While variation exists across occupational status and prestige for both self-employment and venture-founding exit rates, the effect sizes of occupational status on self-employment is considerably lower than that for venture-founding. Figure 12 shows the effects of occupational ranking on the two entrepreneurial transitions. The omitted category here is the bottom most decile of job titles. Concordant with extant theory (e.g. Stuart, Hoang, and Hybels, 1999), venture-founders are far more likely to spawn from high status roles: jobs that we identify in the top decile of the occupational status hierarchy based on Elo rankings of the job-to-job mobility matrix increases the likelihood of venture-founding by 238% that of the base-rate. In contrast, the effect of occupational status on self-employment is considerably smaller. A negative effect peaks at the lower-middle 25 percentile (p50-75); here, the effect size "peaks" at a negative 16% for jobs that are categorized as slightly below average. The effects of higher than average occupational status on self-employment likelihoods are statistically negligible. This suggests that while there is no direct, clear relationship of occupational status with self-employment, individuals who are in the process of "climbing" the career ladder are much less likely to transit into self-employment statuses, reinforcing the opportunity (or lack thereof) driven nature of self-employment entry.

Discussion

Entrepreneurship is not a thing. It is a complex, multivalent, set of phenomena. In consequences, a central assertion of our work is that the theoretical edifice for this field of research must exhibit enough plasticity to account for the heterogeneity of entrepreneurship-related phenomena. Following a number of recent authors, we propose that career theory is ideal for focusing and unifying the literature on the transition to entrepreneurship. In this view, entrepreneurship is simply a phase in the state space of many modern careers. We then demonstrate new sources of data and new empirical methods that can be used to conduct much more nuanced empirical investigations of entrepreneurial phenomena.

There are two, core advances in the empirical analysis. The first is the use of a current-day, population-level resume database to create an immense library of entrepreneurs' identity claims. Identity claims provide an extraordinary wealth of information about the type of endeavor, its timing, entrepreneurial intentions, and possibly even insight into entrepreneurs' self-conceptions and psychological traits. We use a machine-learned classifier to partition these identity claims into types of ventures, which allows us to estimate competing risks models of founding events by type of venture. Our second, significant contribution is to present reliable estimates of the correlations between career histories and the transition to entrepreneurship in the high technology sector—and to show how fundamentally this depends on the type of venture. We find that the underlying determinants of the founding rate of high potential ventures vastly differ from the correlates of the transition to self-employment.

Looking ahead to future research on entrepreneurship, we believe that the availability of larger and richer datasets portends a much more rapid development of empirical understandings of the phenomena. In this paper, we have focused on the high-tech sector for a variety of reasons, including its economic importance, its public visibility, and the availability of data sources that enable us to cross-reference and categorize acts of entrepreneurship. These databases were instrumental in expediently identifying the "ground truth" that is necessary to train a machine to assign uncategorized events to type. However, it is now feasible to assemble and analyze broader datasets and (with a few assumptions) to construct risk sets that adequately reflect populations of could-be founders. The data we have collected and cleaned certainly enable many different sampling and estimation strategies.

Likewise, the types of data we have assembled for this project can be used for many new investigations, including offering a first window into how the entrepreneurial process unfolds in the early years of new ventures. For instance, population resume data would enable us for the first time to study the sequence at which organizational departments are built and the pace of growth of new ventures. We could gain the first real, systematic insights into scaling processes in large, representative samples of new organizations. They also allow us to compare the personal attributes of founders versus early hires or to study the re-entry of entrepreneurs into the paid-employment sector if they depart from their new ventures. We can also study financing rounds conditional on founding and the demographic and human capital correlates of the capital-raising process, and we can create a census and point estimates for proclivity of all major employers to spawn new ventures. These are a few of the many projects that can be undertaken with the increasing rich information about initial acts of entrepreneurship.

Returning to a theoretical lens, we believe that career theory offers the most compelling edifice upon which to unite the various strands and conflicting empirical results of the entrepreneurship literature (cf. Burton, Sorensen, and Dobrev, 2016). We find particularly compelling Hughes's notions of the "phases and phrases" of a career, which highlight the dual and reciprocal social processes by which status transitions occur alongside the evolution of self- and social-identities. It is beyond the scope of this paper to present a theoretical agenda for research on entrepreneurship, but we believe that progress in the field will hinge on rallying around a few, umbrella constructs. The alternative to this would likely be a fragmented literature, in which scholars invoke different stands of theory that map to the idiosyncrasies of the context they study and the (usually implicit) definition of entrepreneurship that matches the research setting.

CHAPTER 3
VICARIOUS EXPERIENCES:
ENTREPRENEURIAL RESPONSES OF ALUMNI TO LIQUIDITY EVENTS

Abstract

I examine the effect of status gain on entrepreneurial entry and success by examining different forms of entrepreneurial activity around liquidity events: initial public offerings (IPOs) and large scale acquisition. Although such events accord both monetary and reputation gains for current organizational members, I isolate the effect of status gain by considering the entrepreneurial activity of the organizational alumni. By examining a prospective sample of over 300,000 resumes of graduates of the top 23 science, technology, engineering and mathematics (STEM) colleges in the United States, I find that these alumni are on average 23% more likely to enter into high potential entrepreneurship and 17% less likely to enter into contract self-employment around liquidity events. However, such forms of status gain confer no significant funding advantages to the nascent venture. This suggests that the status gain only serves to lower the perceived founding threshold of the entrepreneur and is largely ignored by external investors.

What drives entrepreneurial entry? While a successful venture accrues the entrepreneur both wealth and status, the notion of firm initiation is daunting (Stinchcombe, 1965); the entrepreneurial process is one of high risk and dubious rewards. The significant majority of entrepreneurial ventures fail to survive the early years (Shane, 2008; Hayward et al., 2004; Freeman et al., 1983); uncertainty characterizes every stage of the entrepreneurship process (Podolny, 2005; Burt, 1992). The doubt that surrounds nascent ventures not only limits the engagement of crucial resource holders, but also holds back the very act of entrepreneurial entry (Stuart and Sorenson, 2008).

This paper focuses on the effect of organizational status on individual entrepreneurial entry. While organizational affiliations have been demonstrated to accord the entrepreneur significant benefits (Burton et al., 2002), their effect on the entry decision is unknown. In essence, I ask these questions: are individuals associated with a successful startup more likely to engage in entrepreneurship? If so, what sort of entrepreneurial activity do they pursue? Are they subsequently more likely to be successful? These questions are answered by considering two factors: barriers and careers. The former refers to a perceived quality and ability threshold required for entrepreneurial pursuits; the latter addresses the issue of why individuals would consider the risky decision of engaging in entrepreneurship. I propose that different entrepreneurial outcomes represent different forms of career attainment: a gain in status affiliations will affect both the decision and type of entrepreneurial entry.

However, the individual variation in organizational affiliations are often confounded with socio-demographic, human and monetary capital variables: the association with high status organizations through an individual's career is hardly random. I examine time-dependent variations in the prominence of past organizational affiliations through positive transformative events: initial public offerings (IPOs) and large-scale acquisitions. These events represent the most idealized outcomes for technology startups: they create large boons of wealth to organizational members and brands the associated founders, funders and first employees with marks of success (Shane and Cable, 2002). To disentangle the effect of monetary and resource capital from status affiliations, I focus on the effect of such events on the *alumni* of these successful startup companies. While "current" employees will likely either hold on to or exercise Incentive Stock Options (ISOs) upon liquidity, the likelihood of exercising ISOs for alumni who have left the company before liquidity will be considerably lower due to taxation laws and risk considerations. This paper exploits this distinction to examine the status effect of a startup's successful outcome on the entrepreneurial activity of the startup's alumni. I conceptualize such liquidity events as fuzzy localized "status shocks" (Azoulay et al., 2012). These shocks accord the alumnus favorable organizational affiliations without monetary gain.

The research setting here comprise the high technology sector in the United States; the study proceeds through the examination of a large archival dataset of over 3 million resumes, supplemented with funding and company level data from the technology start-up data repositories AngelList and CrunchBase. The resumes comprise two main sources. The first is a snowball sample of public career histories drawn from the high-technology sector on resume repository websites such as LinkedIn and about.me. The second is a prospective sample of graduates from 23 elite STEM colleges and universities in the United States of America provided by a human resource headhunting company People.Co. The data advantage presented is threefold: not only does the data provide a relevant risk-set to avoid sampling on dependent variable, the richness and granularity of unstructured resume data allows us to identify not just successful startups, but also discern serious venture founding and self-employment *attempts* as declared in individual resumes. This allows us not only to avoid survivorship bias but also to disentangle the two types of entrepreneurship as mentioned, lending a certain careful precision to the specification of the dependent variable (Levine and Rubinstein, 2013, Sorenson and Fasiotto, 2011; Ruef, 2010). Finally, the presence of full, complete career histories allows us to construct matched samples that create comparison sets of similarly socially situated individuals with comparable human capital.

I make a number of contributions: the first develops the concept of entrepreneurial entry as career mobility (Sorenson and Sharkey, 2014). I propose that in addition to notions of necessity entrepreneurship as a response to blocked organizational and labor market opportunity, positive organizational and labor market experiences can lead to risk taking and the pursuit of high potential high growth entrepreneurship. Secondly, this paper identifies the effect of the experience of a liquidity event on the alumni of the organization. This separates social capital effects from human and resource capital, allowing me to assess the effect of affiliation in a system that heavily relies on status and prestige as signals (Roberts and Sterling, 2012). Finally, to pre-empt a finding, I show a differential between the perception of a status boost to the founder and relevant resource providers, presenting a cautionary tale to the would-be founder who would act on unfounded hubris accorded by status.

The paper thus proceeds as follows: the first chapter begins with a conceptualization of entrepreneurship as careers. The second discusses the effect of status affiliations on founding advantages and career mobility I then present the data, setting and empirical models, detailing the differential effects of liquidity events on the wealth and status of the founding employees. Finally, I conclude by detailing the implications of these findings on the parties associated with the high technology entrepreneurship process and discussing possible mechanisms.

Theory: Entrepreneurial Entry as Career Transitions

While traditional sociological perspectives on entrepreneurial entry have focused on the intersection of resource and opportunity, a recent parallel framework has emerged that conceptualizes the entrepreneurial process as one of careers (Burton et al., 2016; Sorenson and Sharkey, 2014). The two conceptions are intrinsically linked. Individuals with new venture interest fundamentally begin by gaining a knowledge of possible opportunities and subsequently mobilizing the necessary social and monetary resources to exploit them (Aldrich and Ruef, 2006). In urban life, much of these knowledge and resource pipelines are intrinsically coupled to an individual's participation in and movements between organizations. These roles, movements and associations crystallize into an identity. The dual aspect of role and identity constitutes a person's career (Barley, 1989; Hughes, 1958).

In particular, I draw upon the attainment theory of careers. Attainment theory builds upon the presumption that each stage of a person's career represents an attempt to increase her wealth and social status. Thus the proverbial career ladder: each career state a rung to advance and get ahead. Although first studied within singular organizations, attainment theories have been applied to movements between organizations, observing, charting and explaining careers that extend beyond organizational boundaries (e.g. Bidwell and Briscoe, 2010). Consequently, an individual's career outcome is not simply dependent on her own abilities and resources, but also intrinsically linked with the characteristics of her employer and organizations (e.g. Castilla, 2008; Petersen and Saporta, 2004; Baron and Bielby, 1980).

With this framework, entrepreneurial entry is conceptualized as a state, transient or final, in a person's career (Burton et al. 2016; Sorenson and Sharkey, 2014). The survey and observational evidence is compelling: self-employment and firm founding are more and more incorporated into people's careers (e.g. Kalleberg, 2000). The career theory of entrepreneurship gains empirical support from the accumulating evidence that the vast majority of entrepreneurs hail not from basements and garages, but rather established organizations (Sorenson and Fassiotto, 2011; Hannan, 1986). It follows that a person's organizational experience will play a role in the likelihood of her entrepreneurial entry.

Entrepreneurial entry represents an effort to further one's current career status; entrepreneurship presents an option aside from paid employment to develop an individual's wealth and accrue social status. Sorenson and Sharkey (2014), for instance, examines the limitations of organizational opportunity structures: employees exit their organizations to engage in self-employment when intra-organizational advancement prospects dry out. This has parallels with the classical sociological literature on entrepreneurship, which focused on the entrepreneurial activities of under-privileged population segments otherwise blocked from labor market opportunities (e.g. Aldrich and Waldinger, 1990; Portes and Jensen, 1989). Concordantly, such activity impacts mobility and opportunity structures in the larger organizational ecology (Stuart and Sorenson, 2003; Haveman and Cohen, 1992).

A key consequence of this theory is that entrepreneurial entry is contingent on the comparison between the opportunity cost of leaving a current career position and the expected benefits given an anticipated entrepreneurial outcome. This cost-benefit analysis therefore necessarily takes into account not only an assessment of the likelihood of entrepreneurial success, but also the comparison of potential rewards with the current career status. I thus examine the effect of gaining prominent organizational affiliation on entrepreneurial entry through this cost-benefit framework.

