Who Hires Whom? Entrepreneurial Backgrounds and Labor Market Opportunities*

Josh Feng, *University of Utah – David Eccles School of Business*
Xavier Jaravel, *London School of Economics* and *CEPR*

November 11, 2026

Abstract

What are the implications of unequal access to entrepreneurship for labor markets? Using data from the U.S. Census and LinkedIn profiles, we document that entrepreneurs are significantly more likely to hire workers from similar social backgrounds (gender, race, age, education, etc.). These effects are quantitatively large across several demographic dimensions. For example, the female employee share at all-female-founded startups is 36.4pp higher than at all-male-founded startups, after controlling for industry-by-metro area-by-founding-year fixed effects; the corresponding estimates are 51.2pp for Black workers, 37.3pp for Hispanic workers, and 11.3pp for non-college individuals. Large effects are present in high-growth startups, across industries and occupations, and remain stable across new firm cohorts. In addition, we find that these differences persist for at least 20 years. We use wage data and an AKM research design to untangle whether the relative differences are driven by labor demand or labor supply effects. We find that demand drives the differences: group-specific wage decompositions show that new firms pay higher relative wages to individuals from similar backgrounds to the entrepreneur. Using these estimates, we calibrate a model of entrepreneurship with heterogeneous ability and heterogeneous production functions in terms of labor inputs, and assess the impacts on relative wages from reducing access barriers to entrepreneurship. Across a range of calibrations, we find long-run relative wage gains of

*We thank Naomi Hausman, Ben Jones, Mike Roach, Tim Simcoe, Elena Simintzi, and Melanie Wallskog for providing detailed feedback. We thank participants at the Banff Innovation, Productivity, and Growth Conference, the NBER Productivity Lunch, the NBER Summer Institute, REER, the Strategy Science Annual Conference, and the Wasatch Front RDC seminar for thoughtful comments. Any views expressed are those of the authors and not those of the U.S. Census Bureau. The Census Bureau has reviewed this data product to ensure appropriate access, use, and disclosure avoidance protection of the confidential source data used to produce this product. This research was performed at a Federal Statistical Research Data Center under FSRDC Project Number 2229 (disclosure numbers CBDRB-FY25-P2229-R11972 and CBDRB-FY26-P2229-R12962). This research uses data from the Census Bureau's Longitudinal Employer-Household Dynamics Program, which was partially supported by the following National Science Foundation Grants SES-9978093, SES-0339191 and ITR-0427889; National Institute on Aging Grant AG018854; and grants from the Alfred P. Sloan Foundation.

4–12pp for women, 2–13pp for Black workers, 3–17pp for Hispanic workers, and 1–4pp for non-college workers.

How does the composition of entrepreneurs in an economy impact individuals' labor market opportunities and outcomes? Existing work by Haltiwanger et al. (2013) has documented that young firms play a significant role in net job creation in the economy. In a separate line of work, researchers have documented that specific socioeconomic groups (women, low-SES individuals) are dramatically underrepresented in the pool of entrepreneurs and innovators (e.g., Bell et al., 2019; Guzman and Kacperczyk, 2018; Cook et al., 2026; Bento and Hwang, 2023; Bento, 2024; Glaeser and Hausman, 2020). If entrepreneurs are better at managing or creating organizational structures that make use of workers with similar backgrounds to them, then individuals from these underrepresented backgrounds may face adverse labor market opportunities. To date, there has been little systematic evidence on this issue, given the difficulty in assembling systematic data on the demographics of entrepreneurs and workers.[1]

In this article, we study the impacts of "unequal access" to entrepreneurship and business ownership on workers by compiling two complementary datasets on new firms and the demographics of their founders and workers. Our primary data source comes from the Longitudinal Employer-Household Dynamics (LEHD) data collected by the U.S. Census, which provides broad coverage of new firms across 30 states and the District of Columbia, the vast majority of which are ordinary small businesses rather than startups with high growth potential.[2] To add evidence on high-growth-potential startups, we also compile a complementary dataset based on LinkedIn career history data compiled by Revelio Labs, linked to information on founders from Crunchbase, a tracker of venture-backed startups.[3]

Using both datasets, we first test whether entrepreneurs are more likely to hire individuals of similar socioeconomic backgrounds in the early years of each firm. We then explore heterogeneity by firm characteristics and worker roles, in order to gain further insights into the potential mechanisms driving the high-level results. Next, we use wage data from the

[1]The closest work to ours is Chiplunkar and Goldberg (2024), which we discuss in detail in the literature review below. Our work is also related to Gozen et al. (2024), who analyze differences between male- and female-owned businesses using historical U.S. Census of Manufactures data from the 19th century.

[2]Appendix Figure A1 maps the covered states. Appendix Section A.B compares their population with the United States. Appendix Section B.D.3 reports estimates restricting the Crunchbase--LinkedIn sample to the covered states.

[3]These data also address limitations of the LEHD data by providing direct measures of founders and employee roles.

LEHD to provide evidence on equilibrium wage differences at the firm level, in order to test whether labor supply or labor demand factors drive the heterogeneity in hiring, which is a key factor to assess the impact of unequal access to entrepreneurship on wage inequality. Finally, we use our empirical estimates to calibrate a model where entrepreneurs create firms with heterogeneous production functions in terms of labor inputs. We use the calibrated model to assess the implications of our estimates for labor market opportunities and wage gaps.

As the starting point of our analysis, we establish results on the similarity between founders and their employees. To do so, we compute simple correlations between the demographics of the founders (e.g., one-third female) and the demographics of the employees at the firm in a given year (e.g., two-thirds female), both raw and adjusted by founding year, location, and industry. We do so for a series of demographic variables: age, gender, race, and college education. Throughout, we find large estimates of "homophily"—the tendency for founders to employ workers who share their demographic background—across all demographic variables. For example, within founding year by metropolitan statistical area (MSA) by industry groups, firms founded entirely by women employ a 36.4pp higher fraction of female workers during their first six years than firms founded entirely by men. The corresponding differences are 51.2pp for Black workers, 37.3pp for Hispanic workers, and 11.3pp for non-college individuals. We provide further evidence that homophily is quantitatively relevant for firms with high growth potential, by focusing on the fastest-growing firms in the LEHD and by analyzing the Crunchbase–LinkedIn sample.[4]

Next, we assess how broadly homophily applies across firms and workers. First, we find that homophily is a consistent feature across the vast majority of industries and founding-year cohorts. Within firms, homophily is similar by employee seniority and broad occupation groups (e.g., engineering versus marketing), and slightly increasing in pay rank. Homophily also persists for at least ten years in the LEHD and twenty years in the LinkedIn-Crunchbase sample.[5] Furthermore, as firms age, founder demographics remain strongly predictive of employee composition when controlling for initial employee composition. These findings

[4]Although there are quantitative differences between the Census LEHD and Crunchbase–LinkedIn samples in terms of homophily estimates, we provide evidence that this is driven by differences in the measurement of founders and potential attenuation bias from misclassification of founder demographics in the latter dataset.

[5]Homophily between founder and new hires does decrease significantly over time, but is partially offset by differential retention.

motivate our modeling of hiring homophily as a persistent, founder-based characteristic of firms that applies broadly across the economy.

Then, we use evidence from wages to test whether labor supply or labor demand is the primary driver of the observed founder-employee homophily, because these mechanisms have different implications for the potential wage impacts of broadening access to entrepreneurship. To do so, we run wage decomposition regressions within demographic subgroups (e.g., college-educated versus not) to extract firm fixed effects, following the approach in Card et al. (2016) that extends Abowd et al. (1999). Using these estimates, we then compute the differences in firm fixed effects for each demographic group (e.g., college-educated estimate minus the estimate for non-college-educated) and estimate whether the differences correlate with founder characteristics (e.g., having a college-educated founder). The logic of this test is as follows: if homophily reflected workers accepting lower pay for the amenity of working with founders like themselves ("labor supply" effects) then own-group workers would earn relatively *less* at own-group firms, a compensating differential. Instead, we find large and positive relative wage effects for workers that share demographics with the founders, ranging from 2.9 to 12.9 log points across demographic variables.[6] This points to labor demand effects dominating labor supply effects, i.e., founders being willing to pay more to hire own-group workers.

Finally, we incorporate homophily and wage estimates into a model to assess the impacts of unequal access to firm creation on labor market opportunities and outcomes. The combination of the homophily estimates and wage premium estimates motivates two key model features: i) founders create production functions that make own-group workers relatively more productive; and ii) an upward-sloping labor supply curve at the firm level in order to allow for higher relative own-group pay. We calibrate the model using our core estimates, and then simulate counterfactuals under different compositions of founders to assess the impact on equilibrium wage gaps across worker groups. In our preferred calibration, we find steady-state relative wage gains of 8pp for women and Black workers, 5pp for Hispanic work-

[6]We test the robustness of our estimates in several ways: i) we show that worker-firm interaction models have similar fit and that the degree of assortative matching is similar across startups and other firms (Bonhomme et al., 2019); and ii) we present raw evidence that movers exhibit minimal pre-trends and large earnings gains when moving from a low demographic-specific wage premium firm to one with a high premium.

ers, and 2pp for non-college workers. We also calibrate and compute counterfactuals under alternative parameter inputs and homophily estimates, and find gains ranging from 4–12pp for women, 2–13pp for Blacks, 3–17pp for Hispanics, and 1–4pp for non-college workers.

We contribute to three strands of literature. First, we extend the literature on unequal access to entrepreneurship and business ownership. Existing work has documented that women, minorities, and individuals from low-SES backgrounds are underrepresented among entrepreneurs and inventors (e.g., Bell et al., 2019; Guzman and Kacperczyk, 2018; Bento and Hwang, 2023; Bento, 2024). Census Bureau Annual Business Survey estimates also show substantial increases in the number of Black- and Hispanic-owned employer businesses between 2020 and 2022 (U.S. Census Bureau, 2022, 2024). A related strand focused specifically on the innovation system has documented that innovators tend to innovate for consumers similar to them (e.g., Koning et al., 2021; Einio et al., 2023). Our study complements this work by studying the implications of "unequal access" to entrepreneurship for the *workers* that firms hire, rather than for consumers.

Second, we contribute to the literature on new firm creation and hiring. Seminal work in the area noted the importance of young firms for net job creation (e.g., Haltiwanger et al., 2013). Newer work in the literature has documented additional issues surrounding pay at startups (e.g., Kim, 2018; Sorenson et al., 2021; Roach and Sauermann, 2023) and the behavior and importance of early workers at startups (Bryan et al., 2022; Choi et al., 2023; Bessen et al., 2023; Balsmeier et al., 2025). Our work adds additional evidence to this line of work by highlighting the workers that gain from new firm creation using both administrative data and a complementary source focused on venture-backed startups.

Finally, we contribute to the literature on labor market opportunities and outcomes. A long line of literature has documented the drivers of the gender wage gap (e.g., Goldin and Rouse, 2000; Tate and Yang, 2015; Card et al., 2016; Cortés and Pan, 2023) and the wage differences across racial and ethnic groups (e.g., Chetty et al., 2020). This includes specific work on the impact of managers' social backgrounds and experiences on worker outcomes at firms (Tate and Yang, 2015; Ronchi and Smith, 2024). More recent work has studied female entrepreneurship, hiring, and wage setting in other contexts (Campero and Kacperczyk, 2020; Chiplunkar and Goldberg, 2024; Gozen et al., 2024; Chen et al., 2025;

Ashraf et al., 2025).[7] Beyond gender, Miller and Schmutte (2021) study the mechanisms behind racial co-hiring using administrative data from Brazil, Lerner et al. (2025) study how impact investors influence the wages of minority employees, and Hwang and Wallskog (2026) analyze the impact of entrepreneurship on the labor market opportunities of "marginalized workers," finding that successful young firms do have a positive impact on the careers of these workers. Our work aims to assess the role played by entrepreneurs of different backgrounds, specifically the types of organizations created by the current pool of entrepreneurs. We can thus assess the extent to which unequal access to entrepreneurship careers might induce lasting inequalities across workers in terms of job opportunities and wages across socio-demographic groups.

The remainder of the article is organized as follows. Section I discusses data sources and research design. Section II presents estimates on hiring homophily. Section III provides evidence on heterogeneity and mechanisms. Section IV presents the wage analysis. Section V presents and calibrates a model of entrepreneurship with heterogeneous production functions and access barriers and shows the policy counterfactuals. Section VI concludes.

