Of Hobos and Highfliers:
Disentangling the Classes and Careers of Technology-Based Entrepreneurs*

Weiyi Ng
Haas School of Business
University of California, Berkeley
2220 Piedmont Avenue
Berkeley, CA 94720
weiyi_ng@haas.berkeley.edu

Toby E. Stuart
Haas School of Business
University of California, Berkeley
2220 Piedmont Avenue
Berkeley, CA 94720
tstuart@haas.berkeley.edu

DRAFT
October 2, 2016

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.

I. Introduction

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 paper, 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.

II. 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 workplacess 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 constiute 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.

III. 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:

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

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:

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

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

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

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:

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

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:

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

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.

[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

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:

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

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:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (8)$$

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",*

[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.

"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.

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.

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 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.

Phases: Status 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:

[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.

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

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:

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

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

$$\alpha_{t\tau} = \alpha_0 + \alpha_1 t + \alpha_2 \tau \quad (11)$$

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.

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 degree information college. 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 (JD., MBA., and M.D.) as master's degree with the associated field of study as the major (law, business administration and medicine respectively).

III. 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.)

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

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.

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 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).

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

[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 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.

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 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.

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

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.

Figures 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.

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.

IV. DISCUSSION AND CONCLUSION

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.

==== I cut out this paragraph and things after ===

We point to what may have been the most vibrant time prior to the present in which organization theory's contribution. In our view, this is when organizational ecologists approached the subject through the lens of founding rate studies in which organizational populations spawn new entrants. This led to a burst of articles and genuine advancement of the entrepreneurship literature. We believe it occurred because a community of scholars embraced shared theoretical and empirical approaches, which rapidly advanced an inchoate research agenda.

We believe that career theory is next, and has the benefit of accommodating a few of the other recent, theoretical approaches to entrepreneurship. For instance, organizational theorists have thought at length about how category systems and the legitimacy of inchoate ventures funnel the entrepreneurial process. This market- and company-based approach easily dovetails to

a career theory of entrepreneurship in which the transition to different phases is determined by

… .

References

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

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

Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

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

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.

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.

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.

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., & Mosakowski, E. (1987). The career dynamics of self-employment. *Administrative science quarterly*, 570-589.

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

Elfenbein, D. W., Hamilton, B. H., & Zenger, T. R. (2010). The small firm effect and the entrepreneurial Elo, A. E. (1978). *The rating of chessplayers, past and present*. Arco Pub..

spawning of scientists and engineers. *Management Science*, *56*(4), 659-681.

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

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.

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

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Unsupervised learning. In *The elements of statistical learning* (pp. 485-585). Springer New York.
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.
Hughes, E. C. (1958). *Men and their Work*. Free Press.
Hsu, D. H., Roberts, E. B., & Eesley, C. E. (2007). Entrepreneurs from technology-based universities: Evidence from MIT. *Research Policy*, *36*(5), 768-788.
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.
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.
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.
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.
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.
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., & Stuart, T. (2002). Organizational endowments and the performance of university start-ups. *Management science*, *48*(1), 154-170.
Shane, S., & Khurana, R. (2003). Bringing individuals back in: the effects of career experience on new firm founding. *Industrial and Corporate Change*, *12*(3), 519-543.
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). 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. and J. Sonas. (2012). PlayerRatings: Dynamic Updating Methods for Player Ratings Estimation. R package version 1.0-0.
Stryker, S. (1980). *Symbolic interactionism: A social structural version*. Benjamin-Cummings Publishing Company.
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.
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.
Stinchcombe, A. L. (1965). Social Structure and Organizations in James G. March, (Eds), *Handbook of Organizations* (p142-193). Chicago: Rand McNally.
Stovel, K., Savage, M., & Bearman, P. (1996). Ascription into achievement: Models of career systems at Lloyds Bank, 1890-1970. *American Journal of Sociology*, 358-399.
Stuart, T. E., & Ding, W. W. (2006). When do scientists become entrepreneurs? The social structural antecedents of commercial activity in the academic life sciences. *American Journal of Sociology*, *112*(1), 97-144.
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.
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.
Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, *58*(301), 236-244.
Zhao, P., & Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research*, *7*(Nov), 2541-2563.
Zuckerman, E. W. (1999). The categorical imperative: Securities analysts and the illegitimacy discount. *American journal of sociology*, *104*(5), 1398-1438.