Theory: Status Affiliations on Founding Thresholds

The process of entrepreneurial entry is daunting; this is especially so for high potential ventures. In addition to technical skills required for both the conception and realization of a product, the entrepreneur also needs to identify market opportunities, develop organizational technology, and manage both internal and external resources. In addition, the whole process is mired in uncertainty. Indicators of success and positive reinforcement remains far and few; this uncertainty does not only affect market and investor audiences, but also engenders a form of market failure that would deter the very attempt of entry (Stuart and Sorenson, 2008). The decision to pursue an entrepreneurial venture is therefore not only contingent on the would-be entrepreneur's current resources, skills and social capital, but the perceived barriers of entrepreneurial success. Entrepreneurial activity initiates when an individual's own capabilities and resources are self-assessed to be above a certain perceived founding threshold.

As such, this study examines a particular social capital effect – the experience of a status shock from being affiliated with a successful start-up organization – on the perception of the founding threshold. The experience of a status boost might lower the founding threshold by conferring the would-be entrepreneur superior social signals. As individuals move between organizations, they not only bring with them their skills and human capital, they also transfer prior social capital (Dokko and Rosenkopf, 2009). In addition, the status of both the origin and destination organizations affects an individual's choice of destination firm (Rider and Tan, 2014). As entrepreneurial transitions reflect career mobility events, individuals on who experience status gain will not only reassess the level of the founding threshold, but also the state of the entrepreneurial entry.

Consider that affiliations serve as a cost-effective means for audiences to assess both the entrepreneurs and ventures in an environment of uncertainty. To venture investors, the assessment of the quality of the venture proves to be opaque, not in the least due to information asymmetries between the founding-team and resource holders (Stuart et al., 1999). High status organizational affiliations raise the assessments of quality and ability by potential investors and customers (Podolny, 1994). The validity of these social associations as success signals hinges on an inherent assumption that social status has a loose correlation with underlying quality. When the quality of an innovator's technology cannot be easily evaluated, perceptions of the technological innovation are contingent on the status of organizational affiliates (Podolny and Stuart, 1995). For instance, third party investors rely on the high status affiliations of start-ups to infer the quality of these nascent entrepreneurial ventures (Stuart et al., 1999). It follows that an entrepreneur's prior organizational affiliations represent social capital endowments that are critical to the survival of the firm through its various life stages (Shane and Stuart, 2002). At the individual level, the affiliation with prominent organizations rich in entrepreneurial off-spring accords the entrepreneur significant advantages in attracting external financing (Burton et al., 2002).

The alumni who gain status affiliations will not only have access to advantageous network positions, but also gain beneficial status signaling credentials. It has been demonstrated that entrepreneurs understand the importance of these affiliation signals. At the firm level, start-ups sacrifice equity to buy status: higher status venture capitalists on average command a 10-14% discount on startup equity (Hsu, 2012). This strategy has demonstrated advantages. Concordantly, this strategy has positive payoffs that beget success. Stuart et al. (1999) found that biotechnology startups with high status strategic and equity partners on average go to IPO at faster rates and with higher IPO prices. Nascent firms with prominent ties are also found to charge more for products of similar quality (Roberts and Sterling, 2012).

As such, I propose that alumni of the startup that experiences the liquidity event will gain status signaling credentials. This will in turn decrease the perceived threshold of founding. Thus:

Hypothesis 1: Alumni of a startup that experiences a liquidity event will be more likely to pursue entrepreneurial entry.

However, the decision of entrepreneurial entry is made in comparison to the current career state. This implies that the quality of the entrepreneurial state warrants analysis. This point is crucial as entrepreneurial activity runs the gamut from forced necessity self-employment to high growth, high potential and high innovation type companies; entrepreneurs are drawn from the tail ends of the wage distribution (Elfenbein et al., 2011). Consequently, each activity as a potential state in the career will be viewed differently by potential entrants occupying different career positions: while self-employment is characterized by lower status coupled with lower than median wages and higher uncertainty, venture founding validates risk taking with promises of both a gain in social standing and monetary windfall. This line of thought finds support with studies that show, for example, that entrepreneurs with higher prior incomes and management experience are more likely to have growth intentions for their small business ventures (Cassar, 2006). The choice of entrepreneurial activity is contingent on a comparison with the entrepreneur's current career state: experiencing a rare, positive status shock will in turn raise the ambitions and expectations of entrepreneurial reward.

In this examination of technology entrepreneurship, I make an observation that entrepreneurial activity in this sector predominantly comprises two destination states which I will term self-employment (e.g. stay-at-home programmers, advising, independent consultants and freelance designers) and venture founding (the founding of firms that take advantage of new technologies and production techniques, often with the aid of private equity and venture capital funding) (e.g. Sorenson and Fassiotto, 2011). While the latter is associated with high rewards, prestige and status, the former is often a result of necessity. As such, I decompose Hypothesis 1:

Hypothesis 1a: Alumni of a startup that experiences a liquidity event will be more likely to pursue venture founding.
Hypothesis 1b: Alumni of a startup that experiences a liquidity event will be less likely to pursue self-employment.

Consequently, there are two possible venture outcomes given venture founding. A first possible consequence is that the decision to enter is warranted: resource providers in the form of venture investors value such signaling credentials and reward associations with a prominent, high-status and successful startup. The alumni status emboldens this claim: the individual in question was one of the founding employees of the startup and as such, could be attributed credit for its success. All this then translates to real gains in venture funding advantage.

Hypothesis 2a: Alumni that attempt venture founding post liquidity event will be more likely to receive venture funding.

However, the reduced founding threshold might be purely perceptual. While the gain in status affiliations could embolden the ambitions and career desires of the alumnus, this gain is completely independent of the skills and capability of the entrepreneur: the human and monetary capital of the would-be entrepreneur remains the same. In such a case, the gain in status is only perceived by the entrepreneur herself. The quality of the newly founded venture is orthogonal to the status gain and:

Hypothesis 2b: Alumni that attempt venture founding post liquidity event will be less or no more likely to receive venture funding.

Data and Methods

This study imposes numerous requirements upon the dataset. The first issue is size. Both the dependent variable of venture founding and the independent variable of experiencing a liquidity event as an alumnus are incredibly rare events: the data set has to be of a sufficient size so as to allow for their observation in tandem. Second is the issue of risk-set definition. The dataset needs to observe not only founders, but also comparable individuals who are "at risk" of founding. The design which I will propose adopts a matching strategy to find common support between the focal and comparison groups: this will impose stringent sampling demands that will further shrink the sample. Third, the individuals that constitute this data-set should have common human capital. In addition, the data source should detail the career and educational histories of these individuals, allowing us to observe their organizational affiliations, job roles and social status passages. Finally, we require founding metadata: the events and details of founding, funding, IPOs and acquisitions.

To satisfy these criterion, I consider a prospective sample of resumes of 445,648 individuals that represent a draw of the top 23 Science, Technology, Engineering and Mathematics undergraduates (BA., BS., BEng. etc.) and graduates (at both master's degree and doctoral levels) across the years from 1980-2010. These represent an elite sample of technically savvy individuals that are at risk of high technology, high potential entrepreneurship (Hsu et al., 2007). This constitutes the analysis data set.

While resumes present a treasure trove of employment and education histories, the bulk of this information is in the form of unstructured text. Structuring the data will require a deep dive into both supervised and unsupervised machine learning algorithms, methodologies which demand large amounts of data for precision and face validity. To do so, I supplement the prospective sample with a snowball collected sample of over 2 million public resumes that represent the technology start-up ecology of the United States. The source of this data begins with name-disambiguated public profiles self-reported on AngelList, a technology start-up social networking and investment syndication website. At the point of collection, the AngelList data dates back to 1990. It comprises 437,289 users in the startup social network. Beginning with this group of people, I sample structurally equivalent alters by considering LinkedIn's "people-also-viewed" algorithm. This samples on individuals that are related in the viewing network through the collaborative filtering algorithm. A second degree proximity sample nets us 2,038,064 individuals that constitute a sample of founders, funders, freelancers and first employees of the technology ecosystem. This allows us to cluster and classify the predominant job roles, ranks, skills and educational training that are prevalent in the technology start-up ecology. This constitutes the training data set.

Finally, I use CrunchBase to obtain start-up metadata essential for the studying of liquidity and founding events. CrunchBase documents startup funding stage and characteristics from TechCrunch news and through crowd-sourced community contribution with approximately 50,000 participants. At the point of analysis, CrunchBase lists 320,337 distinct founding events.

Dependent Variable(s): Venture Founding vs. Self-Employment

As different entrepreneurial outcomes represent not only different fates and fortunes but constitute fundamentally different career trajectories and identities (Levine and Rubinstein, 2013), I distinguish between two entrepreneurial destination states: venture founding, as defined by the founding of firms that exhibit Schumpeterian characteristics of growth and innovation; and self-employment, as defined by the activity of freelancing and client contract-based labor. In the former, I am interested in high technology, high potential entrepreneurship that so characterizes the recent (2012-2013) technology start-up boom in the San Francisco Bay Area and the dot-com internet era. These founding events are characterized by an active search for investors through the promise of explosive growth and hockey-stick revenue charts. Thus, venture founders will claim activity and firm descriptions that appeal to potential investors who are looking for high innovation, large growth and disruptive potential. For the latter, I observe that the population of self-employed in this ecology largely constitutes freelancers who either service the design and front-end engineering needs of this population (e.g. website/logo design, application/web development and programming), or provide independent consulting and advisory services. This population will claim legitimacy to potential clients: attempts to demonstrate their skills and services in pursuit of establishing trust and quality. It thus stands with good reason that these two groups of entrepreneurs will employ different lexicons in their self-described job descriptions.

As such, I exploit the differences in claimed career identity to delineate the two groups of entrepreneurs (Barley, 1989). I train a classifier on the job descriptions of two pools of founders with known statuses: ventured *funded* founders and self-declared freelancers. From the training set, we have 546,745 entrepreneurial spells (job titles identified/associated with entrepreneurship), of which 33,495 are known to be venture founders from funding data in CrunchBase and AngelList (either being marked as having received more than 100,000 USD in venture/angel funding), while 133,892 have self-declared their self-employment statuses (e.g. freelancer/self-employed/independent contractor/independent consultant). The remaining 379,348 claim ambiguous founding identities (self-described "founders" who do not appear in the startup metadata sets of AngelList and CrunchBase).

I implement a LASSO (Least Absolute Shrinkage and Selection Operator) Regression Classifier, modeling the job descriptions as a binomial bag-of-words (not taking into account word order) (Tibshirani, 1996). The classifier performs well with a precision rate of 80.7% and a recall score of 74.2%. That is, in a test sample of *n* venture-backed founders, the classifier is able to draw *m* venture-backed founders of which 80.7% of them are accurately classified as venture-backed, and this subset of *m* constitutes 74.2% of *n.* Face validity of the classifier weights supports our assumptions and claims of the two groups that comprise this ecology features that determine venture founding class signals innovation through both buzzwords and intellectual property claims, while those that determine the self-employment class focuses on the establishment of known services. Concordantly, a random draw of founding descriptions lends further face validity to the classifier. In essence, the results here mirror that of Chapter 2 (refer to Figures 1 and 3).

I thus measure my dependent variable of founding in the analysis sample through two steps: the first considers founding spells that are already well defined and classifies them accordingly. Founding spells that fail to meet these two criterion in the prospective sample are then classified using the classifier algorithm developed as described.

The test data set observes 48,373 individuals who have engaged in entrepreneurial entry. Of which, 3,664 individuals have received funding, while 13,079 have self-described as being self-employed. The remaining individuals are thus classified, giving a final total of 14,649 venture founding and 33,751 self-employment events.

The measurement of the founding dependent variable as presented brings a certain precision to the phenomenon at hand. More importantly, it also allows us to identify serious, nascent attempts at entrepreneurship. The entrepreneurial process begins usually far before the appearance of the company in any incorporation or startup metadata set; the observation of founding spells on resumes strikes a key balance between frivolous entrepreneurial tinkering attempts (that are not significant or important enough to constitute a career) and survivorship biased observations of founding success and incorporation. The decision to declare an entrepreneurial spell in a resume therefore encodes a serious attempt at entrepreneurial entry.

Independent Variable: Liquidity Events on Alumni

I consider two liquidity events that represent coveted, successful start-up exits: IPOs and large scale acquisitions. These events are documented on CrunchBase and merged into the analysis data set. IPOs and large acquisitions represent the most idealized outcomes for all involved in high potential technology startups. Traditionally, these startups are funded by venture capitalists: angel investors on steroids. Investors in venture capital funds invest in an asset class designed to double or triple their investment (at bare minimum) within a certain timeline that is the life of the fund. "Selling" companies, either as a public offering or to other companies/private equity represents the most desirable outcome of start-ups for venture capitalists. Such events enable them to gain upside for their fund (Zider, 1998).