I Data Sources and Methodological Approach

We discuss our two key data sources, data construction, and our methodological approach for measuring hiring homophily.

I.A Census LEHD Data

First, we make use of the Longitudinal Employer-Household Dynamics (LEHD) data from the U.S. Census Bureau. The Census data have several advantages. First, they contain administrative wage data, which makes it more complete than LinkedIn data in terms of measuring workers and provides a reliable way to adjudicate between labor supply and labor

[7]The closest work to ours is Chiplunkar and Goldberg (2024), both in terms of empirics and modeling. Our work differs in several ways. First, we study a developed economy context and document evidence across several demographic dimensions, rather than gender alone. Second, we provide detailed wage analysis, which can distinguish labor demand versus labor supply as drivers of homophily. Third, motivated by our wage analysis, we model homophily as driven by heterogeneous production functions rather than differential search costs. Chiplunkar and Goldberg (2024) assume constant market-level wages for each worker group, but, if hiring frictions reflect difficulty in attracting opposite-group workers and firms have some wage setting power, then there would be an opposite-group wage premium. This is inconsistent with our wage estimates.

demand effects. Second, they contain demographic information on age, race/ethnicity, gender, immigration status, and education level, collected from the Decennial Census. Finally, Census data also contain information on industry and location. One disadvantage of using the LEHD data is that they are only available for 30 states and the District of Columbia. We provide a detailed discussion in Appendix A.B, including evidence that the population in the covered states has similar demographics relative to the overall U.S. population.

Following Kerr and Kerr (2016), we flag the top three earners in each firm's first year in the LEHD as founders and record their demographics, filter out firms with earlier entries in other Census datasets (to exclude establishments opened by firms already operating in uncovered states), and compute average demographics of non-founder employees in each year relative to founding.[8] Appendix A provides the construction details.

Table 1 presents summary statistics from a balanced panel of new firms covering the period 1992–2014. We require firms in the balanced panel to have at least one non-founder employee in each of years 0–5 relative to founding, with valid MSA and industry information. This ensures that we can estimate trends in homophily as firms age without incorporating selection effects from exit and can incorporate detailed fixed effects in our regressions.[9] We use this as the core sample throughout.

The summary stats highlight discrepancies between founder and employee demographics: women, Black workers, Hispanic workers, younger individuals, and non-college individuals are underrepresented in the founder statistics relative to their fraction in the new-firm employee pool and the general population. In terms of firm size, the average firm grows 48 log points in terms of employees from year 0 to year 5. Average firm-year employment is 30.

We provide some context for these statistics using estimates in other reports. In terms of employees, for firms within five years of founding, published statistics based on data from 2000–2014 imply ranges of 39.2–55.2 percent female, 4.8–15.9 percent Black, and 9.7–21.7 percent Hispanic. These contain our underlying employee shares. On the founder side, relative to the estimates from the 2014 Annual Survey of Entrepreneurs (ASE), our

[8] We later provide estimates under different assumptions for founder measurement.

[9] As a practical matter, Census disclosure rules limit the number of unique samples we can use for disclosure. This approach also implicitly selects for larger and relatively successful firms. We later provide results using an unbalanced sample of new firms and extend the relative years to 9 years after founding.

Table 1: Summary Statistics – New Firms in LEHD

Panel A: Firm Level

	Mean	S.D.	Pseudo Median
Female Founder Fraction	0.408	0.367	0.333
Black Founder Fraction	0.054	0.170	0.000
Hispanic Founder Fraction	0.105	0.227	0.000
Average Founder Age	38.32	9.142	38.00
College Grad Founder Fraction	0.281	0.296	0.333
Change in Log Employees	0.484	1.049	0.406
Observations		825,000	

Panel B: Firm-by-Year Level

	Mean	S.D.	Pseudo Median
Female Employee Share	0.501	0.332	0.500
Black Employee Share	0.075	0.166	0.000
Hispanic Employee Share	0.131	0.213	0.023
Average Employee Age	32.92	8.722	32.33
College Grad Employee Share	0.226	0.213	0.186
Total Employees	29.06	188.7	10.000
Observations		4,947,000	

Notes: Panel A presents statistics at the new firm level for the balanced sample of firms. These include the demographics of the founders (based on the top three earners in the firm's first year) and the growth in employees from year 0 to year 5. Panel B presents statistics at the firm-by-year level, focusing on employee demographics in each year. Pseudo medians are averages of data points closest to the median. All statistics are unweighted averages. Observation numbers are rounded based on Census disclosure rules. Source: authors' calculations based on U.S. Census data (FSRDC Project Number 2229, Disclosure Review Board (DRB) approval number: CBDRB-FY25-P2229-R11972).

founder sample contains higher shares of female, Black, Hispanic, and non-college founders.[10] These differences could be driven by several factors, including: i) our sample of states is not representative and our sample covers a longer period; ii) our top-three-earner proxy is overly inclusive; and iii) coarse ownership variables in the publicly available ASE data.[11]

I.B LinkedIn and Crunchbase Data

To address issues in founder measurement and to focus on a sample of startups with high growth potential, we build a second sample of startups by matching LinkedIn data compiled by Revelio Labs to Crunchbase startup data.[12] Our Revelio Labs dataset is an April 2023 snapshot of LinkedIn, and our Crunchbase dataset comes from a December 2020 snapshot.

Revelio Labs data provide two main advantages. First, we can identify founders based on LinkedIn job titles, instead of guessing based on the top paid individuals at a firm.[13] Second, they contain raw and standardized information on the role of the individual at a given firm (nature of job and seniority). This allows us to test whether our observed effects are driven by people in certain roles (e.g., scientists vs. human resources) or seniority levels.

However, the data come with several limitations. First, individuals select into having LinkedIn profiles. Recent work by Dorn et al. (2025) documents the overrepresentation of college-educated, management, finance, and IT workers relative to the American Community Survey. Second, the data do not contain wage information. Third, the demographic variables are constructed based on names rather than reported through the decennial Census, and there are fewer readily available characteristics than in the Census.

We again construct two subsamples using the merged dataset. First, we construct a balanced sample of firms that have non-founder employees with LinkedIn profiles in each year up

[10] Based on our calculations using the 2014 ASE, the owners of firms with employees aged five years or less were 31.2 percent female, 3.0 percent Black, and 8.5 percent Hispanic, and 51.8 percent were college graduates (U.S. Census Bureau, 2014). The college founder share exhibits the biggest discrepancy.

[11] The public data only records ownership demographics based on the majority owner (or owners in case of equal ownership).

[12] Revelio Labs provided a crosswalk from their firm ID to Crunchbase firm IDs, which we rely on for our analysis.

[13] Hyatt et al. (2021) show that the standard approach for identifying entrepreneurs in the LEHD data is unreliable. The approach does not work for sole proprietors and partnerships. In addition, for S-corporations, the top three earners usually do include the business owners, but most of the top three individuals are not themselves owners (high recall, low precision).

Table 2: Summary Statistics – New Firms in the Crunchbase-LinkedIn Sample

Panel A: Firm Level

	Mean	S.D.	Median
Female Founder Fraction	0.20	0.32	0.01
Black Founder Fraction	0.08	0.15	0.02
Hispanic Founder Fraction	0.07	0.17	0.01
Change in Log Employees	1.21	1.31	1.16
Observations		44,545	

Panel B: Firm-by-Year Level

	Mean	S.D.	Median
Female Employee Share	0.35	0.26	0.33
Black Employee Share	0.09	0.10	0.06
Hispanic Employee Share	0.09	0.14	0.05
Total Employees	85.14	1910.43	10.00
Observations		267,270	

Notes: Panel A presents statistics at the new firm level for the balanced sample of firms founded in or after 2007 that have non-founder employees with LinkedIn profiles in each of the years 0 to 5 after founding. These include the demographics of the founders (as identified from raw LinkedIn job titles) and the growth in employees from year 0 to year 5. Panel B presents statistics at the firm-by-year level, focusing on employee demographics in each year. Source: Crunchbase merged to LinkedIn (Revelio Labs).

to 5 years after founding, focusing on firms founded after 2007, roughly the time Crunchbase starts tracking startups. We also construct an unbalanced panel of all Crunchbase firms, removing the founding year filter.

Table 2 reports summary statistics for the balanced sample. We again find that founder shares are smaller than average employee demographics, although the differences are much larger for women than for Black and Hispanic workers. Firms in this sample exhibit higher employment growth over the same six-year period and have more employees on average. This is unsurprising given that the Crunchbase sample is skewed towards venture-backed and technology firms and we are ensuring survival in our balanced sample.

Relative to Table 1, the female founder fraction is much lower (20 versus 41 percent). This is consistent with evidence on venture-backed and technology-focused startups in the literature, where estimates range from 13 to 21 percent (Ewens and Townsend, 2020; Cook et

al., 2026).[14] In contrast, our Black founder share (8 percent) is significantly higher than that found in the literature for venture-backed startups and also higher than the corresponding number in the Census sample (5 percent).[15] The most likely driver of the discrepancy is our name-based demographic measures. We revisit this when presenting results using this sample.

I.C Research Design

To estimate whether founders hire individuals from similar backgrounds, we employ a simple specification:

$$\text{WorkerType}_{it} = \alpha + \beta_t \cdot \text{EntrepreneurType}_i + \mu_k + \varepsilon_{it} \tag{1}$$

where i indexes the firm, t the age of the firm (equal to zero in the year in which the firm is founded), and μ_k indexes fixed effects at some level, with one category omitted so that α is the intercept for the reference cell. We run separate regressions at each horizon (one β_t for each age t). "WorkerType" measures the fraction of workers who are of a given characteristic (e.g., female) and "EntrepreneurType" is the corresponding measure within the set of founders. To start, we analyze new firms in the first year of their creation. Following Kerr and Kerr (2016), which stresses the importance of geography in determining startup outcomes, we use both detailed industry (six-digit NAICS) and two-digit industry-by-MSA level and interact them with founding year (cohort) to facilitate the comparison of similar firms.

We note a couple of features of the β coefficient. First, when the founder and worker measures are binary, the coefficient is bounded between -1 and 1, because it is a straight comparison of worker shares. This property breaks when we use founder fractions or in the case of age. Second, the coefficient is on the same scale as standard measures in the broader literature on homophily and segregation (McPherson et al., 2001; Currarini et al., 2009). In the binary case without fixed effects, β equals zero under random matching and one under complete segregation. This corresponds to the isolation index that Gentzkow and Shapiro

[14] Cook et al. (2026) study a sample of U.S. startups in PitchBook founded between 2000 and 2020. Their sample has an average female founder fraction of 13 percent. Ewens and Townsend (2020) study startups seeking funding on AngelList, and find that 21 percent of them have at least one female founder.

[15] Using profile pictures and manual review, Cook et al. (2026) find that 3.1 percent of U.S.-based startups in Pitchbook are Black-owned.

(2011) use to measure ideological segregation.[16] With fractional founder shares, the slope is no longer bounded by one (e.g., if mixed gender founding teams hire all female workers and male-founded startups hire all male workers), but all of our estimates are well inside the -1 to 1 range.

Using the basic homophily regression as the building block, we extend the analysis along several dimensions. First, we assess the robustness of the core estimates to the founder definition, weighting by measures of importance (e.g., by employee income and firm size), and the level of geographic fixed effects. Second, we study persistence beyond the first five years using unbalanced panels. Third, we estimate heterogeneity along worker dimensions (e.g., wage quintile, seniority, occupation) and firm dimensions (e.g., sector, growth) to speak to underlying mechanisms. Fourth, we break down the observed persistence by separately analyzing new hires and by running a horserace between founder and initial employee demographics.

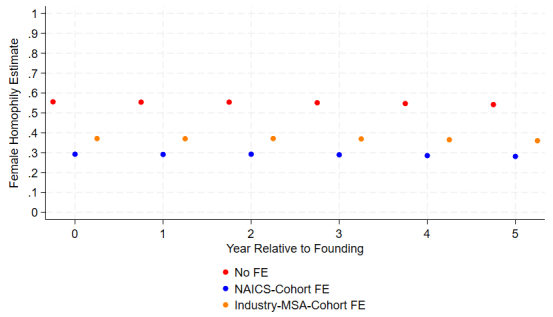
II Hiring Analysis

We start by presenting homophily estimates from the core LEHD sample. We then provide evidence on the importance of homophily by addressing two additional issues: i) whether homophily remains quantitatively similar within startups with high growth potential; and ii) the persistence of homophily as firms age.