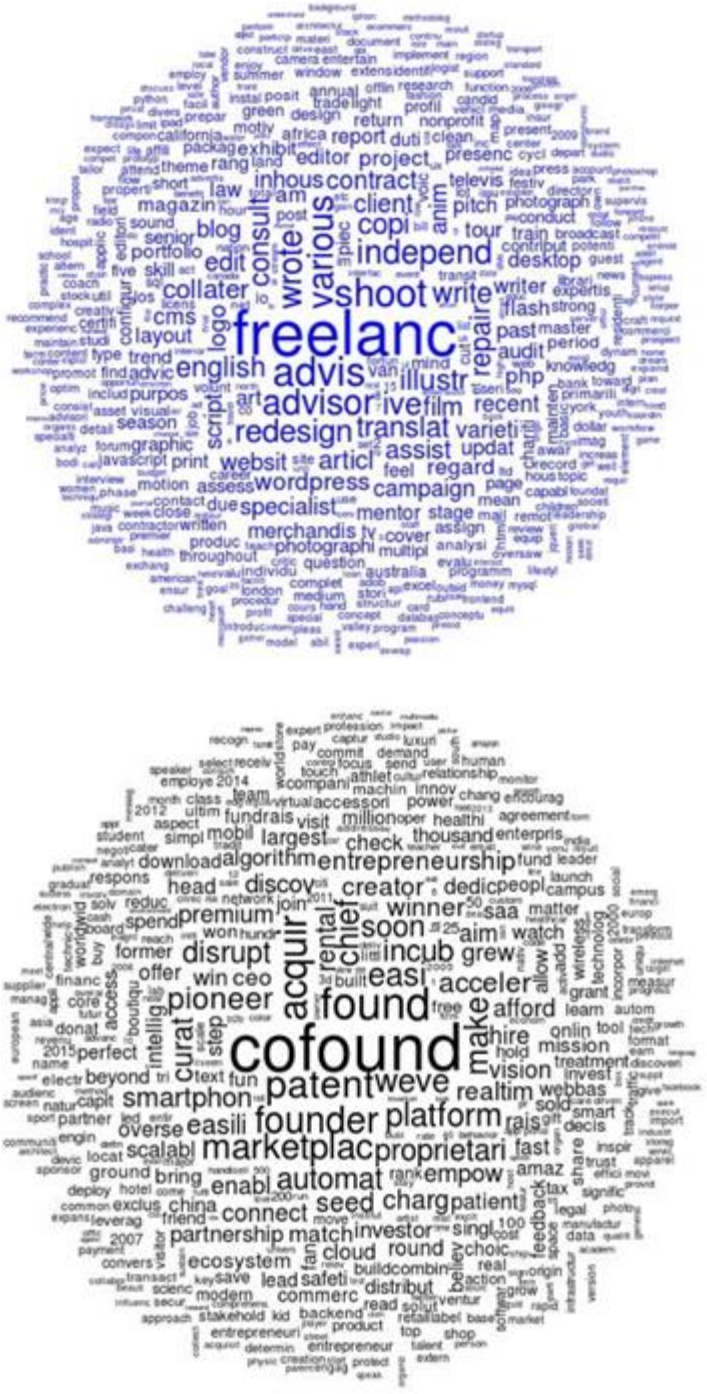


Figure 1: Word clouds representing feature weights for LASSO regression model. Negative weights imply self-employment (above); positive weights imply venture founding (below). Weights shown statistically significant at *p > 0.001* level.