In addition, these events afford liquidity to the founders and employees, allowing them to convert otherwise illiquid assets into cash, rewarding all involved for their early contributions and hard work that piloted the startup to success (Sorenson and Stuart, 2003). Young, cash-strapped technology start-ups seeking to economize on liquid resources use incentive stock options (ISOs) to compensate employees. These forms of compensation have the added double benefit of employee retention (coupled with a vesting scheme) and aligning employee goals with that of the nascent organization. ISO awards give the employee options to buy shares in the company at an exercise price (strike price), a price usually set at the time of employment. Typically, ISOs in venture-backed startups vests linearly over 4 years: should the employee choose to leave before the 4-year schedule, she is entitled to a linearly prorated share of options. Most ISO contracts also feature a one-year cliff: employees who leave before a year of employment will not be entitled to any options whatsoever. Upon leaving the company, the normative ISO contract gives the ex-employee 90 days to exercise their options before the options are forfeited. In the United States, venture backed companies are well known for the organizational depth and intensity of using ISOs as employee compensation (Hand, 2008). A liquidity event therefore allows technology start-up employees to exercise these options to receive the cash compensation they deserve.

These highly desired liquidity events constitute highly visible, positive transformative episodes: rumors of the potential of such events abound in the venture community and permeates technology and business media. Management and technical team members that are associated with high IPO/acquisition potential startups greatly enhances entrepreneurial opportunities available; these form valuable and rare points of data about the ability to guide the nascent firm towards a successful exit. These associations prove to be extremely attractive to resource holders and investors (Shane and Cable, 2002; Zider, 1998).

However, the nature of ISOs and taxation imply that the monetary largesse is not distributed uniformly. Should an employee choose to leave the company before liquidity, the employee has normatively 90 days to exercise the option. The monetary benefits of exercising an option pre-liquidity is a risky one. Not only is the liquidity of the stock not guaranteed, the spread price – a difference between the market value of the stock and the strike price – is subjected to alternative minimum taxation (AMT). In the pre-liquidity scenario and depending on the spread price and the amount of options awarded, this can imply that the exercising of options (buying and holding the shares) will cost the leaving employee large amounts of cash and subsequently be subjected to alternative minimum taxation for theoretical capital gains.

I consider the independent variable of liquidity "treatment" by considering both IPOs and acquisitions of small and medium sized start-ups. In the case of acquisitions, I only consider deals that are valued at more than 100 million USD. This distinguishes successful acquisition exits from illiquid buyouts.

While these events are fixed in time, the buzz and rumors of these transformative events are considerably more diffuse. These forms the entrepreneurial myths of our day: lay-press and technology blogs often project and highlight the hottest technology start-ups that are rumored to go liquid within the next year. Consequently, a sufficient lag between the decision to leave the startup and the startup's liquidity event is needed if we were to go with the assumption that the alumni are less likely to exercise their ISOs.

As such, I only consider the alumni of these successful startups that have left three years before the IPO or acquisition. In addition, I consider the "treatment time" of prominence gain to be 1 year before the actual year of liquidity. The treatment indicator ticks *1* if the liquidity event is experienced as an alumni and *0* otherwise.

Career Covariates

We are interested in the effect of positive transformative events on the trajectory of careers, as such, career related covariates are needed. Again, a large benefit of this data-set is the availability of employment histories; the downside, once again is that these histories are encoded in unstructured text.

Broadly, we are interested in the movement of social positions and statuses of both employment and education. These reflect not just relative statuses, but are also key determinants and controls for human capital. Concretely, this translates into a measurement of the career roles (job titles and education majors) and ranks (occupational prestige and educational degrees). As details of these methods have been reported in Chapter 2, the following represents a brief summary of these measurements.

Career Roles: Job Titles and Educational Majors

The challenge of processing job titles is both the sheer number of unique job titles as well as the issues of synonyms. This data demands clustering. I segment job titles into groups that exhibit high degrees of relation and similarity. This builds upon the insight that people who do the same jobs will describe them in the same way despite differences in job titles. Here, I adopt a bottom-up approach to use unstructured Principal Component Analysis (PCA) and Ward hierarchical clustering (Nagarajan and Hearst, 2009). This allows us to inductively capture main variations of job roles and responsibilities that comprise the technology start-up ecology.

As usual, a large training dataset sharpens the clusters; I consider the 1.73 million unique job titles in the training dataset. From this, the algorithm generates 54 clusters at the bottom of the hierarchy tree. 80% of the job titles are successfully classified into role clusters. I therefore operationalize job roles as a categorical variable with 48 categories: 47 clusters and a comparison "unclassified" category. The list of job roles that comprise the technology start-up ecology is found in Appendix B.

This strategy fails for the classification of education majors: resume norms only demand the listing of a person's degree and major in the report of educational qualifications. To cluster education majors, I instead employ skill tags on LinkedIn, working on the assumption that people with similar educational experiences will gain similar skills. I apply the same clustering process for job titles to education majors. This produces 24 different groupings of education majors. Appendix C describes the different educational group clusters.

Hierarchy: Employment Rankings and Educational Majors

To ascertain the relative hierarchy of almost 2 million unique job titles, I consider the mobility of individuals as they change job titles either within or between companies. Building upon an assumption that job title switches are made in effort to get ahead, I model each mobility transition as a game in which the destination job title wins over the origin. For instance, if a "software developer" switches jobs to become a "product manager", I model this transition as a game in which "product manager" wins.

With an average of 8 employment spells per system in the training dataset, I then determine the relative player ratings of each job using an Elo rating system. These rating systems are first employed to rate competitive chess players. Recent development has seen an extension of these systems to the rating of participants in other competitive arenas such as the National Scrabble Championship, Major League Baseball, and competitive multiplayer online e-sports. I calculate these ratings using the PlayerRatings Package in R (Stephenson and Sonas, 2014). Subsequently, the ratings are binned into 6 quantiles: $[0,10)$, $[10,25)$, $[25,50)$, $[50,75)$, $[75,90)$, $[90,100]$. 20% of the unique job titles lack the frequency required robust ratings. These titles form a comparison "unrated" category.

The public resumes predominantly report post-college degree information; a minority of resumes report pre-college educational activity (high school/vocational training). Since educational degrees are considerably more structured, I systematically classify education degrees reported into 4 categories: Bachelor's degrees, Master's degrees, Doctoral degrees and other using a manual, exhaustive search. As degree information that is not sorted into the first 3 categories exhibit significant heterogeneity and noise, we only consider the effects of bachelor's, master's and doctoral degrees in our model.

Cumulative Career Controls

In addition to the above covariates that reflect static, Markovian transitions, I consider cumulative controls that chart the process of the career. Both internal and external mobility has been shown to be indicative of entrepreneurial aspirations or the lack thereof (e.g. Lazear, 2004, Åstebro and Thompson, 2011). I measure external and internal mobility by cumulatively counting the number of job title switches between and within company respectively. As the distribution of these two variables are highly skewed, I enter the square root transformation of these variables.

In addition, company characteristics of the current employment state has shown to be highly indicative of entrepreneurial spawning (Elfenbein et al., 2011). Particular to this setting is the finding that employees of previous venture backed companies are more likely to start venture-backed companies. I control for this by measuring the cumulative number of years a person has been employed at a venture-backed startup.

Finally, I enter fixed effects for the 23 educational institutions that characterizes the prospective sample.

Models and Design

Careers represent listings of social positions across time and age that forms event histories; as such I test my hypotheses through event history analysis. I model the likelihood of entrepreneurial entry using discrete linear probability hazard models through two broad designs. The first considers the entire dataset of the prospective sample. The second defines a robustness check through a matched comparison group with the treated alumni through Coarsened Exact Matching (CEM) (Iacus et al., 2011). This has the added benefit of allowing us to establish a treatment time clock to examine pre-post time trends. I will detail the model and designs in turn.

Linear Probability Hazard Model

Public resume data imposes no strict structure on time listings of the time of employment and education spells. While some individuals choose to be precise to the day level, a significant portion of the data approximates employment spells by year. As such, I consider entrepreneurial entry using a discrete time hazard rate:

$$P_{ikt} = Pr[T_i = t, K = k \mid T_i \geq t, x_{it}] \tag{12}$$

where P_{ikt} is the probability that individual i enters entrepreneurial state k at a particular age t (Allison, 1982).

I model how this hazard rate relates to our time and explanatory variables through a linear probability model (LPM). A benefit of the LPM over more traditional logistic regression models is the ease of interpretation. Coefficients can be read off and interpreted as additive advantages over the base rate (Angrist and Pischke, 2008). In addition, cross-derivatives of the LPM remains a constant, avoiding otherwise difficult interpretations of interaction results (Ai and Norton, 2003).

As these entrepreneurial and employment spells are listed as sequential events, and as we are interested in how prior career status results in entry, we specify:

$$P_{ikt} = \alpha_{t\tau} + \boldsymbol{\beta}^T \boldsymbol{x}_{it} \tag{13}$$

Where β represents the coefficients of covariates $\boldsymbol{x}_{it}$ to be estimated and $\alpha_{t\tau}$ is a constant specified by:

$$\alpha_{t\tau} = \alpha_0 + \alpha_1 t + \alpha_2 \tau \tag{14}$$

t is the age of the individual (where 0 is college graduation) and τ is the calendar year. These represents the two time-dependent variables: t is the hazard clock; τ is included to account for greater market conditions on entrepreneurial outcomes.

As no birth dates are reported in the data, age t is approximated by the year from graduation. This starts the age clock t at 0. As we are interested in founding events during employment, I standardize the age clock by considering only the post-graduation years, starting the age clock t at 0 for the year of graduation. Note that the myth of the "college dropout" is almost negligible in these schools, with the few college-dropouts listing their expected year of graduation and college attendance to claim affiliation.

The data is both left and right censored. Not everyone lists full education *and* employment histories in their resumes. These gaps between education and employment can result in unobserved left-censoring of employment spells. To prevent biases inherent in left-censored event history data, only individuals that exhibit full career histories (with a maximum "employment gap" of 3 years between college and employment) are considered (Klein and Moeschberger, 2005). In addition, I consider liquidity events that occurred on or before 2012 and cohorts that graduated on or before 2011. This reduces the sample to 354,866 individuals.
The first model examines basic differences between the career covariates of self-employment and venture founding: models 1 and 2 considers the base controls and hazard rate of the two entrepreneurial destination states, estimating equation (2).

I subsequently enter treatment dummies in models 3 and 4. Equation (2) becomes:

$$P_{ikt} = \alpha_{t\tau} + \boldsymbol{\beta}^T \boldsymbol{x}_{it} + \gamma D_{it} \tag{15}$$

where D_{it} is the treatment dummy indicator for individual i at time t and γ the treatment effect. These models include year, age (year from graduation) and college/university affiliation fixed effects. These models assess Hypotheses 1a and 1b broadly. I implement an Ordinary Least Squares (OLS) estimator; all standard errors are clustered at the individual level.

However, I note that the individuals who are at-risk of experiencing a liquidity event as alumni constitute a fundamentally different sample even among this population. Consider that the alumni of startups who go public are very likely to have worked with venture-backed companies. This particular group of people will be embedded in firms with a culture of spin-outs and are gain venture relevant human capital (e.g. Franco and Filson, 2006, Klepper and Sleeper, 2005).

To account for this, I construct a matched control group using Coarsened Exact Matching (CEM) (Iacus et al. 2011). As I am interested in considering pre-treatment trends, I matched on the careers of individuals at the time of exit from the focal company. That is, if company X experiences a liquidity event at year t, and an alumnus Y leaves the company at $t-5$, I consider an equivalent employee who have left a venture-funded startup firm that has yet to experience a liquidity event in my dataset at $t-5$. This equivalent employee Y' at $t-5$ will have worked in a venture-backed startup that is: (1) in the same industry as company X, categorized by CrunchBase (exact), (2) in the same age group (coarsened: binned), and (3) at the same stage of venture financing (by amount raised, coarsened). At the individual level, the control will be from (4) the same college cohort and major (graduation year of college and major cluster, exact), (5) hold the same job title role as Y (job title cluster membership, exact) and (6) have similar prior experience with venture backed companies (coarsened: binned). This creates the matched (with replacement) case-control sample. These matching criteria taxes the dataset with stringent sampling demands. In addition to two rare events (experiencing a liquidity event as an alumni and venture founding), I now impose the requirement of overlapping career support. From the full prospective sample of N individuals, a one-to-one matching narrows the case-control sample to 12,327 individuals.

With this sample, I implement a piece-wise linear probability hazard model, interacting the treatment with time pieces. The treatment clock starts at *zero* at the year of the IPO/acquisition and ticks negatively backwards and positively forward. I observe 6 time pieces: $\leq 7,\ (-7,2],\ (-2,0],\ (0,3],\ (3,7],\ >7$. As such, I implement:

$$P_{ikt} = \alpha_{t\tau} + \boldsymbol{\beta}^T \boldsymbol{x}_{it} + \gamma_1 D_i + \boldsymbol{\gamma}_2 \boldsymbol{\alpha}_{t_d} + \boldsymbol{\gamma}_{12} D_i \cdot \boldsymbol{\alpha}_{t_d} \tag{16}$$

where D_i is the treatment dummy indicator for individual i and $\boldsymbol{\alpha}_{t_d}$ the treatment-relative time piece. Models 5 and 6 estimates equation (16) with just year and age fixed effects. Models 7 and 8 includes career controls; this controls for any differences in post exit career movements between the treatment and control groups. Again, I implement an OLS estimator with clustered standard errors at the individual level.