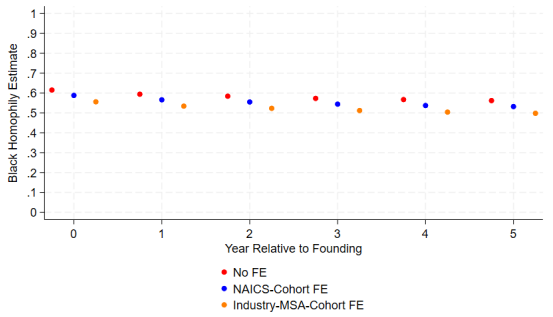
II.A Core Results – LEHD

Figure 1 presents the core estimates of Equation 1 from the Census LEHD balanced sample. The estimate for female founders in year 0 with the most detailed fixed effects is 0.37. The corresponding year-0 numbers for Black founders are 0.556 and for Hispanic founders 0.406. We also find large effects for Asian (0.574), age (0.364), and college-educated (0.118). The detailed controls reduce the quantitative estimates relative to the specification with no controls, which suggests that a significant part of the raw estimate is driven by differences within very similar firms.

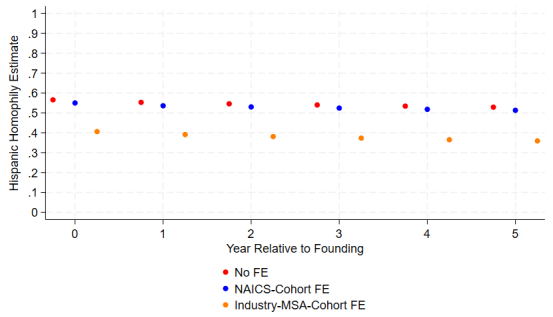
[16] Their measure focuses on one population, whereas we study two separate populations (founders and employees).



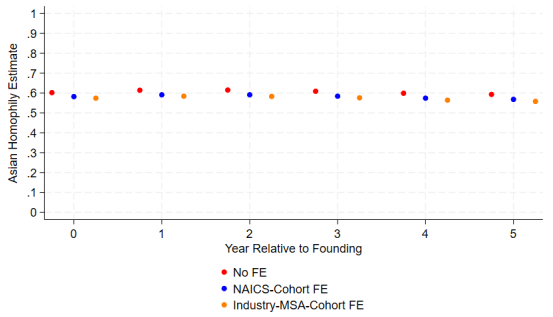
(a) Female



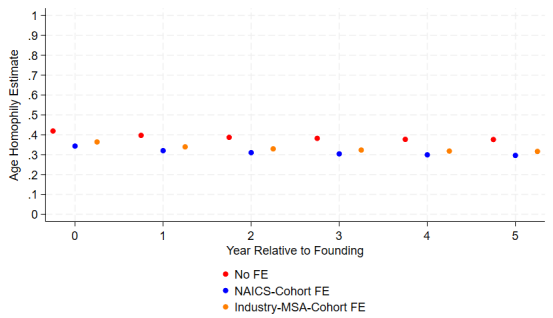
(b) Black



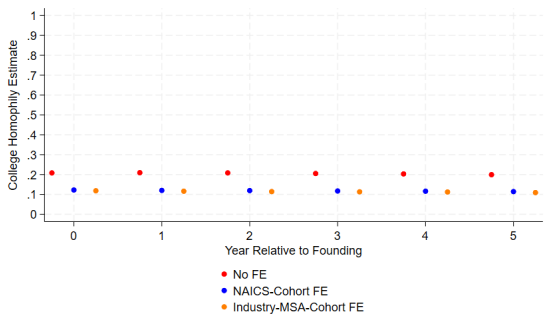
(c) Hispanic



(d) Asian



(e) Age



(f) College Educated

Figure 1: Estimates from Balanced Sample of New Firms

Notes: Each panel plots the coefficient $\hat{\beta}_t$ from Equation 1 using a different demographic variable. The three series correspond to specifications with: (i) no fixed effects; ii) six-digit NAICS-by-founding-year fixed effects; iii) two-digit industry-by-MSA-by-founding-year fixed effects. All standard errors are clustered at the firm level and are small relative to the point estimates, so we have omitted confidence intervals for legibility. Source: authors' calculations based on U.S. Census data (FSRDC Project Number 2229, Disclosure Review Board (DRB) approval numbers: CBDRB-FY25-P2229-R11972 and CBDRB-FY26-P2229-R12962).

We assess the robustness of the results in several ways. First, we test the robustness of our estimates to how many top paid employees we count as founders and report the results in Table A3. For gender homophily, the estimates become smaller when assuming that only the top paid individual is the founder (0.162) or that the top two are founders (0.277), but the point estimates remain quantitatively large. This could be driven by misclassification of early employees as co-founders (Hyatt et al., 2021) or attenuation bias from including too few founders (e.g., the one-founder measure is a noisy version of the founder composition if there are truly three founders).

We also report results using zip code instead of MSA in the very detailed fixed effects specification. We use within MSA (and industry and founding cohort) comparisons in the core results to capture the local labor market, following Kerr and Kerr (2016). However, zip code level analysis could help account for more granular labor markets. Figure A3 shows that the estimates become smaller across all demographic characteristics, but remain quantitatively large.

Our estimates are also stable across founding cohorts. Splitting the LEHD sample into three founding-year groups (1992–1997, 1998–2003, 2004–2009), we find minimal quantitative differences: some coefficients rise slightly (female and age) while others fall (Black, Hispanic, and college). Appendix Figure A4 reports the detailed cohort estimates.

II.B Homophily at Firms with High Growth Potential

The core LEHD analysis gives equal weight to all startups. Potentially, homophily could be minimal in startups with high growth potential, because the stakes are higher. To address this, we provide additional evidence focusing on these startups.

II.B.1 LEHD Weighting and High-Growth Results

First, we use weighted regressions to emphasize larger firms and higher-paid individuals. We compute weighted average employee demographics by weighting by employee earnings. Then, we weight the core regression by the square root of total employees to give more weight to larger firms. Table A4 provides estimates when applying both weighting schemes. The point estimates are again very similar to the core year-zero estimates: 0.360 for female, 0.544

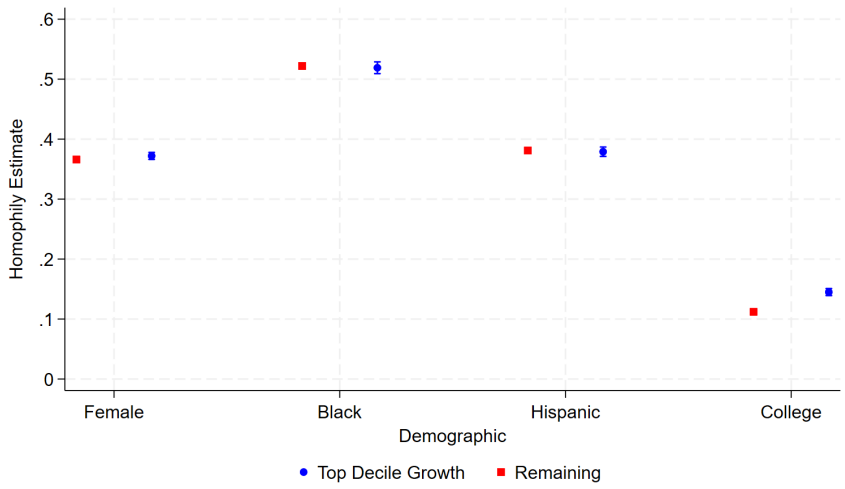


Figure 2: Homophily by Firm Employment Growth (LEHD)

Notes: Pooled homophily coefficients from the balanced LEHD sample (Equation 1), separately for firms in the top decile of employment growth from year 0 to year 5 and the remaining 90 percent. All specifications include industry-by-MSA-by-founding-year fixed effects. Bars show 95 percent confidence intervals. Source: authors' calculations based on U.S. Census data (FSRDC Project Number 2229, DRB approval numbers: CBDRB-FY25-P2229-R11972 and CBDRB-FY26-P2229-R12962).

for Black, and 0.377 for Hispanic.

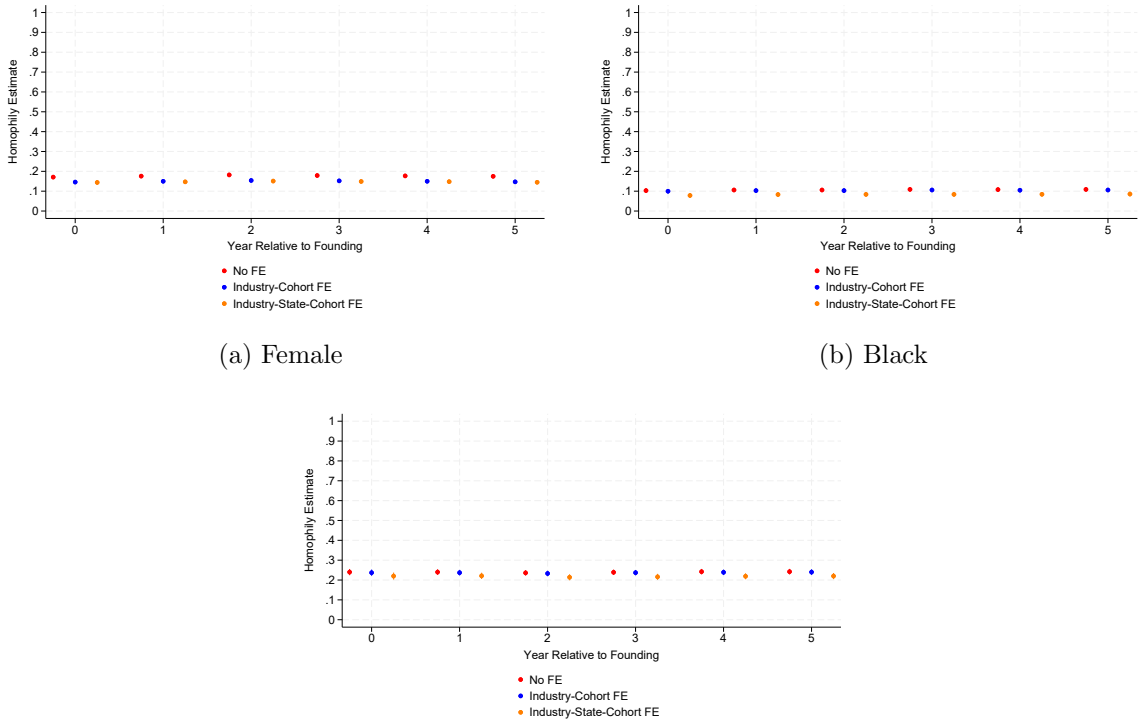
Next, we split firms by realized employment growth over the first six years in the data, separately estimating homophily within the fastest-growing ten percent of firms versus the remaining ninety percent. Figure 2 shows that faster-growing firms exhibit nearly identical homophily estimates to the remaining firms.

II.B.2 Evidence from Crunchbase–LinkedIn Sample

As discussed above, the Crunchbase–LinkedIn sample contains technology-focused and venture-backed startups. Figure 3 presents the core results using the balanced panel from the LinkedIn-Crunchbase linked dataset. We find quantitatively large differences in early employee composition across firms. In the year of founding, the homophily estimates we recover are 0.171 for female, 0.103 for Black, and 0.240 for Hispanic. Homophily persists over several years and remains quantitatively stable over time as startups add more employees.[17]

We again assess the robustness of the core results. First, we attempt to replicate the

[17]Appendix Figure A5 displays the regression intercept (projected share of female employees at startups with all-male founders). The intercept increases for female employees, from 0.31 in year 0 to 0.34 in year 5. For Black and Hispanic employees, the baseline share does not change in a quantitatively significant way.



(a) Female

(b) Black

(c) Hispanic

Figure 3: Estimates from Balanced Sample of New Firms (Crunchbase)

Notes: estimates of Equation 1 using the balanced Crunchbase sample. Each panel plots the coefficient for a different demographic variable. The three series correspond to specifications with: (i) no fixed effects; (ii) industry-by-founding-year fixed effects; (iii) industry-by-state-by-founding-year fixed effects. Standard errors are clustered at the firm level. Source: Crunchbase merged to LinkedIn (Revelio Labs).

industry, geography, and founding-year fixed effects within the Crunchbase sample. The challenge is that the Crunchbase data are not as detailed in terms of industry and location, so we use higher-level industry groups (48) and geography variables (state). The specifications we use thus include industry-by-founding-year FEs and industry-by-state-by-founding-year FEs. Figure 3 reports the estimates from the specification with the detailed fixed effects, alongside the estimates from the baseline specification. The estimates again become slightly smaller in magnitude once we include more detailed controls (0.144 for female, 0.079 for Blacks, and 0.220 for Hispanics in year 0), but remain very close to the baseline estimates.