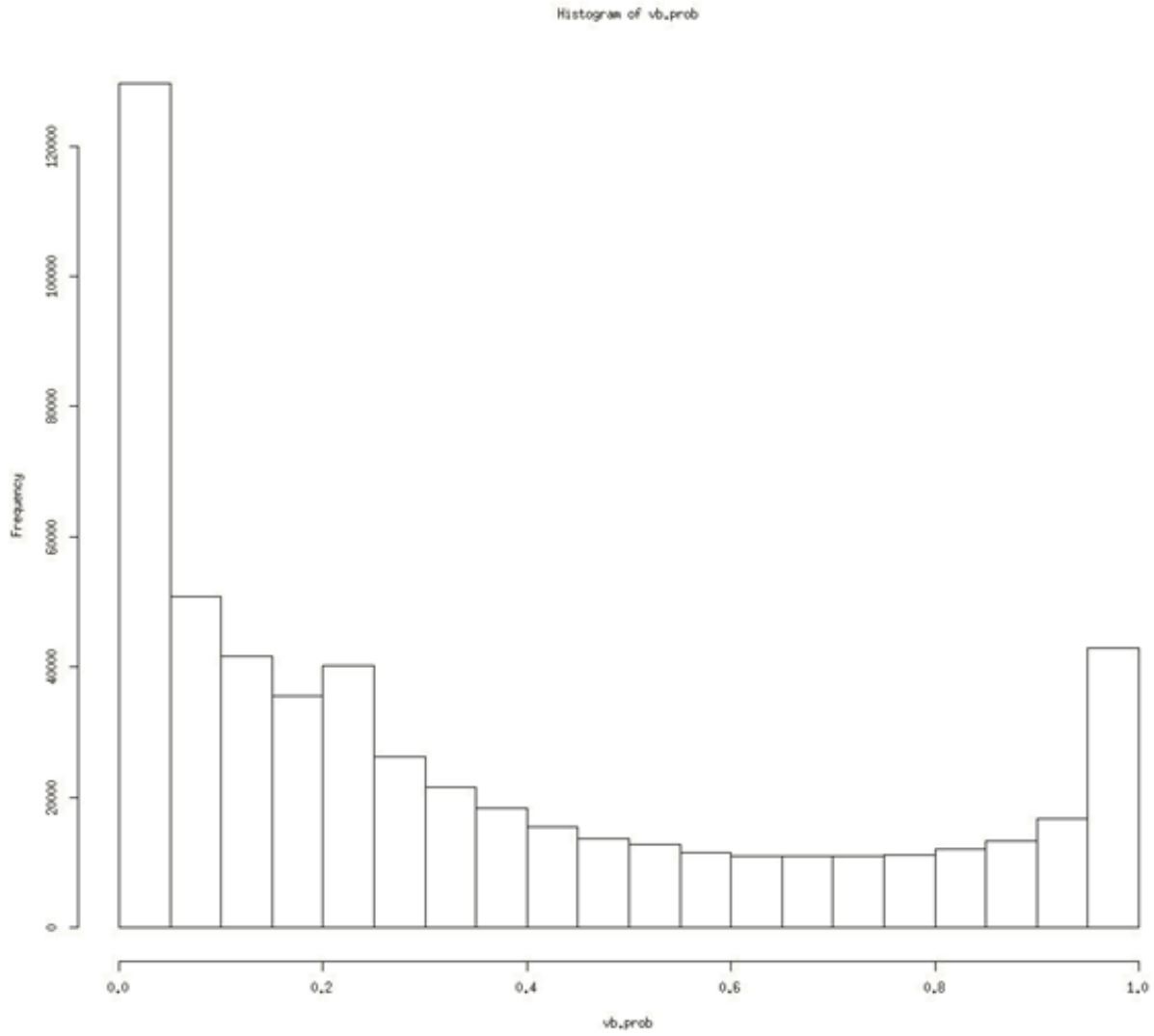


Figure 2: Distribution of predicted scores for ambiguous founding events. Bimodal distribution: scores reflect probability of event being venture founded (as opposed to a null of self-employment).

"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).