Finally, I run Linear Probability Models conditioned on venture founding to assess the likelihood of being founded. Model 9 estimates the base effects of career covariates on venture founding on the prospective sample. Model 10 enters the treatment dummy. Model 11 considers the case-control sample.

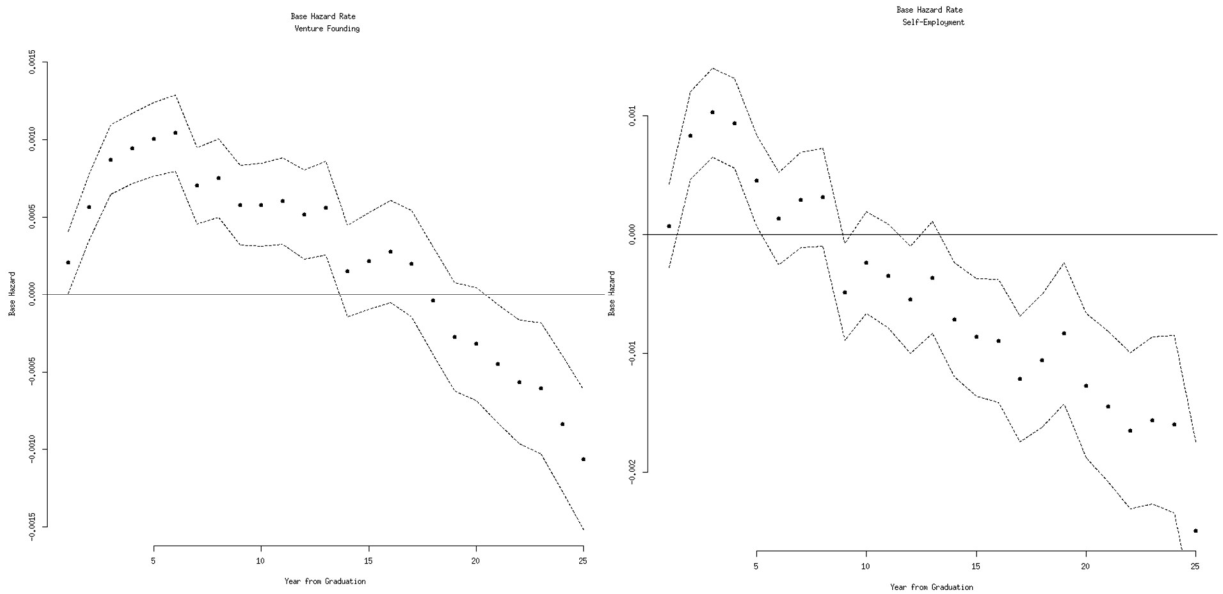


Figure 13. Base hazard rates by age (year from graduation), venture founding (left) and self-employment (right)

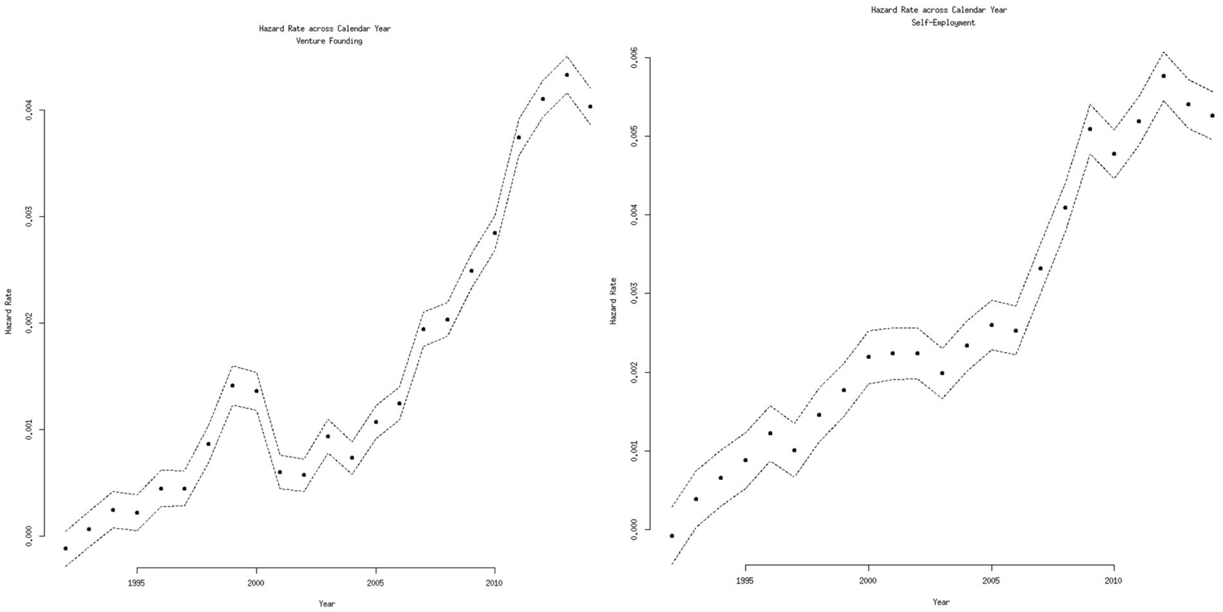


Figure 14. Hazard rates across calendar year, venture founding (left) and self-employment (right)

Results

Baseline Hazard and Descriptives

Table 1 shows the descriptive statistics for the prospective sample. Figures 13 and 14 depicts the baseline hazard rates across age (year from graduation) and calendar year.

Immediately we see differences between the venture founding and self-employment hazards for the prospective sample of elite STEM graduates. While the rate of venture founding peaks at about 6-10 years after graduation, the rate of self-employment sharply decreases after the first 3 years. Self-employment and venture founding hazard rates exhibit different responses to greater market conditions and employment trends. While venture founding directly reflect macro-economic conditions (observe the dot-com and recent technology start-up booms), we see a steady increase in rate of self-employment across the years. These results mirror the findings of the larger technology start-up ecology and are corroborated by evidence that examines the trends in contract labor and self-employment (Ruef, 2010; Kalleberg, 2000).

Table 4 depicts the results for models 1-4. Models 1 and 2 characterizes the fundamental career differences in venture founding and self-employment. The relative-to-base-rate effect sizes for most career covariates are large, suggesting that the role of the career in entrepreneurship in general is significant. The coefficients depict opportunity cost considerations and necessity in the consideration of entrepreneurial activity. The hierarchy variables show that the opportunity cost and human capital investments required for each entrepreneurial activity are starkly different. Individuals with significant higher education investments (Master's/Doctoral degrees) are much less likely to pursue self-employment while more likely to start ventures: the achievement of these post-graduation degrees increases the likelihood of venture funding by about 16% and 12% and reduces the likelihood of self-employment by 14% and 35% respectively.

Correspondingly, the occupancy at each rung of the job hierarchy affects each type of entrepreneurial entry differently. While there is a monotonic relationship of job rank with likelihood of venture founding (peaking at 80% more likely at the top 10% of the job hierarchy), the relationship of job hierarchy on self-employment for the prospective sample has less of a relationship, peaking at 12% more likely at the top of the job hierarchy. These results replicate extant findings that founders of high potential entrepreneurship are more likely to be drawn from the upper echelons of organizations and society.

The effect of career mobility on entrepreneurship is also replicated here. I find support for the claim that entrepreneurs in general are more likely to engage in job hopping: external mobility and internal mobility have large significant positive and negative effects respectively on both venture founding and self-employment. More saliently, I note that the quality of the job experience also matters. Experience at venture-backed companies are strongly predictive of venture founding (peaking at 42% more likely at 6-10 years) and subsequently lead to a lowered propensity for self-employment.

All in all, the career covariates support the base conception of venture funding as an aspirational career state as opposed to one entered as a consequent of blocked opportunity structure.

The Prospective Sample

The increase in status affiliations drives the alumni to venture founding: alumni of companies that experiences a liquidity event are 23% more likely to enter into venture founding and 17% less likely to enter into self-employment. These preliminary findings give support for both hypothesis 1a and 1b. However, I note that the requirements of being an alumnus of a liquid startup imposes subset restrictions on the data. As seen in Tables 4, the experience of working at a venture backed company is highly predictive of venture founding.

In addition, the requirement of having worked at, leave and then experience the liquidity event of a startup will significantly age the alumni of concern. This group of individuals will gain an increase in mobility and exposure to venture-backed start-ups. Tables 2 and 3 shows the cross-sectional descriptives of two subsets of the prospective sample: individuals with venture-

backed experience (Table 2) and the alumni who experienced (vicariously) the liquidity event (Table 3). Observe that there are fundamental differences, not only in the age and venture-backed experience, but also in the hierarchy and educational credentials of these three groups.
The preliminary results as presented in models 3 and 4 needs to be checked for common support. In addition, the entire prospective sample does not allow us to adjust for a treatment clock. To rectify this, I consider the case-control design as specified in Models 5-8.

Table 1: Descriptive Statistics, Prospective Sample

Prospective Sample (N = 354,866)	**min**	**max**	**range**	**median**	**mean**	**std.dev**
College Graduation Year	1980	2011	31	2003	2001.357	8.038
Age (years after college graduation)	4	35	31	12	13.643	8.038
External Mobility (count)	0	39	39	4	4.751	3.237
Internal Mobility (count)	0	17	17	0	0.633	1.120
Venture Backed Work Experience (years)	0	39	39	0	1.237	2.875
Liquidity Experience as Alumni	0	1	1	0	0.089	0.285
Highest Education Level	**Bachelor's**	**Master's**	**Doctoral**			
Proportion	0.479	0.388	0.132			
Highest Job Rank (Percentile)	**<10**	**10-25**	**25-50**	**50-75**	**75-90**	**90-100**
Proportion	0.039	0.118	0.239	0.278	0.182	0.144

Table 2: Descriptive Statistics, Venture Alumni

Venture Alumni Sample (N = 101,523)	**min**	**max**	**range**	**median**	**mean**	**std.dev**
College Graduation Year	1980	2011	31	2004	2002.354	7.528
Age (years after college graduation)	4	35	31	11	12.646	7.528
External Mobility (count)	0	37	37	5	5.363	5.363
Internal Mobility (count)	0	15	15	0	0.660	0.660
Venture Backed Work Experience (years)	1	39	38	3	4.323	3.942
Liquidity Experience as Alumni	0	1	1	0	0.150	0.357
Highest Education Level	**Bachelor's**	**Master's**	**Doctoral**			
Proportion	0.433	0.392	0.175			
Highest Job Rank (Percentile)	**<10**	**10-25**	**25-50**	**50-75**	**75-90**	**90-100**
Proportion	0.028	0.086	0.201	0.320	0.208	0.157

Table 3: Descriptive Statistics, Alumni of Organizations with Liquidity

Liquidity Alumni Sample (N = 31,714)	**min**	**max**	**range**	**median**	**mean**	**std.dev**
College Graduation Year	1980	2011	31	1997	1996.359	8.109
Age (years after college graduation)	4	35	31	18	18.641	8.109
External Mobility (count)	0	36	36	6	6.421	3.533
Internal Mobility (count)	0	15	15	0	0.910	1.386
Venture Backed Work Experience (years)	1	39	38	4	5.287	4.489
Liquidity Experience as Alumni	1	1	0	1	1.000	0.000
Highest Education Level	**Bachelor's**	**Master's**	**Doctoral**			
Proportion	0.497	0.419	0.085			
Highest Job Rank (Percentile)	**<10**	**10-25**	**25-50**	**50-75**	**75-90**	**90-100**
Proportion	0.019	0.068	0.180	0.270	0.235	0.228