Second, we provide results weighting by firm size (employment) to make sure that our results are not driven by smaller or slower-growing firms. We show results for two weighting schemes, using alternatively the square root of the number of employees or the total number of employees as weights. Tables A5 and A6 present the results. We find slightly smaller quantitative differences across the specifications relative to our core estimates. For example, the estimates in year 0 are 0.138, 0.107, and 0.130 for female, Black, and Hispanic in the regressions with linear weighting.

As with the LEHD sample, we re-estimate homophily within the fastest growing startups within the Crunchbase sample. Appendix Table A7 shows similar estimates when splitting the Crunchbase sample by employment growth.[18] The evidence suggests that homophily is not just a small firm phenomenon and persists even as firms grow and hiring processes become formalized.

II.B.3 Reconciling Crunchbase–LinkedIn Estimates and LEHD Estimates

Here, we discuss the sources driving the difference in homophily estimates. Appendix B.D provides detailed evidence.

The Crunchbase estimates are quantitatively smaller than those from the LEHD, and the gap is uneven across groups. Pooling years 0–5 with the most detailed available controls, the LEHD estimates are 0.364, 0.512, and 0.373 for female, Black, and Hispanic, versus 0.147, 0.083, and 0.218 in Crunchbase. For female and Hispanic founders, the Crunchbase estimates are of the same order, whereas the Black estimate is about six times larger in the LEHD.

[18] Appendix Table A8 provides a breakdown by firm size. We find slightly smaller coefficients for the largest quintile of firms by size but estimates that remain quantitatively large.

A leading issue is founder measurement. Our LEHD approach labels the top three earners as founders, whereas the title-based Crunchbase definition typically identifies between one and three founders, with an average of two. As noted earlier, our founder measurement for the LEHD data is likely overly inclusive. In contrast, the Crunchbase-LinkedIn data may miss founders who do not have LinkedIn profiles. Reweighting the LEHD estimates by the Crunchbase founder-count distribution closes about half of the female gap and two-thirds of the Hispanic gap, but only about a quarter of the Black gap.

A second issue is founder demographic measurement. As noted above, Revelio Labs uses name-based demographic classification. The large discrepancies for Black homophily is consistent with the fact that name-based race classification is particularly noisy for Black individuals (Berry et al., 2024; Cook et al., 2026). In general, the Crunchbase–LinkedIn estimates are likely to suffer from attenuation bias, which would further close the gap between estimates.

As a final note, we do not find evidence that geographic coverage is driving the difference. Restricting Crunchbase startups to the LEHD-covered states modestly raises female homophily, leaves Black homophily largely unchanged, and slightly lowers Hispanic homophily (Appendix B.D.3).

Overall, our findings suggest that homophily is quantitatively significant within high-growth-potential startups. Homophily may be smaller in the Crunchbase–LinkedIn sample, but much of the gap can be attributed to measurement issues. We account for a broader range of homophily estimates in our subsequent policy analysis.

II.C Persistence of Founder-Employee Homophily

Finally, we provide evidence that homophily persists as as firms age. Between year 0 and year 5, the estimates with the most detailed fixed effects decline only slightly for female (0.371 to 0.360), Asian (0.574 to 0.558), and college-educated founders (0.118 to 0.109), and somewhat more for Black founders (0.556 to 0.498), Hispanic founders (0.406 to 0.359), and age (0.364 to 0.316). The year zero estimates are not very different than the estimates recovered when pooling across years 0–5 (0.364 for female, 0.512 for Black, and 0.373 for

Hispanic founders).[19]

To provide more evidence on long-run homophily patterns, we make use of the unbalanced panels in each setting. First, we re-run our analysis on an unbalanced panel of new firms from the LEHD and report the results from the most detailed fixed-effects specification in Figure A8, obtaining estimates for up to 9 years after founding. We find similar initial effects (e.g., the female estimate is 0.381 in year 0 of the unbalanced panel, 0.566 for Black and 0.395 for Hispanic). We also find that the effects persist strongly out to the tenth year of the firm, albeit slightly attenuated (e.g., 0.250 in year 9 versus 0.381 in year 0 for female and 0.433 versus 0.566 for Black). Because the results use an unbalanced panel, the decay will reflect both within-firm changes and differential attrition across types of organizations.

Second, we also provide longer-run results using an unbalanced sample of firms from the LinkedIn-Crunchbase linked dataset, reporting estimates for up to 20 years after founding. We include the most detailed version of the fixed effects. Figure A9 shows quantitatively similar estimates to those in the core sample (0.149, 0.085, 0.216 in year 0 for female, Black, and Hispanic, respectively). The effects again persist strongly over time, and the estimates for year 9 are very similar to the estimates for year 0 across all three demographics. Even out to twenty years, the estimates are still quantitatively large (0.135, 0.062, 0.174). We note that, over this time span, the shares of female, Black, and Hispanic workers all increase gradually, but do so in all new firms (e.g., 0.303 to 0.389 for the constant term in the female regressions).

III Heterogeneity and Mechanisms

In this section, we provide evidence on the mechanisms behind the homophily patterns, with a broader goal of assessing whether homophily is a general feature across settings. We first study which workers and which firms drive the estimates by analyzing worker- and firm-side heterogeneity. We then study the dynamics of homophily, asking whether persistence reflects retention or new hiring and whether the founder or the initial team drives it. Finally, we test whether observable production characteristics correlated with founder demographics can

[19] Appendix Table A3 (Panel C) reports the pooled estimates for all demographic characteristics. Appendix Figure A2 also displays the intercepts, which remain stable between year 0 and year 5 within this sample.

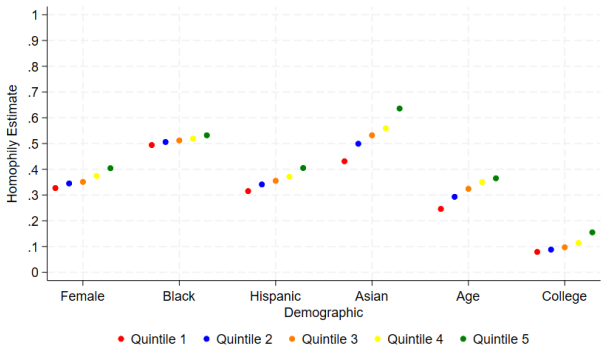


Figure 4: Homophily by Worker Wage Quintile

Notes: estimates of Equation 1 using the balanced LEHD sample and reporting results separately for each demographic and wage quintile (1 is lowest paid, 5 is highest). Workers are sorted into quintiles based on within-firm pay distribution in each year, and average demographics are computed within each quintile for each firm-year cell. All standard errors are small relative to the point estimates; confidence intervals are omitted for legibility. Source: authors' calculations based on U.S. Census data (FSRDC Project Number 2229, Disclosure Review Board (DRB) approval number: CBDRB-FY25-P2229-R11972).

account for the estimates.

III.A Worker-Side Heterogeneity

We first examine heterogeneity across worker types, starting with results by income quintile from the LEHD before moving to seniority and occupation in the LinkedIn data. The last two variables are unavailable in the Census LEHD data. Overall, the evidence suggests that homophily is not driven by particular types of workers.

First, we sort employees into quintiles within each firm's pay distribution in any year, and compute average demographics within each quintile for each firm and firm age cell. This then allows us to estimate homophily coefficients by quintile. Figure 4 reports the results for each demographic and worker wage quintile, with 1 being the lowest paid quintile of workers and 5 the highest. We find large effects across the board, but, for each demographic, we find larger point estimates when analyzing the highest paid workers.

Switching to the Crunchbase sample, we present results related to seniority. Based on role name, Revelio provides an estimate of the seniority of a given role on a scale from 1 (entry-level) to 7 (C-suite). We split the sample of workers into senior (seniority greater than or equal to 4) and junior, and run separate regressions based on seniority. In these regressions, we pool the six years of each firm and cluster standard errors at the firm level.

We find similar point estimates across the two groups for each demographic (Table A14).

Finally, we present results, splitting by broad occupation categories based on Revelio's processing of raw job titles. The seven broad occupation categories are: administrator, engineer, finance, marketing, operations, sales, and scientist. Table A15 presents the results. We find large effects across all occupation categories for gender and race.

III.B Firm-Side Heterogeneity

Here, we use industry information to assess heterogeneity by firm characteristics. We start with the Census sample and present results for service versus non-service industries. We split the sample based on two-digit NAICS codes: service (51 and above); non-service (49 and below). Potentially, the nature of production or management might generate differences across these two groups. Again, the body of evidence points to homophily as being a broad-based phenomenon.

Appendix Figure A10 shows very similar coefficients across service and non-service firms along each demographic dimension. We also use high-level industry groups from the Crunchbase sample to assess heterogeneity across industry groups. Appendix Table A13 shows similar qualitative patterns across four large industry groups (software, hardware, manufacturing, and health care).

In addition, we also estimate the core homophily regressions (with MSA-by-founding-year fixed effects) separately for each six-digit NAICS industry. Table A16 reports the distribution across all industries with any variation and for the top quartile of industries in terms of new firms. The latter sample likely produces more reliable estimates, especially with the very detailed fixed effects.

III.C Dynamics: Founding Team and Retention

Here, we break down the dynamics behind the persistence documented in Section II.C in two ways. We focus on whether founder demographics or initial employees are more predictive of employee demographics over time. This distinction informs our subsequent modeling. If founder influence faded as firms grew, then a model that takes into account early employees would be needed. If founder influence persists, then it can justify a model in which firms

have a fixed production function that depends on founder characteristics.

Figure 5 reports the estimates from a horserace regression incorporating both founder characteristics and initial employee characteristics in the founding year, using the balanced LEHD sample. We find that initial worker demographic composition is strongly predictive of subsequent worker demographics, which is unsurprising given that some initial employees will remain at the firm, but the coefficient declines as firms age. In contrast, the founder coefficient rises initially and remains positive as firms age, suggesting a persistent association between founder characteristics and later employee composition after accounting for the initial team.[20]

As an additional breakdown, we find that persistence is sustained partly through retention rather than new hiring. Homophily among new hires decays substantially faster than overall homophily as firms age, even though both remain quantitatively large (Appendix Figure A12). This suggests that higher retention of workers from similar backgrounds is partly driving the observed persistence.