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

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

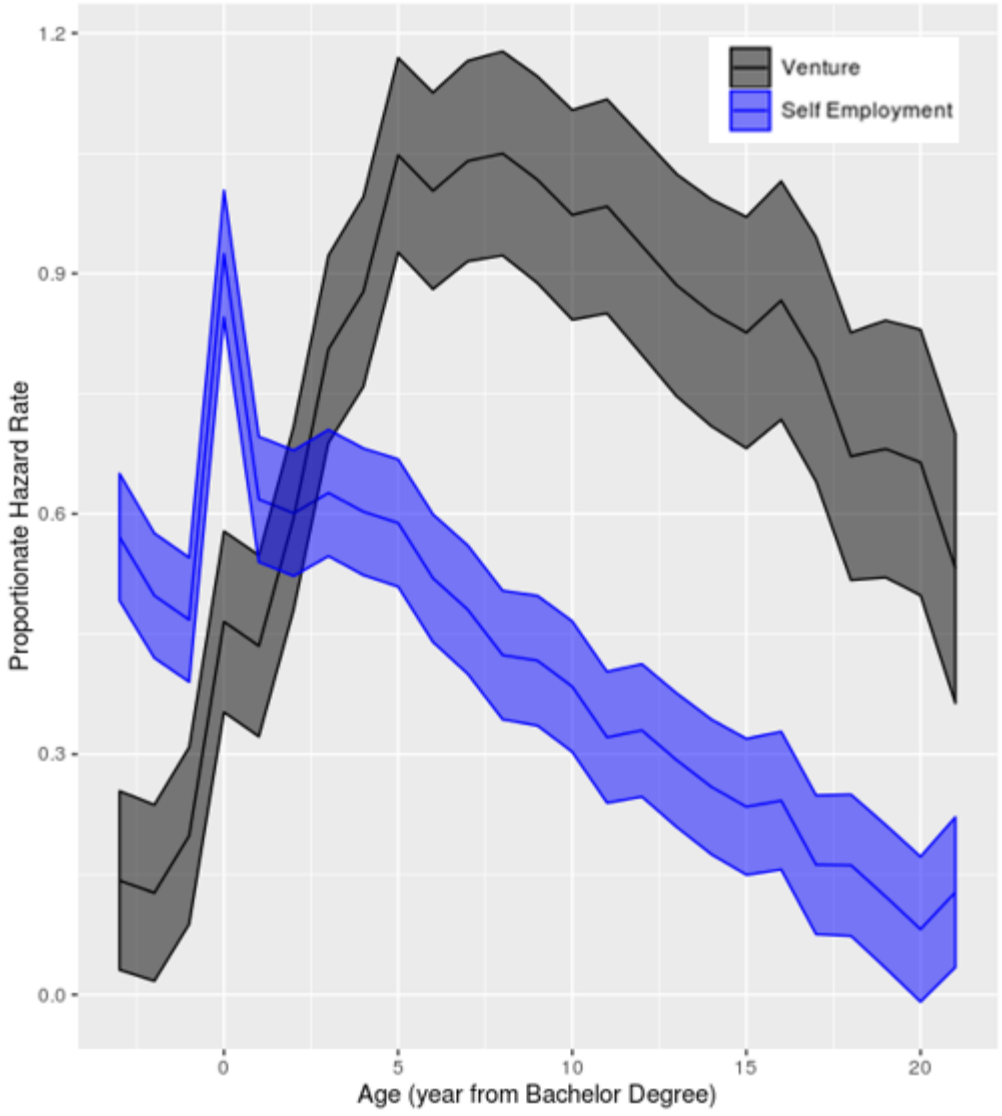


Figure 7. Base-rate normalized age fixed effects across age (time from college graduation) for high potential entrepreneurship vs. self-employment. Fixed effects coefficients are normalized by denominating with the base-rates of high potential entrepreneurship (0.003) and self-employment (0.01) respectively.

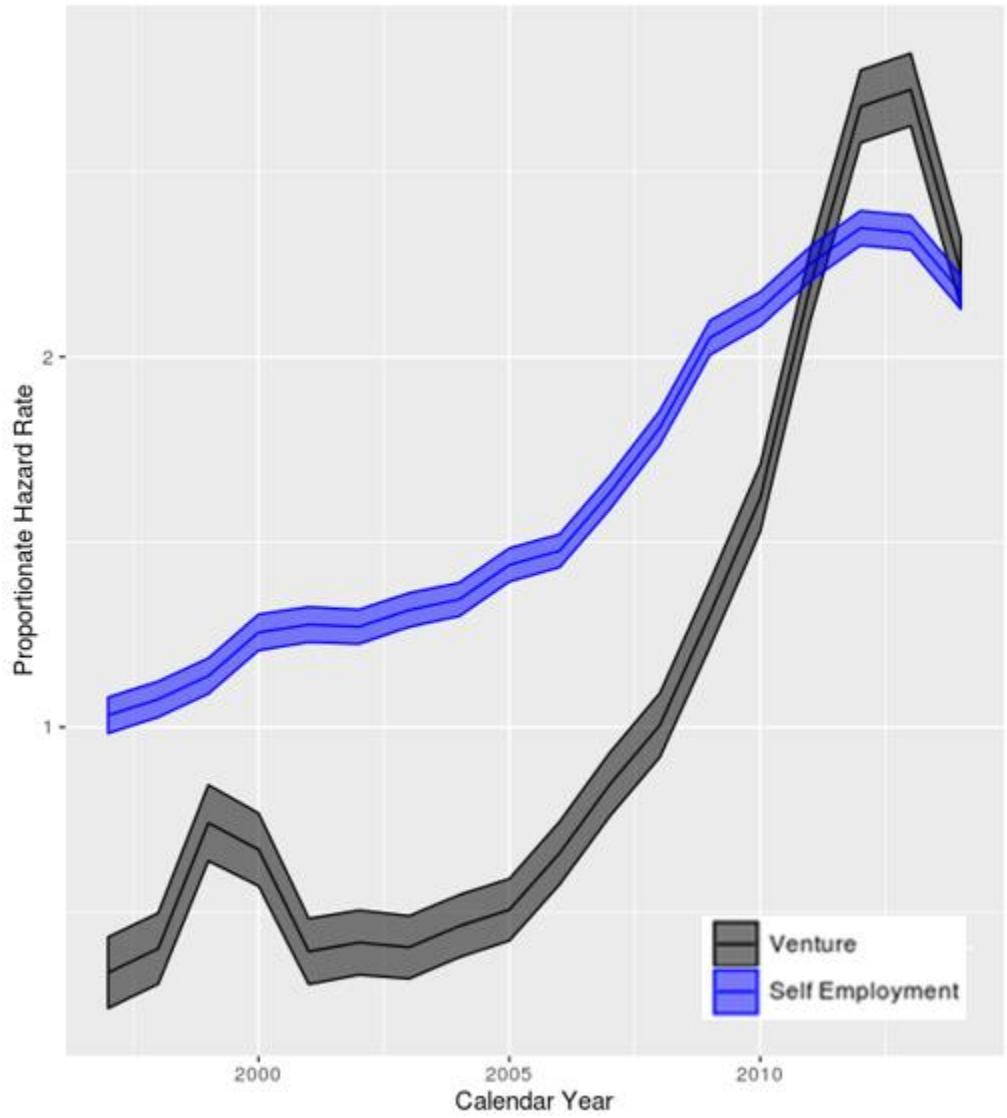


Figure 8. Base-rate normalized calendar year fixed effects across age (time from college graduation) for high potential entrepreneurship vs. self-employment. Fixed effects coefficients are normalized by denominating with the base-rates of high potential entrepreneurship (0.003) and self-employment (0.01)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