Table 4: Prospective Sample

	Model 1	Model 2	Model 3	Model 4
Dependent Variable	**Venture Founding**	**Self-Employment**	**Venture Founding**	**Self-Employment**
Base Rate	2.69E-03	6.37E-03	2.69E-03	6.37E-03
External Mobility (sqrt)	2.35E-03 ***	5.44E-03 ***	2.34E-03 ***	5.47E-03 ***
	(3.83E-05)	(6.17E-05)	(3.83E-05)	(6.21E-05)
Internal Mobility (sqrt)	-4.77E-04 ***	-1.39E-03 ***	-4.92E-04 ***	-1.37E-03 ***
	(4.45E-05)	(6.51E-05)	(4.47E-05)	(6.52E-05)
Venture Experience: 1-2 years	7.94E-04 ***	-9.26E-04 ***	7.72E-04 ***	-8.91E-04 ***
	(1.22E-04)	(1.65E-04)	(1.22E-04)	(1.65E-04)
Venture Experience: 3-5 years	1.10E-03 ***	-5.09E-04 **	1.06E-03 ***	-4.52E-04 *
	(1.42E-04)	(1.85E-04)	(1.42E-04)	(1.85E-04)
Venutre Experience: 6-10 years	1.16E-03 ***	-2.22E-04	1.10E-03 ***	-1.37E-04
	(1.84E-04)	(2.37E-04)	(1.84E-04)	(2.37E-04)
Venture Experience: >11 years	4.59E-04 *	-1.05E-03 ***	4.06E-04 *	-9.58E-04 ***
	(1.84E-04)	(2.48E-04)	(1.84E-04)	(2.48E-04)
Masters Degree	4.36E-04 ***	-9.14E-04 ***	4.38E-04 ***	-9.17E-04 ***
	(6.05E-05)	(9.16E-05)	(6.05E-05)	(9.16E-05)
PhD Degree	3.44E-04 ***	-2.26E-03 ***	3.60E-04 ***	-2.29E-03 ***
	(1.02E-04)	(1.42E-04)	(1.02E-04)	(1.42E-04)
Not in Employment	-3.47E-04 ***	-9.23E-04 ***	-3.47E-04 ***	-9.24E-04 ***
	(9.79E-05)	(1.44E-04)	(9.79E-05)	(1.44E-04)
Title Hierarchy: 0-10% quantile	-7.47E-04 ***	-3.05E-04	-7.42E-04 ***	-3.12E-04
	(1.19E-04)	(2.49E-04)	(1.19E-04)	(2.49E-04)
Title Hierarchy:10-25% quantile	-1.05E-03 ***	-9.86E-04 ***	-1.05E-03 ***	-9.92E-04 ***
	(8.14E-05)	(1.58E-04)	(8.15E-05)	(1.58E-04)
Title Hierarchy: 25-50% quantile	-3.09E-04 ***	-3.03E-04 *	-3.06E-04 ***	-3.09E-04 *
	(7.93E-05)	(1.33E-04)	(7.94E-05)	(1.33E-04)
Title Hierarchy: 50-75% quantile	-2.61E-04 **	-6.02E-04 ***	-2.60E-04 **	-6.05E-04 ***
	(8.59E-05)	(1.39E-04)	(8.59E-05)	(1.39E-04)
Title Hierarchy: 75-90% quantile	8.62E-04 ***	4.60E-04 *	8.58E-04 ***	4.67E-04 **
	(1.32E-04)	(1.79E-04)	(1.32E-04)	(1.79E-04)
Title Hierarchy: 90-100% quantile	2.19E-03 ***	7.81E-04 **	2.18E-03 ***	8.03E-04 **
	(2.16E-04)	(2.57E-04)	(2.16E-04)	(2.57E-04)
Treatment: Liquidity			6.30E-04 ***	-1.07E-03 ***
			(1.51E-04)	(1.98E-04)
Fixed Effects:				
Calendar Year	✓	✓	✓	✓
Age	✓	✓	✓	✓
Education Major	✓	✓	✓	✓
Job Title	✓	✓	✓	✓
Education Institution	✓	✓	✓	✓
n	5227797	5227797	5227797	5227797
N	354666	354666	354666	354666
df	5132230	4998050	5132229	4998049
F	116.9***	162***	116.4***	161***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Case-Control Sample

Table 5 depicts the results for the case control sample. Models 5 and 6 considers just the main treatment effect interacted with time pieces; models 7 and 8 includes career covariates.

Without controls, models 5 and 6 finds a positive effect of treatment in the case-control sample on venture founding during the years surrounding the liquidity event as well as in the later years of the career. I find no significant effect of the treatment on self-employment. Although there is an absence of a pre-trend (likelihoods of venture founding and self-employment of the two groups not different before liquidity event), there is a fundamental difference in founding rates between the treatment and control groups. This suggests that there is considerable career divergence of the alumni and their matched "twin" after their respective startup exits. Model 7 and 8 controls for the divergence of careers by entering career covariates.

In Model 7, I observe that while treatment effect differences in the later stages of time disappear for venture founding, the effect remains localized around the year of the liquidity event. In addition, two other findings support this matching strategy. First, we see no pattern in the pre-trend: there are no differences in the likelihood of venture-founding between the treatment and control groups before the liquidity event. Second, we find no significant differences in the likelihood of founding between treatment and control groups. When examining time trends, the effect when localized to the years surrounding the liquidity event, is considerably larger: alumni of high status startups are almost 80% more likely to pursue venture founding during the years surrounding the liquidity event.

Model 8 shows that the propensity for self-employment in the later years is reduced in the treatment group. Again I observe no significant pre-trend and inter-group differences in the likelihood of self-employment. In sum, these findings again support Hypotheses 1a and 1b.

Table 5: Case-Control Sample

	Model 5	Model 6	Model 7	Model 8
Dependent Variable	**Venture Founding**	**Self-Employment**	**Venture Founding**	**Self-Employment**
Base Rate	2.21E-03	4.77E-03	2.21E-03	4.77E-03
Treatment: Liquidity	4.77E-04 **	6.71E-04 .	1.61E-04	1.95E-04
	(1.66E-04)	(3.96E-04)	(1.69E-04)	(3.99E-04)
Time Piece -7 to -2 years	4.32E-04	2.81E-04	4.17E-04	1.39E-05
	(3.74E-04)	(5.60E-04)	(3.80E-04)	(5.63E-04)
Time Piece -1 to 0 years	-3.86E-04	2.02E-03 *	-4.55E-04	1.51E-03
	(4.53E-04)	(9.52E-04)	(4.51E-04)	(9.46E-04)
Time Piece 1 to 3 years	9.53E-04 .	1.97E-03 *	7.84E-04	1.26E-03
	(5.26E-04)	(9.81E-04)	(5.26E-04)	(9.62E-04)
Time Piece 4 to 7 years	1.46E-03 *	1.29E-03	1.13E-03 .	4.50E-04
	(5.89E-04)	(8.56E-04)	(5.87E-04)	(8.53E-04)
Time Piece > 7 years	2.74E-04	2.53E-03 *	-6.08E-05	1.57E-03
	(5.04E-04)	(1.17E-03)	(5.20E-04)	(1.15E-03)
Treatment × Time Piece -7 to -2 years	1.25E-04	1.04E-03	-2.32E-04	4.26E-04
	(5.62E-04)	(7.24E-04)	(5.62E-04)	(7.27E-04)
Treatment × Time Piece -1 to 0 years	2.38E-03 ***	8.13E-04	1.72E-03 *	-2.24E-04
	(7.01E-04)	(1.16E-03)	(7.01E-04)	(1.17E-03)
Treatment × Time Piece 1 to 3 years	5.65E-04 ***	3.42E-04	-1.84E-04	-7.62E-04
	(6.61E-04)	(9.95E-04)	(6.66E-04)	(9.91E-04)
Treatment × Time Piece 4 to 7 years	8.45E-04	4.75E-04	4.82E-05	-7.78E-04
	(6.93E-04)	(8.48E-04)	(7.00E-04)	(8.59E-04)
Treatment × Time Piece > 7 years	1.74E-03 **	-6.30E-04	8.78E-04	-2.25E-03 *
	(5.89E-04)	(1.06E-03)	(5.98E-04)	(1.08E-03)
External Mobility (sqrt)			2.04E-03 ***	5.07E-03 ***
			(1.82E-04)	(2.80E-04)
Internal Mobility (sqrt)			-3.28E-05	-6.11E-04 *
			(2.03E-04)	(2.48E-04)
Venture Experience: 1-2 years			1.49E-03 *	1.51E-04
			(6.16E-04)	(7.43E-04)
Venture Experience: 3-5 years			1.79E-03 *	8.47E-04
			(7.35E-04)	(8.65E-04)
Venutre Experience: 6-10 years			1.97E-03 *	-1.18E-04
			(9.60E-04)	(1.00E-03)
Venture Experience: >11 years			4.63E-04	1.21E-04
			(1.24E-03)	(1.45E-03)
Masters Degree			3.81E-04	-4.44E-04
			(2.56E-04)	(3.46E-04)
PhD Degree			1.88E-05	-1.24E-03 .
			(4.98E-04)	(7.07E-04)
Not in Employment			-3.37E-04	-1.02E-03
			(4.44E-04)	(6.44E-04)
Title Hierarchy: 0-10% quantile			-8.22E-04 *	-2.08E-03 *
			(3.71E-04)	(9.93E-04)
Title Hierarchy:10-25% quantile			-8.44E-04 **	-1.15E-03 .
			(3.18E-04)	(6.70E-04)
Title Hierarchy: 25-50% quantile			-6.01E-04 *	-9.89E-04 *
			(2.95E-04)	(5.02E-04)
Title Hierarchy: 50-75% quantile			1.62E-04	-4.03E-04
			(3.42E-04)	(5.28E-04)
Title Hierarchy: 75-90% quantile			1.00E-03 .	4.49E-04
			(5.49E-04)	(6.41E-04)
Title Hierarchy: 90-100% quantile			2.05E-03 *	1.51E-03
			(8.50E-04)	(9.42E-04)
Fixed Effects:				
Calendar Year	✓	✓	✓	✓
Age	✓	✓	✓	✓
Job Title			✓	✓
n	273010	273010	273010	273010
N	12327	12327	12327	12327
df	272928	272928	272860	272860
F	6.727***	3.821***	8.57***	8.093***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 6: Venture Funding Rates

	Model 9	Model 10	Model 11
Sample	**Prospective**	**Prospective**	**Case-Control**
Dependent Variable	**Venture Funding**	**Venture Funding**	**Venture Funding**
Base Rate	0.223	0.223	0.222
External Mobility (sqrt)	-6.33E-04	-1.73E-04	1.21E-02
	(6.94E-03)	(6.98E-03)	(4.01E-02)
Internal Mobility (sqrt)	-1.26E-02 .	-1.24E-02	3.81E-03
	(7.43E-03)	(7.44E-03)	(3.38E-02)
Venture Experience: 1-2 years	6.98E-02 ***	7.01E-02 ***	7.35E-02
	(1.19E-02)	(1.19E-02)	(6.05E-02)
Venture Experience: 3-5 years	9.37E-02 ***	9.44E-02 ***	1.54E-01 *
	(1.44E-02)	(1.45E-02)	(6.76E-02)
Venutre Experience: 6-10 years	7.98E-02 ***	8.06E-02 ***	7.20E-02
	(2.12E-02)	(2.12E-02)	(8.73E-02)
Venture Experience: >11 years	1.66E-01 ***	1.67E-01 ***	1.05E-01
	(4.45E-02)	(4.46E-02)	(2.02E-01)
Masters Degree	3.28E-02 ***	3.27E-02 ***	9.28E-02 *
	(8.40E-03)	(8.40E-03)	(4.10E-02)
PhD Degree	1.42E-01 ***	1.42E-01 ***	3.11E-01 **
	(1.54E-02)	(1.54E-02)	(1.16E-01)
Not in Employment	-2.41E-02 .	-2.41E-02 .	-7.84E-02
	(1.42E-02)	(1.42E-02)	(6.53E-02)
Title Hierarchy: 0-10% quantile	-2.37E-02	-2.38E-02	2.92E-01
	(2.97E-02)	(2.97E-02)	(2.13E-01)
Title Hierarchy:10-25% quantile	-3.41E-02 .	-3.42E-02 .	-2.75E-01 *
	(1.98E-02)	(1.98E-02)	(1.26E-01)
Title Hierarchy: 25-50% quantile	1.86E-04	1.43E-04	-1.17E-01
	(1.44E-02)	(1.44E-02)	(8.16E-02)
Title Hierarchy: 50-75% quantile	-1.18E-02	-1.18E-02	-6.00E-02
	(1.42E-02)	(1.42E-02)	(6.41E-02)
Title Hierarchy: 75-90% quantile	2.77E-02 .	2.78E-02 *	-9.61E-02
	(1.41E-02)	(1.41E-02)	(5.60E-02)
Title Hierarchy: 90-100% quantile	2.89E-02 .	2.92E-02 .	-9.95E-02
	(1.55E-02)	(1.55E-02)	(5.60E-02)
Treatment: Liquidity		-9.33E-03	7.10E-02
		(1.51E-02)	(2.55E-01)
Time Piece -7 to -2 years			2.89E-02
			(2.09E-01)
Time Piece -1 to 0 years			6.27E-02
			(2.28E-01)
Time Piece 1 to 3 years			1.24E-01
			(2.13E-01)
Time Piece 4 to 7 years			1.21E-01
			(2.17E-01)
Time Piece > 7 years			-1.95E-02
			(2.23E-01)
Treatment × Time Piece -7 to -2 years			-3.82E-02
			(2.65E-01)
Treatment × Time Piece -1 to 0 years			-8.85E-02
			(2.84E-01)
Treatment × Time Piece 1 to 3 years			-1.09E-01
			(2.68E-01)
Treatment × Time Piece 4 to 7 years			-8.51E-02
			(2.66E-01)
Treatment × Time Piece > 7 years			3.54E-02
			(2.70E-01)
Fixed Effects:			
Calendar Year	✓	✓	✓
Age	✓	✓	✓
Education Major	✓	✓	
Job Title	✓	✓	✓
Education Institution	✓	✓	
n	13740	13740	604
df	13650	13651	519
F	5.84***	5.8***	1.648*

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Funding Likelihoods

The results of Model 9-11 are shown in Table 6. Contingent on venture-founding, Model 9 examines the base career covariates on the likelihood of venture funding on all venture founding attempts in the prospective sample. In particular, differences in effect sizes of Model 9 and Model 1 highlights differences in the career antecedents of receiving venture capital and venture entry. For this sample, the base rate of receiving venture funding (defined as either a Series A or more than 100k in Angel Investments) is 0.223. Given this, the two largest predictors of venture funding are: (1) experience in a venture backed company (peaking at a 77% advantage for >11 years of venture backed experience) and (2) obtaining a doctoral degree (a 65% advantage). Investors evidently also gravitate towards founders who are drawn from the higher ranks of organizations (36% more likely to fund top 10% of job titles), but are relatively agnostic towards diversity of experiences (low effect sizes for both internal and external mobility).