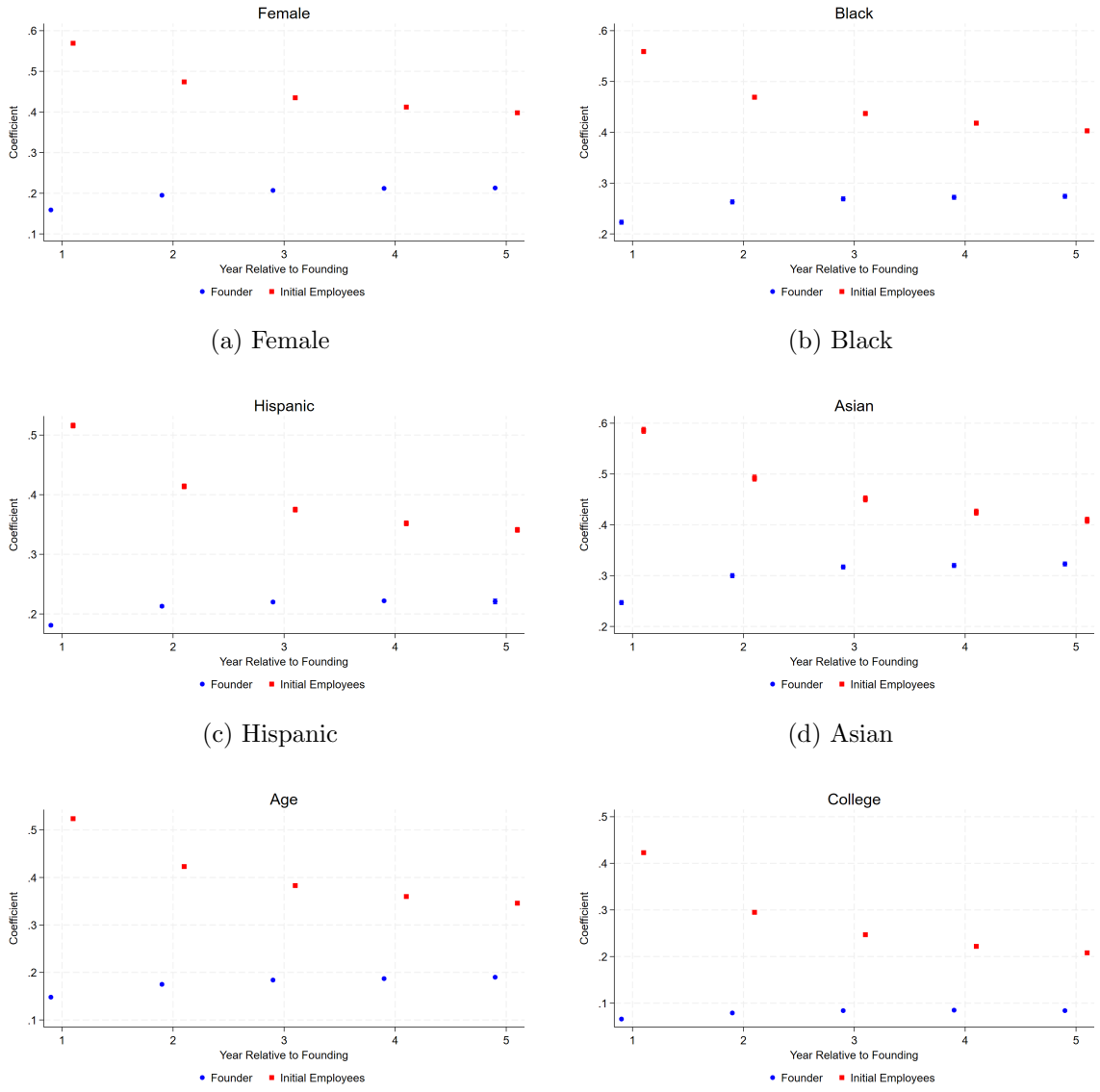
III.D Production Structure

Finally, we study whether observable firm characteristics correlated with founder demographics, rather than founder demographics per se, can account for some of the homophily estimates.

We use Census data to measure the production structure of new firms. We test whether founders from different backgrounds create firms with different average wages, labor share, and trading behavior. To do so, we focus on the 2007 and 2012 Economic Census, the last two during our sample period, and compute statistics for each new firm in the first Economic Census they are surveyed, restricted to firms within 5 years of founding. We again draw comparisons within detailed industry-by-MSA-by-founding-year groups.

We find significant differences in terms of average wage: female, Black, and Hispanic founded firms pay significantly lower average wages ($15k, $11k, and $6.5k, respectively), whereas college-educated founders pay more ($19.8k). We find smaller differences in terms of

[20]We also report corresponding estimates from the Crunchbase--LinkedIn sample, where job titles allow more direct identification of founders (Appendix Figure A11). The qualitative patterns are similar.



(a) Female (b) Black

(c) Hispanic (d) Asian

(e) Age (f) College Educated

Figure 5: Horserace Estimates Using the LEHD Sample

Notes: estimates of Equation 1 using the balanced LEHD sample, with an additional variable measuring the initial employee composition. All standard errors are small relative to the point estimates; confidence intervals are omitted for legibility. Source: authors' calculations based on U.S. Census data (FSRDC Project Number 2229, Disclosure Review Board (DRB) approval number: CBDRB-FY26-P2229-R12962).

labor share and trading intensity, although some significant differences are present. Appendix Table A17 provides detailed estimates.

We then control for these characteristics and re-run the homophily analysis. Table A18 presents the estimates. Because the sample is different to the main balanced sample (restricted to startups founded between 2002 and 2012 and to firms with wage and labor share information), we report specifications with and without controls. We find very minimal differences in homophily when controlling and not controlling for production structure characteristics, other than for college-founded startups. This suggests that the set of production structure measures available to us are not driving the observed homophily coefficients.

These results support simplifying the model to focus on the relative productivity of different worker demographics based on similarity to founder and hold other aspects of firms' production technologies fixed.

IV Wage Analysis

Given our core results, we turn to assessing whether the differences in employment composition are driven by labor supply or labor demand factors. Estimating the relative importance of labor supply and labor demand in driving founder-employee homophily is key to understand the wage impacts of diversifying access to entrepreneurial career, which we return to in the model section below.

IV.A Methodological Approach

Our approach is based on a simple prices and quantities argument. If founders are more productive with workers who share their background, the relative demand curve for own-group labor is shifted out at own-group firms, raising *both* the relative employment and the relative wage of those workers. If instead own-group workers accept lower pay for the amenity of working with similar founders, the relative supply curve is shifted out, raising relative employment but *lowering* the relative wage. The sign of the own-group wage premium therefore distinguishes the two mechanisms. The AKM decomposition below implements this test at the firm level.

To do so, we make use of wage data from the LEHD and employ a two-step approach.

First, we estimate regressions in the style of Abowd et al. (1999) (hereafter, AKM) across all employees and firms, except founders at new firms. Because of computational limitations, we do so state-by-state. The difference from a standard approach is that we compute firm and individual fixed effects for individuals from each demographic group (and the complementary group of individuals), in the style of Card et al. (2016). Formally, we run the following regression for each state s:

$$\log(Wage_{i(ds)t}) = \alpha_i + \psi_{jsd} + \lambda \cdot X + \epsilon_{it} \tag{2}$$

where the outcome is the log wage of individual i in year t, who belongs to demographic group d and works in state s; ψ_{jsd} denotes firm j's fixed effect for demographic group d estimated from state s data, and α_i the individual fixed effect. This allows us to collect a set of estimates at the firm-demographic-state level. As noted above, for each demographic group d, we also compute the firm fixed effect for the complementary set of workers (d'). The additional controls include age, age squared, and tenure at the current firm.

For the second step, we compute the difference in firm FEs for each demographic group d (compared to the complement d') and regress the difference against founder fraction from that group. Unlike the first step, we now focus on new firms (using the same sample of firms as the unbalanced LEHD sample). Formally, we run the following regression:

$$\psi_{jsd} - \psi_{jsd'} = \beta_0 + \beta_1 \times FounderFraction_{jd} + \nu_{jds} \tag{3}$$

where d is a specific demographic group. We cluster standard errors at the firm level to account for firms having establishments in multiple states.

IV.B Estimates

Table 3 presents the estimates from Equation 3. Across demographic groups, we find sizable *positive* coefficients, ranging from 2.9 to 12.9 log points (the dependent variable is the difference in log-wage firm fixed effects, so coefficients are in log points throughout this section), with the largest effects for female workers. This means that workers who share a demographic with the founders earn wages that are higher than expected relative to their pay at other firms, relative to the same differential for individuals who do not share similar

Table 3: AKM Decomposition Regressions

	Firm Fixed-Effect Difference				
Female Founder Fraction	0.129***				
	(0.002)				
Black Founder Fraction		0.068***			
		(0.004)			
Hispanic Founder Fraction			0.057***		
			(0.002)		
Asian Founder Fraction				0.029***	
				(0.003)	
College Grad Founder Fraction					0.049***
					(0.005)
Constant	0.062***	0.117***	-0.316***	0.128***	-0.140***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
Observations	1,775,000	791,000	1,802,000	1,104,000	584,000

Notes: The dependent variable is the difference in firm fixed effects between the indicated demographic group and its complement $\hat{\psi}_{jsd} - \hat{\psi}_{jsd'}$. Firm fixed effects are estimated separately by state and demographic group following Equation 2, controlling for age, age squared, and firm tenure. Standard errors are in parentheses and are clustered at the firm level. $^*p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Source: authors' calculations based on U.S. Census data (FSRDC Project Number 2229, Disclosure Review Board (DRB) approval number: CBDRB-FY26-P2229-R12962).

socio-demographic characteristics. These results show that founder-employee homophily is driven by labor demand effects rather than labor supply effects.[21]

For context, we note that the premium we estimate here is relative rather than absolute. Table A20 presents estimates from regressing absolute firm effects for a given group against that group's share of the founders. Across the demographics, own-group workers have lower firm effects at own-group-founded firms (e.g., -0.256 for women at female-founded firms). This is consistent with the lower average wages paid by firms with female, Black, and Hispanic founders documented in Section III. The relative premium is still the more relevant estimate for our test, and male workers at female startups are also paid significantly less than those at male-founded startups.

Our analysis has a few limitations. First, the LEHD data do not contain information about non-wage compensation (e.g., options). However, differential equity compensation is unlikely to be driving our estimates. Based on survey evidence, only 1.7 percent of private-industry employees received stock option grants in 1999 (U.S. Bureau of Labor Statistics,

[21] Our finding that demand dominates stands in contrast to Campero and Kacperczyk (2020), who find that labor supply effects dominate using job application data from high-tech startups. The discrepancy could reflect differences in sample, or the fact that they observe only applications and interviews rather than realized wages.

2000; Schildkraut, 2004). Second, some firms do not hire any individuals from certain backgrounds, so we cannot compute the group-d firm effect and must drop them from the corresponding analysis. However, our estimates are likely to understate the underlying difference across the full sample of firms.[22] Finally, due to computational constraints, we run the model in a state-by-state manner. This excludes across-state movers, which reduces precision and may have differing impacts by demographic group (e.g., men and high-skilled workers are more geographically mobile). As long as the noise introduced by the computational choice is uncorrelated with founder demographics, then it would not bias our estimates.

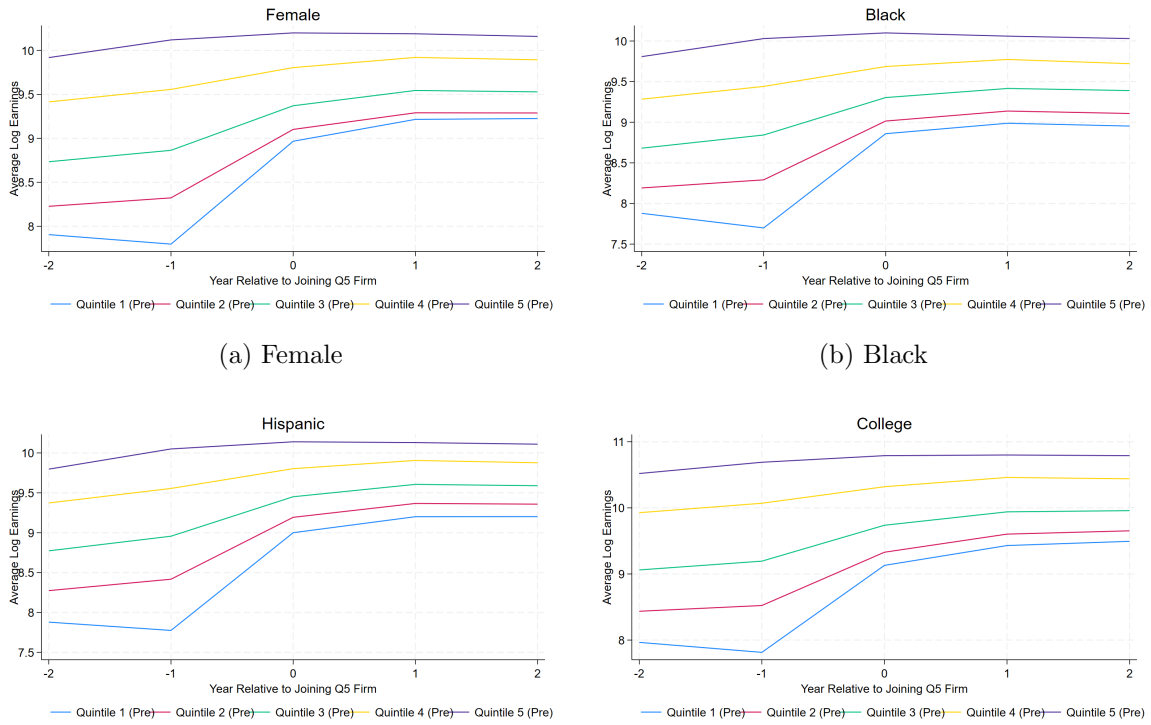
IV.C Robustness Checks

Appendix B.J presents additional evidence supporting the validity of the AKM-by-group approach. First, we show estimates using firm effects estimated using a wage regression without any controls. The premia are larger without controls (0.197, 0.157, 0.101, 0.091, and 0.093 for female, Black, Hispanic, Asian, and college-educated founders, versus 0.129, 0.068, 0.057, 0.029, and 0.049 with controls). This suggests that the controls in our core analysis are not manufacturing significant estimates that are not present under simpler specifications, and that age and tenure matter in estimating firm fixed effects.