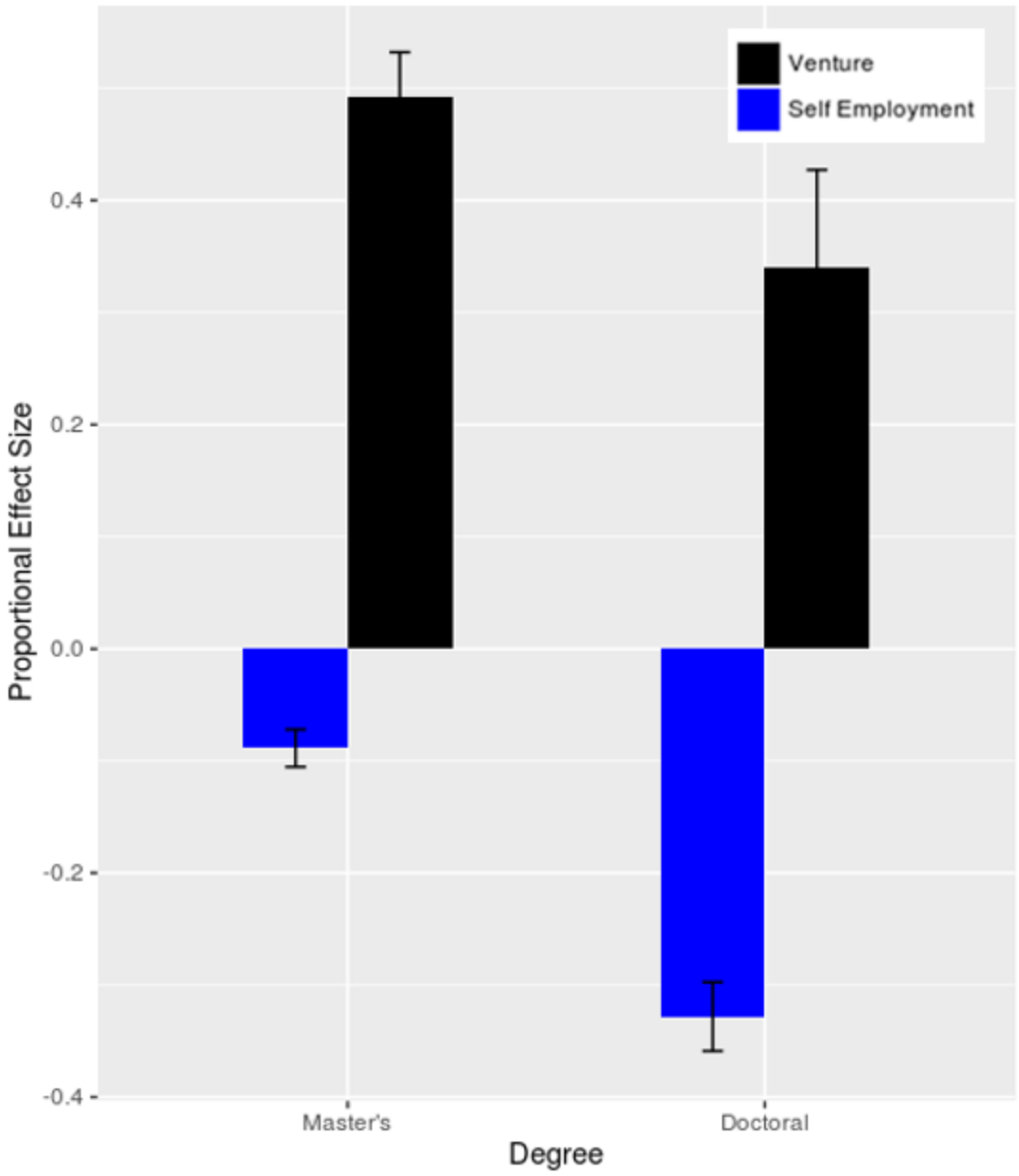


Figure 9. Base-rate normalized effects of higher education on venture founding (dark gray) vs. self-employment (light gray).