Model 10 enters the treatment dummy variable; this shows no effect on the status gain. This result is replicated in the case-control sample in Model 11. These findings reject Hypothesis 2a and lends support Hypothesis 2b: while the status gain spurs individuals to pursue venture founding, venture investors are agnostic to such gains in affiliations.

Discussion

The study here presents a novel research design in effort to identify the effect of organizational affiliation to a high-status start-up on entrepreneurial entry. This study examines the effect of liquidity events on the response of both the would-be entrepreneur and venture investors. To identify the effect of status gain that is distinct from resource confounds, I consider the response of start-up alumni who vicariously experiences the liquidity event. In effort to control for selection and human capital differences, I implement career matching and control for potential career divergence. The findings suggest that while positive status events in one's career is likely to push an individual into venture founding in the years immediately preceding and following the liquidity event, and deters the individual from self-employment in the long run.

However, the results suggest that venture investor reception to ventures born out of this gain is status is lukewarm at best. These results suggest that prior findings of advantages accorded to prominence and status of organizational affiliations in the arena of high potential entrepreneurship might be overstated, not just to researchers but also, perhaps dangerously, to potential entrepreneurs. This flies in the face of numerous studies that suggests otherwise (e.g. Stuart and Cable, 2002; Burton et al., 2002).

One distinction of this study the claim that there is little or no difference in the underlying quality of the entrepreneur immediately before the liquidity event. Prior examinations of social capital and entrepreneurial prominence focuses primarily the functional aspect of social capital. A high-status social affiliation can be activated for knowledge and resource gains; this is again, the pipes and prisms conception of a social network (Podolny, 2001). High status organizational affiliations have functional roles that provide informational advantages and resource flow. Consider that the social capital accorded by being situated in a prominent organization will be beneficial in the assessment and identification of opportunities. Burt (1992) highlights the role of structural holes: gaps in an egocentric network primed for exploitation in pursuit of information and resource gain. Strategic social positioning thus represents an advantage in not just information availability but also the manipulation and activation of social resources; these social positions are in turn mirrored by organizational affiliations.

In the case of liquidity shocks, consider that the alumnus finds herself not just gaining the bragging rights of having worked at a prominent and successful start-up, but also surrounded with ex-colleagues that are now both wealthy and influential. In network analysis jargon, the liquidity event increases the eigenvector centrality of the alumni while remaining orthogonal to their degree centrality. A valuable extension is to examine heterogeneous responses of venture founders that respond to such a status shock. Venture founders could be merely emboldened by the bragging rights. In this case, the lowering of the founding threshold is purely perceptual and unwarranted. Conversely, venture founders can react to the status shock by activating their now superior networks, involving their ex-colleagues in the venture founding process. This functional response to the increase in status affiliations will likely increase the capabilities of the founding team, increasing the quality of the venture and likely accrue funding advantages.

CHAPTER 4
CONCLUSION AND FUTURE DIRECTIONS

I have introduced the career framework of entrepreneurship and demonstrated its application to the illumination of entrepreneurial antecedents. This framework addressed the inherent issues in the question of entrepreneurial antecedents (Chapter 1). In Chapter 2, I elaborated and developed the framework in more detail, ascribing an urgency to the questions detailed in the former chapter, as well as demonstrating how the career framework provides a parsimonious path forward. Chapter 2 also detailed the data collected and empirically described the landscape of the high technology start-up labor ecology in the United States. With the theoretical, methodological and data set up as laid out, I then demonstrated the potency of this conceptual framework and the empirical possibilities in Chapter 3. Here, I demonstrated how positive career events might affect career trajectories. Using several matching strategies to attempt causal inference, I revisited the question of the effect of status on entrepreneurial propensities, and I demonstrated that the data at hand and the career frame work in tandem allows for a deeper more nuanced probing of such topics.

In addition, I have described and introduced a novel new dataset: public resume collections of particular human ecologies. The conceptual contributions and methodologies as described here merely form the beginnings of a research program; these chapters suggest a clear pipeline of research ahead. The potential branches and foliage of research directions are many, I shall only give a couple as examples here.

1) Resume data proves to be extremely rich. The unstructured nature of the text allows for deeper mining and information retrieval. For instance, there is information inherent in people's names: from people's first and last names we can statistically infer their ethnicity and gender. The latter has been a hot topic of discussion. With the defining of good risk-sets, I can establish for certain the gender gap in high-technology entrepreneurship and also disentangle career pipelines and blockages that might cause such a gap. Here, the distinction between self-employment and venture founding becomes especially important when I consider that the two groups occupy the two ends of the wage distribution and therefore possess different fates and fortunes.

2) While the dissertation focuses on the background of entrepreneurs with respect to the first founding attempt, I am interested in what happens to entrepreneurs post founding event. The data as presented observes not only the various outcomes of entrepreneurial founding but also entrepreneurial experiences (e.g. funding events, growth, and greater market fluctuations). This allows me to examine the heterogeneity of post-entrepreneurship careers as a consequence of entrepreneurial experiences.

3) The large scale of the data as presented here suggests that predictive modeling is possible. Although not documented here, I have successfully pursued this line of thought elsewhere.[6] However, predictive models not only demonstrate industry applications and explanatory power, they can provide means for identification. For instance, Goel et al. (2016) used the large amounts of data generated by New York's stop-and-frisk policy to generate ex-ante probabilities of a suspect carrying a weapon (a rare event) through saturated machine learning models (~ 8000 features across over 300,000 data points). Discrimination is then

[6] See: Miller, C. K. (2015, July 2). The Next Mark Zuckerberg Is Not Who You Might Think. *New York Times*. Retrieved from https://www.nytimes.com/2015/07/02/upshot/the-next-mark-zuckerberg-is-not-who-you-might-think.html

identified by considering the heterogeneity of ex-ante probability distributions between racial groups and neighborhoods. I am interested in extending such big data identification techniques to test hypotheses related to rare entrepreneurship events.

4) Finally, beyond entrepreneurship, the data at hand can illuminate other enterprise efforts such as innovation and invention. Consider that the extant data can be merged with patenting information. Disambiguating through time, name and assignee, merging resume data with patenting information allows us to properly measure inventor mobility. Current measures of inventor mobility as assessed purely through patent information are insufficient: findings show that when compared with public resumes, inventor mobility assessed through patents have error rates of 83% false negatives and 12% false positives (Ge et al. 2016). Using the data at hand, I can not only clarify mobility effects on innovation, but also more directly assess the impact of careers on innovation outcomes, and innovation on entrepreneurship.

REFERENCES

Abbott, A., & Hrycak, A. (1990). Measuring resemblance in sequence data: An optimal matching analysis of musicians' careers. *American Journal of Sociology*, 144-185.

Ai, C., & Norton, E. C. (2003). Interaction terms in logit and probit models. *Economics Letters*, 80(1), 123-129.

Aldrich Howard, E., & Ruef, M. (2006). *Organizations Evolving*. Sage Publications, London.

Aldrich, H. E., & Waldinger, R. (1990). Ethnicity and entrepreneurship. *Annual Review of Sociology*, 111-135.

Allison, P. D. (1982). Discrete-time methods for the analysis of event histories. *Sociological Methodology*, 13(1), 61-98.

Angrist, J. D., & Pischke, J. S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press. Princeton, NJ.

Arthur, M. B., & Rousseau, D. M. (1996). *The Boundaryless Career*. Oxford University Press.

Åstebro, T., & Thompson, P. (2011). Entrepreneurs, Jacks of all trades or Hobos?. *Research Policy*, *40*(5), 637-649.

Azoulay, P., Stuart, T., & Wang, Y. (2013). Matthew: Effect or fable?. *Management Science*, *60*(1), 92-109.

Barnett, W. P., Baron, J. N., & Stuart, T. E. (2000). Avenues of Attainment: Occupational Demography and Organizational Careers in the California Civil Service. *American Journal of Sociology*, *106*(1), 88-144.

Barley, S. R. (1989). Careers, identities, and institutions: the legacy of the Chicago School of Sociology, in Arthur, M. B., Hall, D. T., & Lawrence, B. S. (Eds.), *Handbook of Career Theory* (pp 41-65). Cambridge University Press, Cambridge.

Baron, J. N., & Bielby, W. T. (1980). Bringing the firms back in: Stratification, segmentation, and the organization of work. *American Sociological Review*, 737-765.

Baron, J. N., Hannan, M. T., & Burton, M. D. (1999). Building the iron cage: Determinants of managerial intensity in the early years of organizations. *American Sociological Review*, 527-547.

Borjas, G. J., & Bronars, S. G. (1989). Consumer Discrimination and Self-Employment. *Journal of Political Economy*, *97*(3), 581-605.

Bidwell, M., & Briscoe, F. (2010). The dynamics of interorganizational careers. *Organization Science*, *21*(5), 1034-1053.

Burt, R. (1992). The social structure of competition. N. Nohria, R. Eccles, eds. *Networks and Organizations: Structure, Form, and Action*, 57-91.

Burton, M. D., Sørensen, J. B., & Beckman, C. M. (2002). Coming from good stock: Career histories and new venture formation. *Research in the Sociology of Organizations*, *19*(1), 229-262.

Burton, M. D., Sørensen, J. B., & Dobrev, S. D. (2016). A Careers Perspective on Entrepreneurship. *Entrepreneurship Theory and Practice*, *40*(2), 237-247.

Carroll, G. R., & Khessina, O. M. (2005). The ecology of entrepreneurship. In *Handbook of Entrepreneurship Research* (pp. 167-200). Springer US.

Carroll, G. R., & Mosakowski, E. (1987). The career dynamics of self-employment. *Administrative Science Quarterly*, 570-589.

Cassar, G. (2006). Entrepreneur opportunity costs and intended venture growth. *Journal of Business Venturing*, *21*(5), 610-632.

Dahl, M. S., & Sorenson, O. (2012). Home sweet home: Entrepreneurs' location choices and the performance of their ventures. *Management Science*, *58*(6), 1059-1071.

Dokko, G., & Rosenkopf, L. (2010). Social capital for hire? Mobility of technical professionals and firm influence in wireless standards committees. *Organization Science*, *21*(3), 677-695.

Elfenbein, D. W., Hamilton, B. H., & Zenger, T. R. (2010). The small firm effect and the entrepreneurial spawning of scientists and engineers. *Management Science*, *56*(4), 659-681.

Elo, A. E. (1978). *The Rating of Chessplayers, Past and Present*. Arco Pub.

Ferber, M., & Waldfogel, J. (1998). The long-term consequences of non-standard work. *Monthly Labor Review*, *121*(5), 3-12.

Franco, A. M., & Filson, D. (2006). Spin-outs: knowledge diffusion through employee mobility. *The RAND Journal of Economics*, 37(4), 841-860.

Freeman, J. (1986). Entrepreneurs as organizational products: Semiconductor firms and venture capital firms. G. Libecap, ed. *Advances in the Study of Entrepreneurship, Innovation, and Economic Growth*, Vol. 1. JAI Press, Greenwich, CT, 33–58.

Freeman, J., Carroll, G. R., & Hannan, M. T. (1983). The liability of newness: Age dependence in organizational death rates. *American Sociological Review*, 692-710.

Ge, C., Huang, K. W., & Png, I. P. (2016). Engineer/scientist careers: Patents, online profiles, and misclassification bias. *Strategic Management Journal*, *37*(1), 232-253.

Goel, S., Rao, J. M., & Shroff, R. (2016). Precinct or Prejudice? Understanding Racial Disparities in New York City's Stop-and-Frisk Policy. *Annals of Applied Statistics*, *10*(1), 365-394.

Goffman, E. (1959). The moral career of the mental patient. *Psychiatry*, *22*(2), 123-142.

Hand, J. R. (2008). Give everyone a prize? Employee stock options in private venture-backed firms. *Journal of Business Venturing*, *23*(4), 385-404.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Unsupervised learning. In *The Elements of Statistical Learning* (pp. 485-585). Springer New York.

Haveman, H. A., & Cohen, L. E. (1994). The ecological dynamics of careers: The impact of organizational founding, dissolution, and merger on job mobility. *American Journal of Sociology*, 104-152.