Second, we follow Bonhomme et al. (2019) and Lerner et al. (2024) in allowing for non-linear match effects. We sort firms into ten bins and estimate an assortative matching parameter for each bin. Table A21, Panel A reports model fit statistics, which suggest that the two versions of the model have similar fit. In addition, we also find quantitatively small differences in the assortative matching slope between new firms and the remaining firms in the sample (Table A21, Panel B). Both pieces of evidence suggest that matching is unlikely to be driving our core estimates.

Third, we also validate the firm-by-demographics estimates, by showing descriptive wage trends in the style of Card et al. (2013). Figure 6 presents the results for movers to quintile five (highest firm effect) firms, breaking movers down by their pre-move firm's quintile.

[22] We are attempting to estimate the relative difference across all firms in the sample. Our actual sample consists of firms that have realized hires across groups. If this selection is driven by firms paying relatively equally across worker groups, our estimates would be understating the true difference in the overall population of firms.



(a) Female

(b) Black

(c) Hispanic

(d) College

Figure 6: Descriptive Evidence for Movers Moving to Quintile Five Firms

Notes: descriptive event studies showing average log earnings for different mover groups and different demographics. Source: authors' calculations based on U.S. Census data (FSRDC Project Number 2229, Disclosure Review Board (DRB) approval number: CBDRB-FY26-P2229-R12962).

We find minimal pre-trends, and larger jumps for individuals moving from lower quintile firms, which supports the exogenous mobility assumption underlying the group-specific firm effects.[23]

V A Model of Entrepreneurs with Heterogeneous Production Functions and Rent Sharing

Our evidence suggests that labor demand effects drive the observed employee composition differences and that founder effects persist over time. Using this as motivation, we build an equilibrium model where entrepreneurs create firms with heterogeneous production functions, calibrate the model using the summary stats and estimates presented above, and run counterfactuals to assess the impacts of reducing access barriers to entrepreneurship.

[23]Disclosure constraints prevent us from showing the full set of pairwise moves.

V.A Model

Our model follows the general structure of Chiplunkar and Goldberg (2024), hereafter CG, but simplifies away some aspects that are more important in a developing economy context (e.g., informal businesses). Our main modification is to put heterogeneity across founder types in the production function rather than in hiring costs, which helps us generate the own-group wage premium documented in Section IV.

The economy is populated by two demographic groups, majority (m) and minority (f), with population sizes N_m and N_f (only the ratio N_f/N_m matters for the calibration). Throughout, we use "minority" (f) as shorthand for the group of interest—the group underrepresented among founders—and "majority" (m) for the reference group, even in cases where group f is not a numerical minority of workers (as with non-college workers, who are the majority of the workforce but underrepresented among founders). Each individual draws ability z from a log-normal distribution and chooses whether to become an entrepreneur or supply labor at the market wage.[24]

Production. An entrepreneur of type $g \in \{m, f\}$ with ability z produces output using a composite labor bundle:

$$y_g(z) = z \cdot \ell_g(z)^{\rho}, \quad \rho \in (0, 1).$$

The composite labor bundle aggregates workers of both types via a CES technology,

$$\ell_g = \left[A_{gm}^{1/\gamma} \, \ell_{gm}^{\frac{\gamma-1}{\gamma}} + A_{gf}^{1/\gamma} \, \ell_{gf}^{\frac{\gamma-1}{\gamma}} \right]^{\frac{\gamma}{\gamma-1}},$$

where ℓ_{gh} is the quantity of type-h workers hired, $\gamma > 1$ is the elasticity of substitution between worker types, and the parameters $A_{gh} \geq 0$ govern each founder type's relative productivity from different worker types. Because A is indexed based on both entrepreneur and worker group, this setup allows entrepreneurs of backgrounds to create systematically different production functions. Higher values of A_{gg} relative to A_{gh} ($h \neq g$) capture in-group hiring bias (homophily): founders are able to elicit more productivity from workers from

[24]This follows the structure of the Chiplunkar and Goldberg (2024) model, but removes self-employment and non-participation as choices for individuals.

similar backgrounds (e.g., better management, different processes, etc.).[25]

Wage setting. Rather than assuming a standard competitive market (in which every worker of a given type earns a single wage and the productivity advantage A_{gg} would be dissipated entirely into *quantities*), we allow firms to have wage-setting (monopsony) power. Following Card et al. (2018), workers are assumed to have idiosyncratic preferences over jobs, which translates to each firm facing an upward-sloping labor supply curve for each worker type, $\ell_{gh} = D_h\, w_{gh}^{\varepsilon}$, where D_h is a market-level shifter and $\varepsilon > 0$ is the firm labor-supply elasticity.

We impose a common firm labor-supply elasticity ε across worker types within each two-group calibration, so that the group-specific firm-effect differential we target (β_1) is attributed to differences in production weights A_{gh} rather than to differential labor-supply elasticities across groups. The firm then sets a constant markdown on the marginal revenue product of each worker type,

$$w_{gh} = \frac{\varepsilon}{1+\varepsilon}\, MRP_{gh}.$$

Because the marginal revenue product is higher for workers the founder is relatively more productive with (higher A_{gh}), own-group workers command a higher wage at own-group firms. This is the source of the firm-level wage premia (the group-specific AKM firm effect) documented in Section IV.

Entry into entrepreneurship. Entry is governed by two conditions. First, the individual must have access to entrepreneurial opportunities: a fraction $\phi_g \in (0, 1]$ of each group is exposed, where $\phi_m = 1$ (the majority group is normalized to face no access frictions). We model access frictions as a random exposure shock that is independent of ability z within each group. Second, conditional on access, individuals become entrepreneurs when profits exceed their employment outside option. We measure this outside option by W_g, the employment-weighted average wage of workers in group g. The entry condition is then

$$\pi_g(z) \geq W_g,$$

[25]As noted above, our productivity-based model of heterogeneity stands in contrast to the search cost approach in CG. In addition, we allow for monopsony power, whereas CG keep to group-specific market clearing wages. If one assumes that search costs reflect difficulty in attracting opposite-group workers, then combining the CG model with monopsony would generate an opposite-group wage premium, which runs counter to our empirical estimates.

where $\pi_g(z)$ is the monopsony profit derived in Appendix C. Profit is increasing in z, so this implies an endogenous ability cutoff z_g^*; individuals with $z < z_g^*$ or without access supply labor as workers. A worker's wage depends only on their type h, not on their ability z, which matters solely for the entrepreneurship decision.

Equilibrium. Given the markdown wages, an equilibrium is a set of market-level aggregates, the labor-supply shifters (D_m, D_f) and worker values (W_m, W_f), such that (i) each worker market clears, with total labor demanded by all active firms equal to the supply of that type (population net of entrepreneurs), and (ii) the worker values are consistent with the realized distribution of wages. The firm-level demands, wages, and the cutoffs z_g^* follow from the firm's markdown problem and the entry condition, as detailed in Appendix C.

V.B Calibration and Counterfactuals

We calibrate the model in three steps. First, we impose two normalizations and set two structural parameters externally. Second, we estimate the remaining five parameters using seven empirical moments. Finally, we remove the minority group's access friction and recompute the equilibrium. Table 4 reports the calibrated parameters and counterfactual outcomes for each group.

As a first step, we normalize $A_{ff} = 1$. On top of this, we also impose $A_{fm} = A_{mm} + A_{mf} - A_{ff}$, so that the two founder types have the same sum of productivity weights. This restriction aims to parsimoniously capture relative productivity differences across founder types without allowing for a significant shift in absolute productivity across the founder groups.[26]

We set $\rho = 0.725$ and $\gamma = 4$, based on estimates in the literature.[27] The first parameter governs decreasing returns to effective labor, and the second is the elasticity of substitution between majority and minority workers. We take population shares N_m and N_f from Table

[26]The equal row sum restriction does translate to equal effective labor costs across founder types in the case when wages are equal. This is not the case for our baseline equilibrium, because of the access friction.

[27]For ρ, Hopenhayn and Rogerson (1993) recover 0.64 using a U.S. labor-only firm-dynamics model. At the higher end, the effective labor curvature implied by the capital-labor model in Atkeson and Kehoe (2005) is approximately 0.81 if capital adjusts freely at a fixed rental rate. We use the midpoint of these two estimates. Acemoglu et al. (2004) find U.S. gender elasticities of 3.2–4.2 in their main specifications, and Bils et al. (2024) find a long-run lower bound of about 4 for substitution across education groups. We discuss sensitivity to parameter choice below.

1, Panel B. We solve the model by simulating ten million majority individuals per group and N_f/N_m times as many minority individuals.

Ability distribution and the entry margin. We assume $\log z \sim N(\mu_z, \sigma_z^2)$. Because z enters output $z\,\ell^\rho$ as a Hicks-neutral shifter, changing μ_z rescales output, wages, and the value of entrepreneurship and work by a common factor. It therefore leaves the entry cutoffs and entrepreneurship rate unchanged. We set $\mu_z = 0$. σ_z governs the dispersion of firm size and affects how sharply entry responds when the access friction ϕ_f is relaxed.

We discipline σ_z by adding both the mean and standard deviation of log firm employment. Both moments are necessary, because dispersion among active firms depends on the entry cutoff.

We compute both moments using the new firms in our analysis sample, which is also the sample used for the other calibration targets. Under log-normality, the (pseudo-)median employment of 10.0 and mean employment of 29.06 in Table 1, Panel B, imply a mean of log employment of $\ln 10 = 2.30$ and a standard deviation of $\sqrt{2\ln(29.06/10)} = 1.46$.[28]

Entry by group g equals ϕ_g times the mass above the ability cutoff z_g^*. Given σ_z and the population ratio N_f/N_m, market clearing pins the majority entry rate under the normalization $\phi_m = 1$. The targeted minority founder share then disciplines ϕ_f.

Parameters and Moments. After these normalizations and parametrizations, the model has five free parameters: A_{mm}, A_{mf}, ϕ_f, the firm labor-supply elasticity ε, and the dispersion of log ability σ_z. We calibrate them to seven empirical moments: the minority worker share at minority-founded firms, the minority worker share at majority-founded firms, the minority founder share, the minority-to-majority wage ratio, the AKM firm-effect differential β_1 from Table 3, and the mean and standard deviation of log firm employment.

Calibration fit. For each demographic group, we target the seven moments above using the five parameters described above.[29] Appendix Table A22 compares the fitted moments with their targets. The model closely matches the five group-specific moments, with some

[28] The median and mean of a lognormal are e^{μ} and $e^{\mu+\sigma^2/2}$, so $\mu = \ln(\text{median})$ and $\sigma = \sqrt{2\ln(\text{mean}/\text{median})}$.

[29] For example, the female calibration targets the female employee shares at female-founded firms (0.717) and at male-founded firms (0.353), the female founder share (0.408), the raw female-to-male wage ratio (0.80, from the U.S. Bureau of Labor Statistics), the AKM firm-effect differential (0.129), and the mean and standard deviation of log firm employment (2.30 and 1.46). Appendix Table A22 reports the full set of moments.

Table 4: Model Calibration and Counterfactuals

	Calibration								Counterfactuals	
Group	A_{mm}	A_{mf}	A_{fm}	A_{ff}	ϕ	ε	σ_z	W_f/W_m	W'_f/W'_m	ΔRelative Wages
Female	1.41	0.25	0.66	1.00	0.54	12.8	1.33	0.81	0.88	7.83pp
Black	2.90	0.04	1.94	1.00	0.50	48.9	1.17	0.73	0.81	8.14pp
Hispanic	4.30	0.12	3.42	1.00	0.59	38.0	1.29	0.73	0.78	5.08pp
Non-college	2.22	0.46	1.69	1.00	0.47	17.6	1.25	0.60	0.62	2.21pp

Notes: Each row calibrates the model separately for the indicated group (f) versus all other individuals (m). A_{gh} are the CES productivity weights governing demand composition across worker types. ϕ is the share of the minority group with access to entrepreneurship ($\phi_m = 1$ by normalization). ε is the firm labor-supply elasticity, which governs the wage markdown $\varepsilon/(1+\varepsilon)$. σ_z is the calibrated dispersion of log ability, targeted jointly with the entry margin to the mean and standard deviation of log firm employment. W_f/W_m is the calibrated equilibrium ratio of worker values (the employment-weighted average wage W_h of each group). Counterfactual columns report the equilibrium when $\phi_f = 1$ (equal access). Parameters: $\rho = 0.725$, $\gamma = 4$, $A_{ff} = 1$ (normalization), $A_{fm} = A_{mm} + A_{mf} - A_{ff}$ (row-sum normalization). Moments are simulated on a population of ten million majority individuals per group, with N_f/N_m times as many minority individuals.

discrepancies in the employee composition moments for Black and non-college workers. The main systemtatic misfit across the runs is in the two firm-size moments. The fitted standard deviation of log firm employment ranges from 1.25 to 1.33, compared with a target of 1.46. The calibrated ability dispersion is also similar across the four comparisons, with σ_z ranging from 1.17 to 1.33.