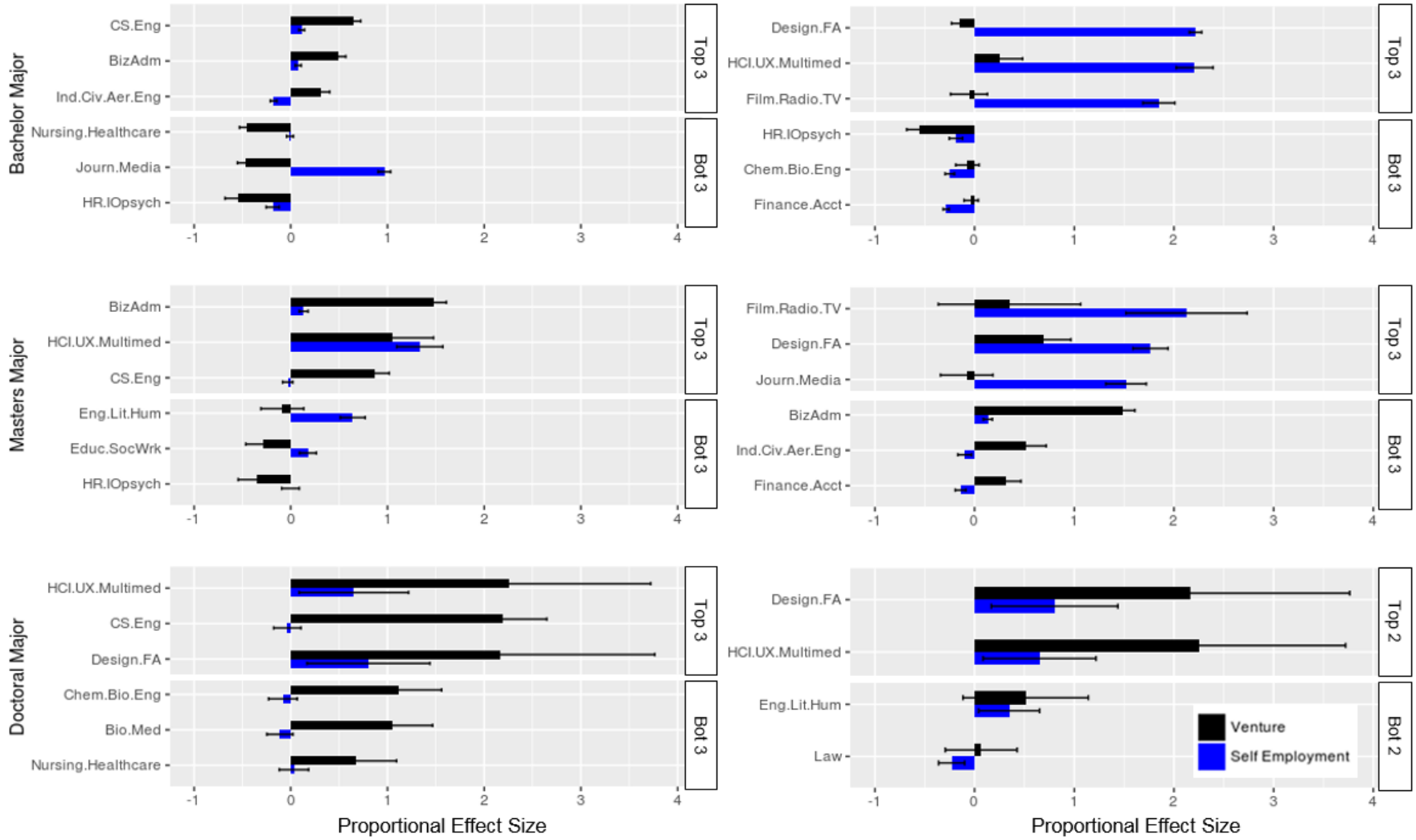


Figure 10. Top and bottom effect sizes of education majors on likelihood of venture founding (left) vs. self-employment (right). Black bars correspond to venture effects; blue bars, self-employment. For clarity, only the top and bottom 3 significant effects for each category are shown *(p < 0.05)*. Effect 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.