Hayward, M. L., Shepherd, D. A., & Griffin, D. (2006). A hubris theory of entrepreneurship. *Management Science*, 52(2), 160-172.

Helleputte, T. (2015). LiblineaR: Linear Predictive Models Based On the Liblinear C/C++ Library. R package version 1.94-2.

Hornik K., Mair P., Rauch J., Geiger W., Buchta C. and Feinerer I. (2013). The textcat Package for *n*-Gram Based Text Categorization in R. *Journal of Statistical Software, 52*(6), 1-17.

Hsu, D. H. (2004). What do entrepreneurs pay for venture capital affiliation?. *The Journal of Finance*, *59*(4), 1805-1844.

Hsu, D. H., Roberts, E. B., & Eesley, C. E. (2007). Entrepreneurs from technology-based universities: Evidence from MIT. *Research Policy*, *36*(5), 768-788.

Hughes, E. C. (1958). *Men and their Work*. Free Press.

Iacus, S. M., King, G., & Porro, G. (2011). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, mpr013.

Kalleberg, A. L. (2000). Nonstandard employment relations: Part-time, temporary and contract work. *Annual Review of Sociology*, 341-365.

Klepper, S., & Sleeper, S. (2005). Entry by spinoffs. *Management Science*, *51*(8), 1291-1306.

Lazear, E. P. (2004). Balanced skills and entrepreneurship. *The American Economic Review*, *94*(2), 208-211.

Levine, R., & Rubinstein, Y. (2013). *Smart and Illicit: Who Becomes an Entrepreneur and Do They Earn More?* (No. w19276). National Bureau of Economic Research.

Lounsbury, M., & Glynn, M. A. (2001). Cultural entrepreneurship: Stories, legitimacy, and the acquisition of resources. *Strategic Management Journal*, *22*(6-7), 545-564.

Mead, G. H. (1934). *Mind, Self, and Society*. Chicago: University of Chicago Press.

Nagarajan, M. and M.A. Hearst 2009. "An Examination of Language Use in Online Dating Profiles." *3rd International AAAI Conference on Weblogs and Social Media* (ICWSM 2009). San Jose, CA.

Navis, C., & Glynn, M. A. (2010). How new market categories emerge: Temporal dynamics of legitimacy, identity, and entrepreneurship in satellite radio, 1990–2005. *Administrative Science Quarterly*, *55*(3), 439-471.

Newman, M. E. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, *46*(5), 323-351.

Patterson, O. (2014). Making sense of culture. *Annual Review of Sociology*, *40*, 1-30.

Petersen, T., & Saporta, I. (2004). The Opportunity Structure for Discrimination1. *American Journal of Sociology*, *109*(4), 852-901.

Phillips, D. J. (2002). A genealogical approach to organizational life chances: The parent-progeny transfer among Silicon Valley law firms, 1946–1996. *Administrative Science Quarterly*, *47*(3), 474-506.

Podolny, J. M. (1994). Market uncertainty and the social character of economic exchange. *Administrative Science Quarterly*, 458-483.

Podolny, J. M. (2001). Networks as the Pipes and Prisms of the Market1. *American Journal of Sociology*, *107*(1), 33-60.

Podolny, J. M. (2005). *Status Signals: A Sociological Theory of Market Competition.* Princeton University Press.

Portes, A., & Jensen, L. (1989). The enclave and the entrants: Patterns of ethnic enterprise in Miami before and after Mariel. *American Sociological Review*, 929-949.

Rao, H. (1994). The social construction of reputation: Certification contests, legitimation, and the survival of organizations in the American automobile industry: 1895–1912. *Strategic Management Journal*, *15*(S1), 29-44.

Rao, H. (1998). Caveat emptor: The construction of nonprofit consumer watchdog organizations. *American Journal of Sociology*, *103*(4), 912-961.

Rider, C. I., & Tan, D. (2014). Labor market advantages of organizational status: A study of lateral partner hiring by large US law firms. *Organization Science*, *26*(2), 356-372.

Roberts, P. W., & Sterling, A. D. (2012). Network progeny? Prefounding social ties and the success of new entrants. *Management Science*, *58*(7), 1292-1304.

Rock, P. (1979). *The Making of Symbolic Interactionism*. London: Macmillan.

Ruef, M. (2010). *The Entrepreneurial Group: Social Identities, Relations, and Collective Action*. Princeton University Press.

Ruef, M., Aldrich, H. E., & Carter, N. M. (2003). The structure of founding teams: Homophily, strong ties, and isolation among US entrepreneurs. *American Sociological Review*, 195-222.

Shane, S. A. (2008). *The Illusions of Entrepreneurship: The Costly Myths That Entrepreneurs, Investors, And Policy Makers Live by*. Yale University Press.

Shane, S., & Cable, D. (2002). Network ties, reputation, and the financing of new ventures. *Management Science*, *48*(3), 364-381.

Shane, S., & Stuart, T. (2002). Organizational endowments and the performance of university start-ups. *Management Science*, *48*(1), 154-170.

Sørensen, J. B., & Fassiotto, M. A. (2011). Organizations as fonts of entrepreneurship. *Organization Science*, *22*(5), 1322-1331.

Sørensen, J. B., & Sharkey, A. J. (2014). Entrepreneurship as a mobility process. *American Sociological Review*, *79*(2), 328-349.

Sorenson, O., & Stuart, T. E. (2001). Syndication networks and the spatial distribution of venture capital investments. *American Journal of Sociology*, *106*(6), 1546-1588.

Sorenson, O., & Stuart, T. E. (2008). 12 Entrepreneurship: a field of dreams?. *The Academy of Management Annals*, *2*(1), 517-543.

Spilerman, S. (1977). Careers, labor market structure, and socioeconomic achievement. *American Journal of Sociology*, 551-593.

Stephenson, A., & Sonas, J. (2014). PlayerRatings: Dynamic Updating Methods for Player Ratings Estimation. R package v1. 0.

Stinchcombe, A. L. (1965). Organizations and social structure. *Handbook of Organizations*, *44*(2), 142-193. Rand McNally: Chicago, IL.

Stryker, S. (1980). *Symbolic Interactionism: A Social Structural Version*. Benjamin-Cummings Publishing Company.

Stuart, T. E., & Ding, W. W. (2006). When do scientists become entrepreneurs? The social structural antecedents of commercial activity in the academic life sciences1. *American Journal of Sociology*, *112*(1), 97-144.

Stuart, T. E., Hoang, H., & Hybels, R. C. (1999). Interorganizational endorsements and the performance of entrepreneurial ventures. *Administrative Science Quarterly*, *44*(2), 315-349.

Stuart, T. E., & Podolny, J. M. (1996). Local search and the evolution of technological capabilities. *Strategic Management Journal*, *17*(S1), 21-38.

Stuart, T. E., & Sorenson, O. (2003). Liquidity events and the geographic distribution of entrepreneurial activity. *Administrative Science Quarterly*, *48*(2), 175-201.

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

Swidler, A. (1986). Culture in action: Symbols and strategies. *American Sociological Review*, 273-286.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*(301), 236-244.

Weber, K., Heinze, K. L., & DeSoucey, M. (2008). Forage for thought: Mobilizing codes in the movement for grass-fed meat and dairy products. *Administrative Science Quarterly*, *53*(3), 529-567.

Zhao, P., & Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research*, *7*(Nov), 2541-2563.

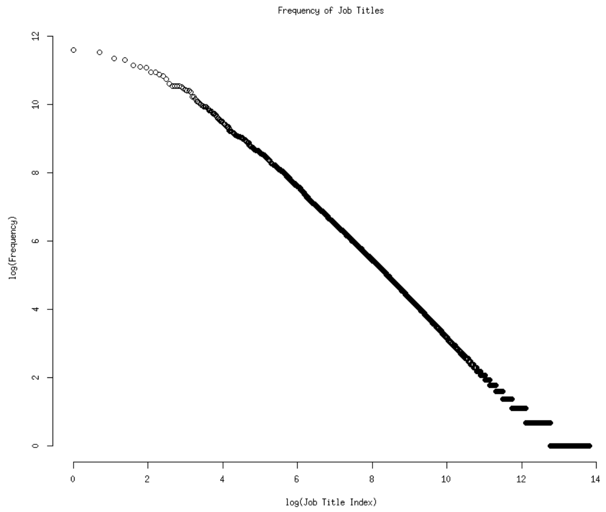
Zider, B. (1998). How venture capital works. *Harvard Business Review*, *76*(6), 131-139.

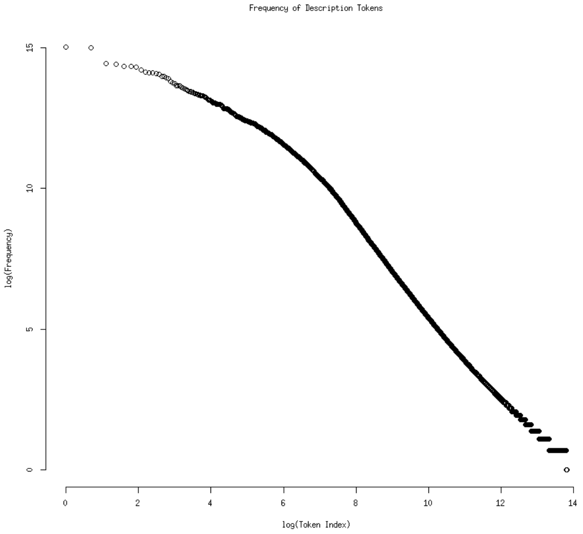
Zuckerman, E. W. (1999). The categorical imperative: Securities analysts and the illegitimacy discount. *American Journal of Sociology*, *104*(5), 1398-1438.

APPENDIX A.
JOB TITLE PCA AND CLUSTERING

Distribution of Job Titles and Descriptions

A key note is that the frequency of words (titles and descriptions) are extremely skewed and resemble power law distributions. The figures below show the log-log frequency distributions of the top 1 million job titles (left) and job description tokens (right). Observe that they obey pseudo-power law distributions. After stop-word removal and stemming of the job descriptions, we are left with a total of 222,623,815 processed monogram tokens across 2,155,556 unique monogram tokens.



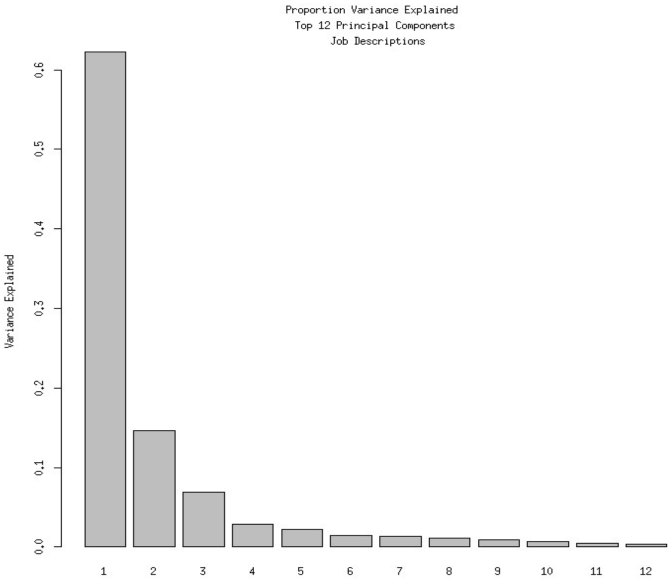


The top 16,679 most frequent monogram tokens are selected, which represent 94.68% of the total number. The distribution of frequency of these tokens are plotted as below (ordered from most to least frequent). These features are indexed and the term-document matrix is created.

Principal Component Analysis:

The main intuition of the learning algorithm here is that similar job titles will be described in the same way and as such, have correlated text vectors. The term-document matrix is aggregated at the level of each title and normalized. To reduce noise, titles that occur more than 30 times in the data are used. This comprise the first 16,679 job titles.

The PCA of the job titles is calculated using the `princomp` command in R. The top 12 principal components represent 0.953399 of the total variation in the text descriptions. The following graph and table shows the variances of the top 12 PC, and also the highest and lowest 20 weighted title tokens for the first 4 principal components (as an example).