Sensitivity of parameter estimates to moments. Using the approach in Andrews et al. (2017), we analyze the local sensitivity of the jointly estimated parameters to small changes in each moment. To do so, we compute approximate percentage changes in each parameter for a 1% increase in each target moment. The production weights are generally most sensitive to the homophily and wage ratio moments. The access parameter (ϕ) responds strongly to the founder share and the wage ratio. Labor-supply elasticity (ε) is sensitive to the AKM differential and homophily moments. Ability dispersion (σ_z) responds primarily to the mean and standard deviation of log firm employment. Appendix Section C.C presents detailed results.

Counterfactuals. For our main counterfactual, we remove access barriers by setting $\phi_f = 1$ and recompute the equilibrium for each demographic comparison. Relative wages increase by 7.83pp for women, 8.14pp for Black workers, 5.08pp for Hispanic workers, and 2.21pp for non-college workers. These are quantitatively significant increases relative to the

respective initial wage gaps.

Two channels generate these gains. First, expanding minority entrepreneurship raises relative demand for minority workers because minority-founded firms employ them more productively. Second, wages under monopsony reflect match-specific productivity, so more matches between workers and entrepreneurs from the same group also increases realized wages.

V.C Robustness and Interpretation

The estimates above are based on a specific set of parameters and moments. Here, we provide a range of counterfactual estimates based on the literature and our estimates.

First, we use both the one-founder and three-founder Census homophily estimates to capture uncertainty about founding team measurement. The one-founder estimates are also broadly similar to the estimates from the Crunchbase–LinkedIn sample, outside of the estimates for Black founders.[30] We interact these two values with $\rho \in \{0.65, 0.70, 0.75, 0.80\}$ and $\gamma \in \{2, 3.5, 5\}$, based on other values used in the literature.[31]

Across the different parameter and moment combinations, we find counterfactual relative wage gains of 4–12pp for women, 2–13pp for Black workers, 3–17pp for Hispanic workers, and 1–4pp for non-college workers. Our preferred estimates sit in the middle of these ranges. The ranges reflect three forces. Smaller homophily coefficients reduce the wage effects of expanding minority entrepreneurship. Smaller substitution elasticity amplifies the projected wage impacts whereas larger elasticities dampen them. Variation in the returns-to-scale parameter, ρ, generally has a smaller effect in the available comparisons.

We note two issues in interpreting the output from this exercise. First, the homophily and β_1 moments are measured among *new* firms, while the wage-ratio moment is the *economy-wide* raw gap (e.g., 0.80 for gender, from the U.S. Bureau of Labor Statistics). If forces unrelated to founder-side demand account for part of the raw gap, the calibration may

[30] The Census homophily coefficients based on one founder are 0.162 for women and 0.194 for Hispanic workers. The corresponding detailed-control Crunchbase estimates are 0.147 and 0.218, respectively. As discussed above, measurement of Black founders is likely to be more reliable in the Census data (which draws on survey and administrative data) than the Crunchbase data (name-based classification).

[31] The range for ρ covers the range of U.S. estimates described above. For γ, the lower endpoint is close to the value used by Chiplunkar and Goldberg (2024) and the upper endpoint extends beyond the preferred value.

attribute too much of that gap to the productivity matrix A and the access friction. This would tend to increase the estimated demand-side effect. However, the sensitivity exercises show that other plausible inputs and parameter values can move the estimates in either direction.

Second, the counterfactual is a *steady state* in which firms drawn under $\phi_f = 1$ replace the entire stock of firms. Young firms account for a small share of aggregate employment at a given time, so the transition would be gradual. The estimates should therefore be interpreted as long-run effects. At shorter horizons, the effects would depend on the rate of firm entry and the employment share of young firms.

VI Conclusion

In this article, we have provided systematic evidence of large and persistent differences in the employee composition across similar new firms with different founder backgrounds. In addition, we have provided evidence that founders pay employees from similar backgrounds more than expected, showing that labor demand dominates labor supply in driving the relative equilibrium employment patterns. To capture these patterns, we build a quantitative model of entrepreneurship where founders create heterogeneous production functions and share rents with workers; we use it to assess the implications of removing access barriers to entrepreneurship for specific demographic groups. We find quantitatively large effects.

The combination of our empirical results and quantitative framework can thus help evaluate the gains from increased participation in entrepreneurship by individuals from underrepresented groups. There are several policies and initiatives that aim to increase access to entrepreneurship. The Small Business Innovation Research program already aims to promote entrepreneurship from women, underrepresented minorities, and individuals from "HUBZones," historically underutilized business zones. Other policy initiatives include the SUCCESS Act of 2018 that directed the USPTO to make recommendations for promoting the participation of women, minorities, and veterans in entrepreneurship and innovation. In addition, venture capital funds have recently increased their hiring of Black and female venture capital partners, which research has shown to increase funding for female and Black entrepreneurs (Calder-Wang et al., 2021; Cook et al., 2026). The evidence presented adds to

our understanding of the distributional effects of policies aimed at broadening access to innovation and entrepreneurship careers. For example, facilitating careers in entrepreneurship for women might provide an important avenue to reduce the gender wage gap.

References

Abowd, John M., Francis Kramarz, and David N. Margolis, "High wage workers and high wage firms," *Econometrica*, 1999, *67*, 251–333.

Acemoglu, Daron, David H. Autor, and David Lyle, "Women, War, and Wages: The Effect of Female Labor Supply on the Wage Structure at Midcentury," *Journal of Political Economy*, 2004, *112* (3), 497–551.

Andrews, Isaiah, Matthew Gentzkow, and Jesse M. Shapiro, "Measuring the Sensitivity of Parameter Estimates to Estimation Moments," *The Quarterly Journal of Economics*, 2017, *132* (4), 1553–1592.

Ashraf, Nava, Alexia Delfino, Edward L. Glaeser, and Irene Solmone, "The Inverted U-Shaped Relationship between Female Entrepreneurship and Economic Development," *AEA Papers and Proceedings*, 5 2025, *115*, 500–508.

Atkeson, Andrew and Patrick J. Kehoe, "Modeling and Measuring Organization Capital," *Journal of Political Economy*, 2005, *113* (5), 1026–1053.

Balsmeier, Benjamin, Lee Fleming, Matt Marx, and Seungryul Ryan Shin, "Startups, Unicorns, and the Local Influx of Inventors," *Review of Economics and Statistics*, 2025.

Bell, Alexander, Raj Chetty, Xavier Jaravel, Neviana Petkova, and John Van Reenen, "Who Becomes an Inventor in America? The Importance of Exposure to Innovation," *The Quarterly Journal of Economics*, 2019, *134*, 647–713.

Bento, Pedro, "Female entrepreneurship in the U.S. 1982–2012: Implications for welfare and aggregate output," *Journal of Monetary Economics*, 1 2024.

_ **and Sunju Hwang**, "Barriers to black entrepreneurship: Implications for welfare and aggregate output over time," *Journal of Monetary Economics*, 3 2023, *134*, 16–34.

Berry, Alexander, Elizabeth M. Maloney, and David Neumark, "The Missing Link? Using LinkedIn Data to Measure Race, Ethnic, and Gender Differences in Employment

Outcomes at Individual Companies," Working Paper 32294, National Bureau of Economic Research July 2024. Originally issued March 2024; revised July 2024.

Bessen, James, Felix Poege, and Ronja Röttger, "Competing for Talent: Large Firms and Startup Growth," 2023.

Bils, Mark, Barış Kaymak, and Kai-Jie Wu, "Labor Substitutability among Schooling Groups," *American Economic Journal: Macroeconomics*, 2024, *16* (4), 1–34.

Bonhomme, Stéphane, Thibaut Lamadon, and Elena Manresa, "A Distributional Framework for Matched Employer Employee Data," *Econometrica*, 2019, *87*, 699–739.

Bryan, Kevin A, Mitchell Hoffman, and Amir Sariri, "Information Frictions and Employee Sorting Between Startups," 2022.

Calder-Wang, Sophie, Paul Gompers, and Patrick Sweeney, "Venture Capital's "Me Too" Moment," Technical Report 2021.

Campero, Santiago and Aleksandra Kacperczyk, "Asymmetric Gender Homophily in the Startup Labor Market," *Advances in Strategic Management*, 2020, *41*, 329–359.

Card, David, Ana Rute Cardoso, and Patrick Kline, "Bargaining, sorting, and the gender wage gap: Quantifying the impact of firms on the relative pay of women," *Quarterly Journal of Economics*, 5 2016, *131*, 633–686.

_ , _ , Jörg Heining, and Patrick Kline, "Firms and Labor Market Inequality: Evidence and Some Theory," *Journal of Labor Economics*, 2018, *36*, S13–S70.

_ , Jörg Heining, and Patrick Kline, "Workplace Heterogeneity and the Rise of West German Wage Inequality," *Quarterly Journal of Economics*, 2013, *128*.

Chen, Jun, Song Ma, and Feng Zhang, "Anti-Harassment Policy and the Startup Labor Market," Technical Report 2025.

Chetty, Raj, Nathaniel Hendren, Maggie R. Jones, and Sonya R. Porter, "Race and economic opportunity in the United States: An intergenerational perspective," 5 2020.

Chiplunkar, Gaurav and Pinelopi Koujianou Goldberg, "Aggregate Implications of Barriers to Female Entrepreneurship," *Econometrica*, 2024.

Choi, Joonkyu, Nathan Goldschlag, John Haltiwanger, and J Daniel Kim, "Early Joiners and Startup Performance," *The Review of Economics and Statistics*, 2023.

Cook, Lisa D., Matt Marx, and Emmanuel Yimfor, "Funding Black High-Growth Startups," *The Journal of Finance*, June 2026, *81* (3), 1619–1660.

Cortés, Patricia and Jessica Pan, "Children and the Remaining Gender Gaps in the Labor Market," *Journal of Economic Literature*, 2023, *61* (4), 1359–1409.

Currarini, Sergio, Matthew O. Jackson, and Paolo Pin, "An Economic Model of Friendship: Homophily, Minorities, and Segregation," *Econometrica*, 2009, *77* (4), 1003–1045.

Dorn, David, Florian Schoner, Moritz Seebacher, Lisa Simon, and Ludger Woessmann, "Multidimensional Skills on LinkedIn Profiles: Measuring Human Capital and the Gender Skill Gap," Discussion Paper 136/25, ROCKWOOL Foundation Berlin December 2025. December 2025 revision. Also circulated as CESifo Working Paper 11846.

Einio, Elias, Josh Feng, and Xavier Jaravel, "Social Push and the Direction of Innovation," Technical Report 2023.

Ewens, Michael and Richard R. Townsend, "Are Early Stage Investors Biased Against Women?," *Journal of Financial Economics*, 2020, *135* (3), 653–677.

Gentzkow, Matthew and Jesse M. Shapiro, "Ideological Segregation Online and Offline," *The Quarterly Journal of Economics*, 2011, *126* (4), 1799–1839.

Glaeser, Edward L. and Naomi Hausman, "The Spatial Mismatch between Innovation and Joblessness," in "Innovation Policy and the Economy," University of Chicago Press, 2020.

Goldin, Claudia and Cecilia Rouse, "Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians," *American Economic Review*, 2000, *90*, 715–741.

Gozen, Ruveyda, Richard Hornbeck, Anders Humlum, and Martin Rotemberg, "Historical Differences in Female-Owned Manufacturing Establishments: The United States, 1850–1880," 2024.

Guzman, Jorge and Aleksandra Kacperczyk, "Gender Gap in Entrepreneurship," 2018.