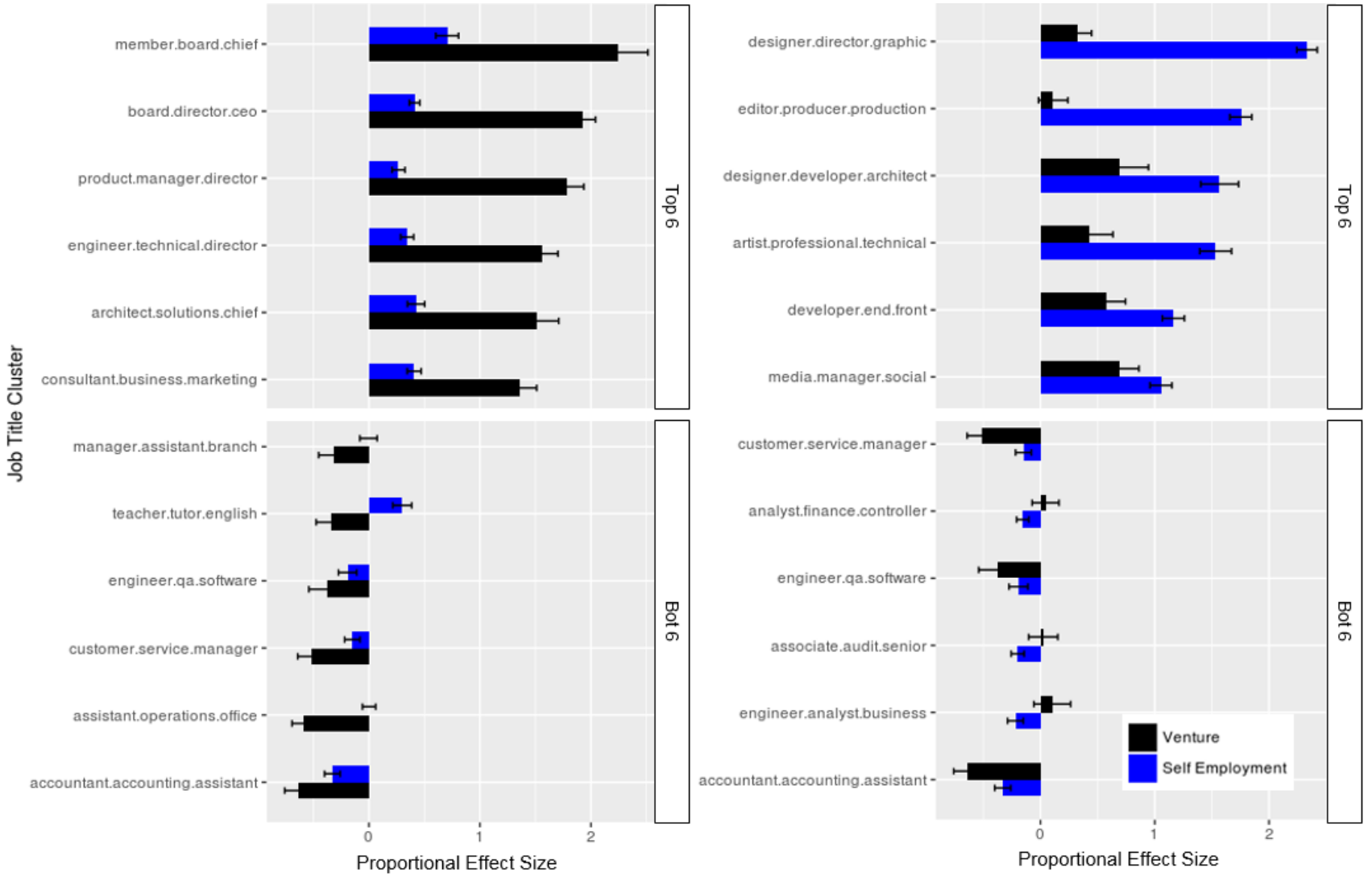


Figure 11. Top 6 and bottom 6 job title categories that lead to venture founding (left) and self-employment (right). Black bars correspond to venture effects; blue bars, self-employment. Comparison category is HR Assistant/Manager. Effect sizes here are normalized with a denominator of the base rates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