PC	1		2		3		4	
Rank	**Title.tokens**	**Weight**	**Title.tokens**	**Weight**	**Title.tokens**	**Weight**	**Title.tokens**	**Weight**
1	business development intern	-1.39E-02	and intern marketing sales	-0.01342	manager service	-0.01573	sales specialist	-0.01838
2	management trainee	-1.37E-02	intern sales	-0.01335	technician	-0.01555	representative sales	-0.01773
3	business development manager senior	-1.35E-02	area manager sales	-0.01224	administrator systems	-0.01497	consultant sales	-0.01763
4	business development manager	-1.35E-02	coordinator sales	-0.01213	engineer support technical	-0.01486	sales	-0.01756
5	analyst business intern	-1.34E-02	management trainee	-0.01192	administrator system	-0.01473	engineer sales	-0.01718
6	director senior	-1.34E-02	manager sales territory	-0.01176	support technical	-0.0147	sales vp	-0.01711
7	business development director	-1.34E-02	rep sales	-0.01157	clerk	-0.01449	rep sales	-0.01711
8	business development vp	-1.32E-02	manager marketing sales	-0.01149	customer manager service	-0.01438	salesman	-0.0168
9	business development head	-1.32E-02	district manager sales	-0.01148	supervisor	-0.01435	inside representative sales	-0.0166
10	manager senior	-1.31E-02	manager national sales	-0.01146	intern it	-0.01427	inside sales	-0.01636
11	business consultant development	-1.31E-02	manager sales senior	-0.01146	it specialist	-0.01404	manager sales territory	-0.01633
12	business development	-1.31E-02	and manager marketing sales	-0.01139	engineer field	-0.01389	executive sales senior	-0.01622
13	commercial director	-1.30E-02	manager sales	-0.01139	it manager	-0.01366	head sales	-0.01619
14	country manager	-1.30E-02	manager territory	-0.01135	administrator office	-0.01359	manager national sales	-0.01576
15	senior vp	-1.30E-02	representative sales	-0.01132	accounting intern	-0.01355	manager regional sales	-0.01575
16	general manager vp	-1.29E-02	district manager	-0.01127	customer representative service	-0.01351	area manager sales	-0.01569
17	gm	-1.29E-02	director marketing sales	-0.01119	administrator	-0.01346	director sales	-0.01556
18	coo	-1.29E-02	and marketing sales	-0.01109	assistant hr	-0.0134	and marketing sales vp	-0.01526
19	business development executive	-1.29E-02	branch manager	-0.01107	officer	-0.01333	manager territory	-0.0152
20	business manager	-1.28E-02	account manager national	-0.01106	administrator network	-0.01333	manager sales senior	-0.01519
-20	physician	-1.99E-03	engineer ii software	0.019279	copywriter	0.018512	assistant professor	0.018802
-19	clerk law	-1.98E-03	c developer	0.019282	designer graphic web	0.018667	assistant project	0.018912
-18	staff writer	-1.97E-03	engineer junior software	0.019343	art director	0.018694	assistant director	0.01892
-17	translator	-1.95E-03	design engineer software	0.019353	designer graphic senior	0.018728	advisor resident	0.019015
-16	3d artist	-1.90E-03	engineer senior software	0.019396	designer visual	0.019051	secretary	0.019064
-15	advisor resident	-1.86E-03	developer ui	0.019417	designer interactive	0.019067	human intern resources	0.019148
-14	rn	-1.85E-03	engineer r&d	0.019433	creative director	0.019116	assistant resident	0.019432
-13	lifeguard	-1.74E-03	developer junior software	0.019629	designer freelance	0.019233	teacher	0.019781
-12	assistant resident	-1.67E-03	intern r&d	0.019697	designer graphic	0.019317	assistant graduate teaching	0.019801
-11	teacher	-1.66E-03	developer software	0.01975	consultant digital marketing	0.019321	intern pr	0.020552
-10	camp counselor	-1.61E-03	engineering intern mechanical	0.019834	intern marketing	0.019364	assistant teaching	0.020719
-9	assistant graduate teaching	-1.37E-03	engineer software	0.019891	design intern	0.019793	development intern	0.020743
-8	nurse registered	-1.26E-03	engineering intern	0.020892	designer freelance graphic	0.0198	editorial intern	0.020862
-7	animator	-1.17E-03	development engineer intern software	0.021313	co creative director founder	0.020358	intern public relations	0.021695
-6	assistant teacher	-1.04E-03	developer intern web	0.021637	intern marketing media social	0.020451	assistant teacher	0.022533
-5	substitute teacher	-9.21E-04	engineering intern software	0.023186	art director freelance	0.020459	assistant s teacher	0.022925
-4	assistant teaching	-8.69E-04	development intern software	0.023356	creative director founder	0.021128	coordinator program	0.023183
-3	english teacher	-6.21E-04	intern software	0.023497	creative intern	0.021721	assistant graduate	0.024294
-2	assistant s teacher	-1.01E-04	engineer intern software	0.023928	digital intern marketing	0.02214	communications intern	0.026476
-1	tutor	5.72E-05	developer intern software	0.024543	design graphic intern	0.022745	assistant program	0.026934

Observe that the component rotation weights make intuitive sense. PC1, for instance, separates business-types and "miscellaneous" odd jobs. PC2 separates front-of-house sales and software development. PC3 separates IT and technical supervision vs. design. PC4 separates sales management and program/research interns.

APPENDIX B
WARD HIERARCHICAL CLUSTERING

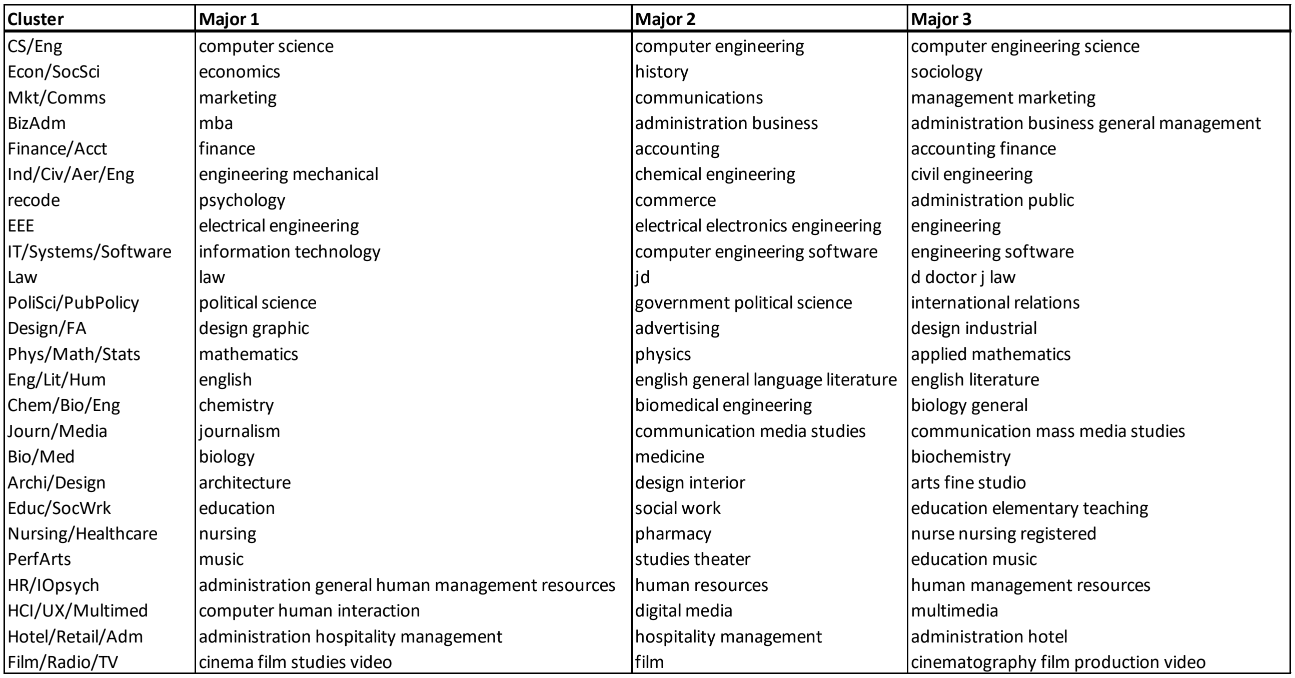
The Euclidean distances of the job title's positions in the first 12 principal components are calculated. Following which the job titles are clustered. The base of the tree gives 47 clusters. On top of this, using the grep routine, we manually assign 5 distinct clusters: (1) consultants (search on "consult") (2) business owners (search on "owner", "investor") (3) freelance and self-employed (as described in paper) (4) interns (search on "intern", "summer", "extern") and (5) founders (search on "found", "entrepreneur").

Unclustered jobs are then assigned to the clusters based on their distances of the *job title* to all the job titles in each cluster.

The top 3 job titles of each cluster is shown below, together with the assigned description.

Cluster Number	1	2	3	Description
1	"engineer software"	"developer software"	"engineer senior software"	developer.software.engineer
2	"ceo"	"president"	"director"	board.director.ceo
3	"manager project"	"manager senior"	"manager project senior"	manager.project.assistant
4	"assistant research"	"researcher"	"assistant graduate research"	research.researcher.fellow
5	"developer web"	"developer ios"	"developer end front"	developer.end.front
6	"associate"	"cfo"	"attorney"	investment.associate.attorney
7	"account executive"	"account manager"	"business development manager"	account.business.manager
8	"manager"	"general manager"	"manager operations"	manager.hr.human
9	"cto"	"engineer senior"	"engineering vp"	engineer.technical.director
10	"manager product"	"manager product senior"	"director management product"	product.manager.director
11	"vp"	"business development director"	"director operations"	vp.business.director
12	"manager marketing"	"director marketing"	"marketing vp"	marketing.manager.director
13	"designer graphic"	"designer"	"creative director"	designer.director.graphic
14	"associate sales"	"representative sales"	"sales"	sales.representative.senior
15	"analyst"	"associate research"	"analyst senior"	analyst.research.scientist
16	"assistant teaching"	"instructor"	"lecturer"	assistant.professor.adjunct
17	"manager sales"	"director sales"	"sales vp"	sales.manager.marketing
18	"analyst business"	"analyst business senior"	"engineer process"	engineer.analyst.business
19	"engineer"	"design engineer"	"engineer mechanical"	engineer.design.mechanical
20	"administrative assistant"	"manager office"	"assistant"	assistant.operations.office
21	"assistant manager"	"manager store"	"branch manager"	manager.assistant.branch
22	"director executive"	"member"	"mentor"	member.board.chief
23	"producer"	"editor"	"assistant production"	editor.producer.production
24	"manager program"	"leader team"	"engineering manager"	manager.engineering.program
25	"teacher"	"tutor"	"english teacher"	teacher.tutor.english
26	"coordinator marketing"	"assistant marketing"	"associate marketing"	marketing.communications.coordinator
27	"associate senior"	"advisor financial"	"auditor"	associate.audit.senior
28	"analyst financial"	"controller"	"director finance"	analyst.finance.controller
29	"customer representative service"	"server"	"specialist"	customer.service.manager
30	"engineer systems"	"it manager"	"engineer system"	engineer.systems.analyst
31	"recruiter"	"recruiter technical"	"executive"	manager.recruiter.services
32	"fellow"	"lifeguard"	"ambassador"	00miscellanous
33	"engineer project"	"senior"	"captain"	00miscellanous
34	"account director"	"account coordinator"	"account supervisor"	director.account.client
35	"supervisor"	"development director"	"trainer"	development.director.training
36	"copywriter"	"community manager"	"manager media social"	media.manager.social
37	"coordinator"	"coordinator program"	"assistant director"	coordinator.director.assistant
38	"administrator system"	"technician"	"engineer network"	administrator.engineer.it
39	"photographer"	"author"	"actor"	artist.professional.technical
40	"cio"	"architect solutions"	"contractor independent"	architect.solutions.chief
41	"architect"	"artist"	"designer developer web"	designer.developer.architect
42	"accountant"	"assistant legal"	"accountant staff"	accountant.accounting.assistant
43	"chief editor in"	"communications director"	"editor managing"	manager.communications.editor
44	"engineer qa"	"engineer test"	"engineer software test"	engineer.qa.software
45	"support technical"	"engineer support technical"	"application engineer"	support.technical.engineer
46	"staff"	"host"	"employee"	chef.assistant.cook
47	"stage"	"commercial"	"junior"	foreign.non.english

APPENDIX C
EDUCATION MAJOR CLUSTERS



Cluster	Major 1	Major 2	Major 3
CS/Eng	computer science	computer engineering	computer engineering science
Econ/SocSci	economics	history	sociology
Mkt/Comms	marketing	communications	management marketing
BizAdm	mba	administration business	administration business general management
Finance/Acct	finance	accounting	accounting finance
Ind/Civ/Aer/Eng	engineering mechanical	chemical engineering	civil engineering
recode	psychology	commerce	administration public
EEE	electrical engineering	electrical electronics engineering	engineering
IT/Systems/Software	information technology	computer engineering software	engineering software
Law	law	jd	d doctor j law
PoliSci/PubPolicy	political science	government political science	international relations
Design/FA	design graphic	advertising	design industrial
Phys/Math/Stats	mathematics	physics	applied mathematics
Eng/Lit/Hum	english	english general language literature	english literature
Chem/Bio/Eng	chemistry	biomedical engineering	biology general
Journ/Media	journalism	communication media studies	communication mass media studies
Bio/Med	biology	medicine	biochemistry
Archi/Design	architecture	design interior	arts fine studio
Educ/SocWrk	education	social work	education elementary teaching
Nursing/Healthcare	nursing	pharmacy	nurse nursing registered
PerfArts	music	studies theater	education music
HR/IOpsych	administration general human management resources	human resources	human management resources
HCI/UX/Multimed	computer human interaction	digital media	multimedia
Hotel/Retail/Adm	administration hospitality management	hospitality management	administration hotel
Film/Radio/TV	cinema film studies video	film	cinematography film production video