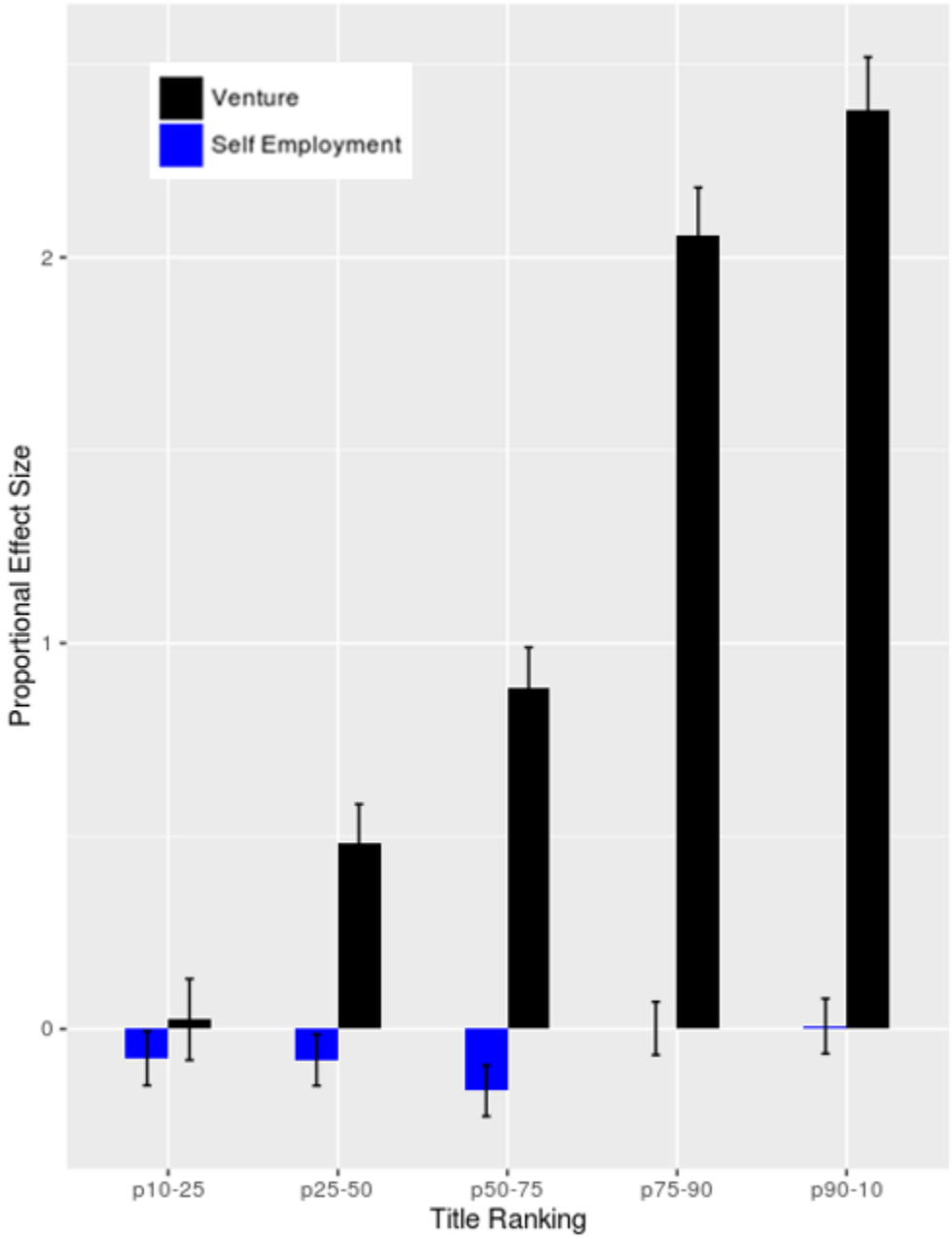
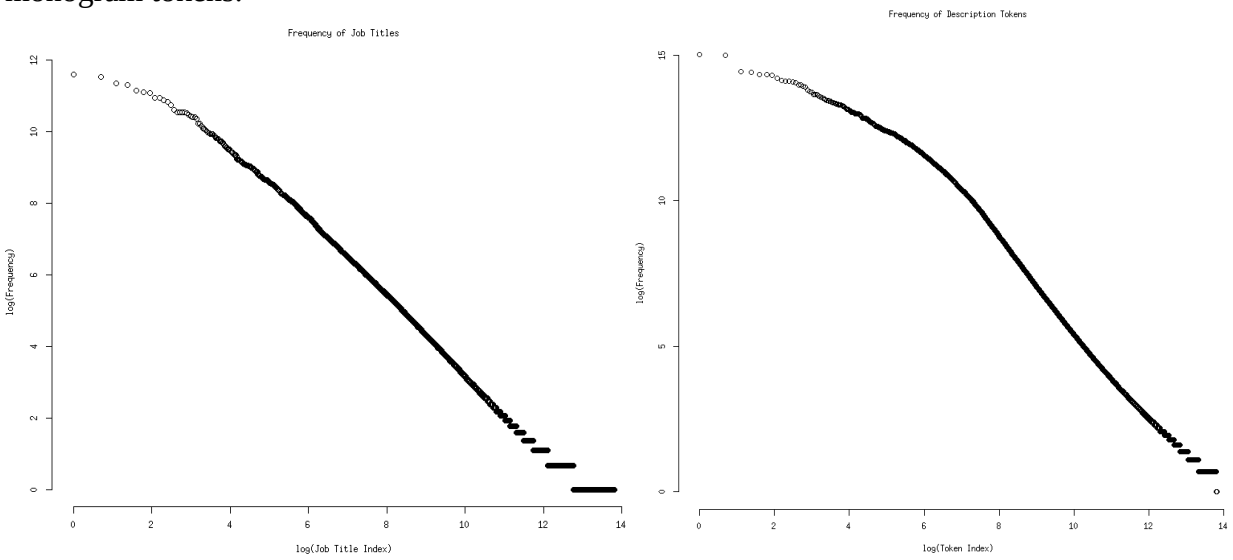


Figure 12. Effect of job title seniority on entrepreneurial transitions across 5 seniority percentile bins. Comparison group is the first decile of job title Elo rankings. Effect sizes here are normalized with a denominator of the base rates of venture founding and self-employment respectively.

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.

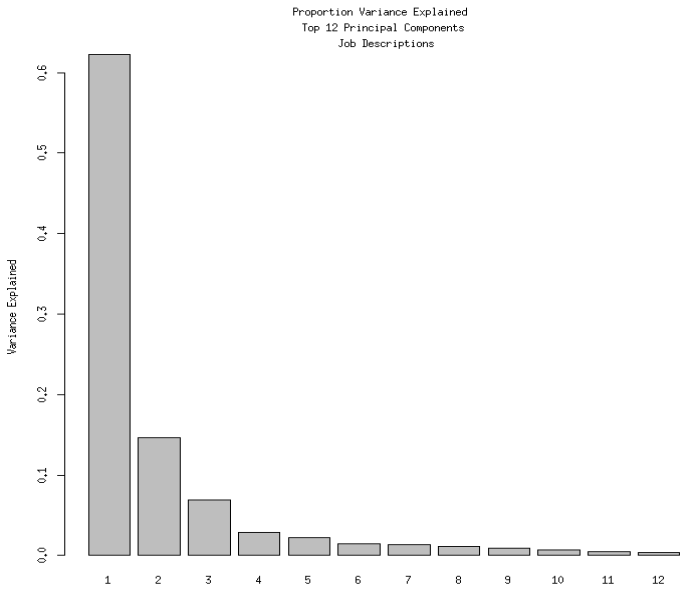


The top 16,679 most frequent monogram tokens are selected, which represent 94.68% of the total number. The distribution of frequency of these tokens are plotted as below (ordered from most to least frequent). These features are indexed and the term-document matrix is created.

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