Haltiwanger, John, Ron S Jarmin, and Javier Miranda, "Who Creates Jobs? Small versus Large versus Young," *The Review of Economics and Statistics*, 2013, *95*, 347–361.

Hopenhayn, Hugo and Richard Rogerson, "Job Turnover and Policy Evaluation: A General Equilibrium Analysis," *Journal of Political Economy*, 1993, *101* (5), 915–938.

Hwang, Kylie and Melanie Wallskog, "The Role of Young Firms as Employers of Marginalized Workers," Technical Report 2026.

Hyatt, Henry, Seth Murray, and L. Kristin Sandusky, "Business Income Dynamics and Labor Market Fluidity," *IZA Journal of Labor Economics*, 1 2021, *10*.

Kerr, Sari Pekkala and William R Kerr, "Immigrant Entrepreneurship," in "Measuring entrepreneurial businesses: Current knowledge and challenges" 2016, pp. 187–249.

Kim, J. Daniel, "Is there a startup wage premium? Evidence from MIT graduates," *Research Policy*, 4 2018, *47*, 637–649.

Koning, Rembrand, Sampsa Samila, and John-Paul Ferguson, "Who do we invent for? Patents by women focus more on women's health, but few women get to invent," *Science*, 2021, *372*, 1345–1348.

Lerner, Josh, Henry Manley, Carolyn Stein, and Heidi Williams, "The Wandering Scholars: Understanding the Heterogeneity of University Commercialization," Technical Report 2024.

_ , Markus Lithell, and Gordon M Phillips, "Impact Investing and Worker Outcomes," 2025.

McPherson, Miller, Lynn Smith-Lovin, and James M. Cook, "Birds of a Feather: Homophily in Social Networks," *Annual Review of Sociology*, 2001, *27*, 415–444.

Miller, Conrad and Ian M Schmutte, "The Dynamics of Racial Co-Hiring," 2021.

Roach, Michael and Henry Sauermann, "Can Technology Startups Hire Talented Early Employees? Ability, Preferences, and Employee First Job Choice," *Management Science*, 8 2023.

Ronchi, Maddalena and Nina Smith, "Daddy's Girl: Daughters, Managerial Decisions, and Gender Inequality," 2024.

Schildkraut, Jeffrey L., "Stock Options: National Compensation Survey Update," *Compensation and Working Conditions*, 2004. U.S. Bureau of Labor Statistics. Posted June 30, 2004; revised September 29, 2004.

Sorenson, Olav, Michael S. Dahl, Rodrigo Canales, and M. Diane Burton, "Do startup employees earn more in the long run?," *Organization Science*, 7 2021, *32*, 587–604.

Tate, Geoffrey and Liu Yang, "Female leadership and gender equity: Evidence from plant closure," *Journal of Financial Economics*, 7 2015, *117*, 77–97.

U.S. Bureau of Labor Statistics, "Stock Options Most Likely for the Higher Paid," The Economics Daily October 2000. October 12. Reports the 1999 Pilot Survey on the Incidence of Stock Options in Private Industry.

U.S. Census Bureau, "Annual Survey of Entrepreneurs: Characteristics of Business Owners, 2014," Public-use data tables 2014. Reference year 2014. Tables SE1400CSCBO07 (education) and SE1400CSCBO08 (age). Accessed October 2, 2026.

_ , "Census Bureau Releases New Data on Minority-Owned, Veteran-Owned and Women-Owned Businesses," Annual Business Survey data release November 2022. November 10. Reports the 2021 Annual Business Survey, reference year 2020.

_ , "Census Bureau Releases New Data on Minority-Owned, Veteran-Owned and Women-Owned Businesses," Annual Business Survey data release December 2024. December 19. Reports the 2023 Annual Business Survey, reference year 2022.

For Online Publication

Appendix to "Who Hires Whom? Entrepreneurial Backgrounds and Labor Market Opportunities"

Josh Feng, *University of Utah – David Eccles School of Business*
Xavier Jaravel, *London School of Economics*
October 2026

A Data Construction

A.A LEHD

We follow the methodology in Kerr and Kerr (2016) to identify new firms and their founders in the LEHD data. Specifically, we use the Longitudinal Business Database (LBD) to identify firms in their first year of existence. We filter out companies that had earlier entries in the LBD, as these represent existing firms that appear as new in the LEHD only because of state coverage limitations.

For each new firm, we flag the top three earners in the company's first year in the LEHD data and classify them as founders. Next, we compute average demographics of non-founder employees in each year relative to the firm's founding. Demographic information comes from the Decennial Census files linked to the LEHD.

For our primary analysis, we construct a balanced panel of firms that have at least one non-founder employee with valid MSA and industry information in each of years 0 through 5 relative to founding. The filtering leaves approximately 825,000 firms founded between 1992 and 2014. We also construct an unbalanced panel, which we use to study longer-run dynamics.

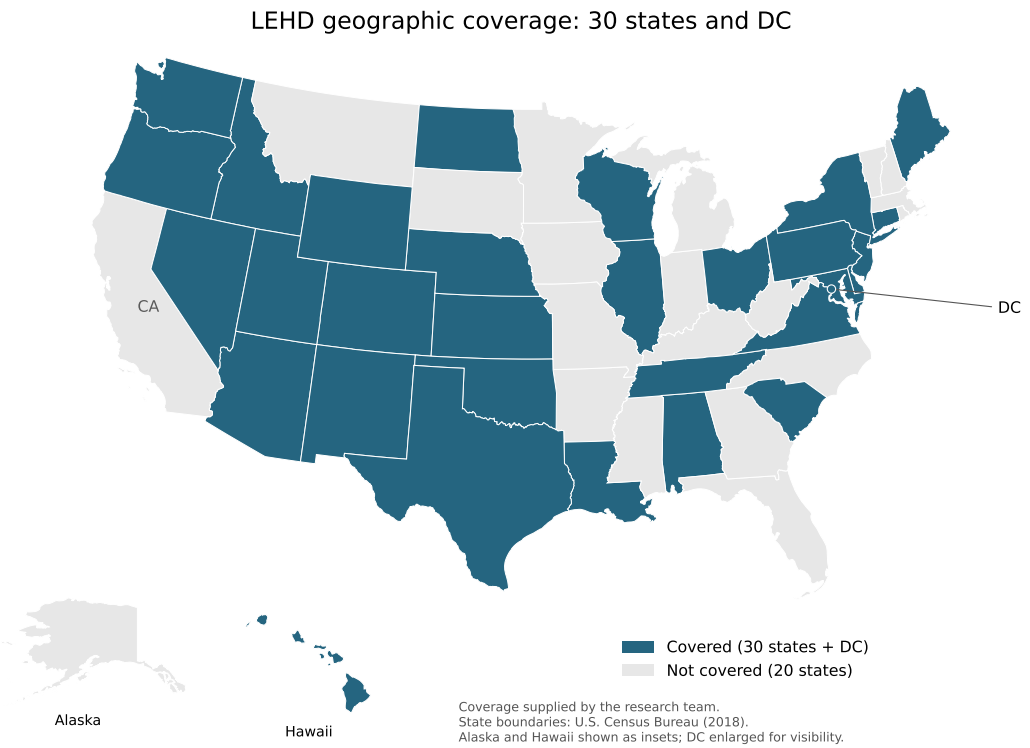


Figure A1: Geographic Coverage of the LEHD Sample

Notes: The states covered are Alabama, Arizona, Colorado, Connecticut, Delaware, Hawaii, Idaho, Illinois, Kansas, Louisiana, Maine, Maryland, Nebraska, Nevada, New Jersey, New Mexico, New York, North Dakota, Ohio, Oklahoma, Oregon, Pennsylvania, South Carolina, Tennessee, Texas, Utah, Virginia, Washington state, Wisconsin, Wyoming, plus the District of Columbia.

A.B Geographic Coverage and Representativeness

Figure A1 shows the 30 states and the District of Columbia covered by our LEHD data. This coverage is based on state-level data sharing policies, and is not representative. Here, we provide some context using public Census population statistics.

Table A1 compares residents of the covered states with the population of the 50 states and DC using the 2010–2014 ACS five-year estimates. The states in our sample cover 57.8 percent of U.S. residents. The covered population is close to the national population in its gender and age composition. There are slight discrepancies in college attainment (29.83 in the covered population versus 29.28 percent in the overall), per capita income ($29,144 versus $28,555), and poverty rates (15.12 versus 15.59 percent). Foreign-born, Asian, and Hispanic residents are underrepresented by 1.05, 0.54, and 0.75 percentage points, respectively.

Table A1: Covered-State Population Compared with the United States

	Covered	U.S.	Difference	California
Female	50.82	50.81	0.02	50.32
Age 18-64	62.79	62.77	0.02	63.67
Age 65+	13.55	13.75	-0.20	12.13
White alone, non-Hispanic	63.71	62.77	0.95	39.16
Black alone	12.88	12.60	0.28	5.94
Asian alone	4.46	5.00	-0.54	13.48
Hispanic, any race	16.14	16.90	-0.75	38.18
Foreign-born	12.02	13.07	-1.05	27.03
Bachelor's or higher, age 25+	29.83	29.28	0.56	31.00
Graduate degree, age 25+	11.37	11.01	0.36	11.41
Below poverty threshold	15.12	15.59	-0.47	16.38
Labor force participation, age 16+	64.45	63.90	0.55	63.83
Unemployment, civilian labor force	8.64	9.18	-0.54	10.99
Per capita income (2014 dollars)	29,144	28,555	589	29,906

Notes: "Covered" includes the 30 states and DC summarized in Figure A1. The second column ("U.S.") includes all 50 states and DC. Each cell is in percentage points, except per capita income, which is in 2014 dollars. Source: 2010–2014 ACS five-year Summary File.

A.C Crunchbase

Our second sample merges startup information from Crunchbase with employee work histories from LinkedIn, as compiled by Revelio Labs.

We focus on startups in Crunchbase that were founded in or after 2007, roughly the time Crunchbase began systematically tracking new firms. Our Crunchbase data comes from a December 2020 snapshot, while the LinkedIn data comes from Revelio Labs' April 2023 snapshot.

We use a crosswalk provided by Revelio Labs to link Crunchbase firm identifiers to Revelio company IDs (RCIDs). Founders are identified based on raw job titles in LinkedIn profiles (e.g., "Founder," "Co-Founder," "Founding CEO"). Non-founder employees working at each firm in a given year are identified based on the start and end dates reported in their LinkedIn work histories.

Unlike the LEHD, demographic variables are constructed from names rather than self-reported data. As discussed in the main text, name-based classification introduces measurement error that likely attenuates our estimates relative to the Census sample, where

demographics are directly reported.

For our primary analysis, we construct a balanced sample of 44,545 firms that have non-founder employees with LinkedIn profiles in each of years 0 through 5 after founding. We also construct an unbalanced panel to study longer-run dynamics out to 20 years after founding. We remove the founding year filter for the unbalanced sample.

A.D LEHD Sample Sizes

Table A2 reports rounded firm counts for the primary LEHD samples and subsamples used in the heterogeneity analyses.

Table A2: LEHD Sample Sizes

Sample	Subsample	Firms
Balanced Sample		825,000
	Service	649,000
	High Growth	78,000
	Economic Census Sample	336,000
Unbalanced Sample		3,582,000

Notes: Rounded firm counts for LEHD samples and subsamples used in the analysis. The balanced sample requires firms to have at least one non-founder employee in each of years 0 through 5 relative to founding. Source: authors' calculations based on U.S. Census data (FSRDC Project Number 2229, Disclosure Review Board (DRB) approval number: CBDRB-FY25-P2229-R11972).

B Additional Results

This section presents additional results that complement the main analysis. We first report robustness checks for the core LEHD estimates, then evidence on growth potential and firm size, followed by the Crunchbase reconciliation and persistence results. Additional heterogeneity, dynamics, production, and wage results follow.

B.A Core LEHD Estimates and Robustness

B.A.1 Projected Worker Fractions (LEHD)

Figure A2 presents the core LEHD estimates in terms of projected worker fractions. For each demographic characteristic and year relative to founding, we plot the estimated intercept,

which corresponds to firms with no founders from a given group, and the intercept plus the homophily coefficient, which corresponds to firms whose founders are all from that group. Because founder fractions are not binary, the projections for firms with all founders from a group rely on the linear specification.

B.A.2 Founder Definition (LEHD)

We assess robustness to how we measure founders.

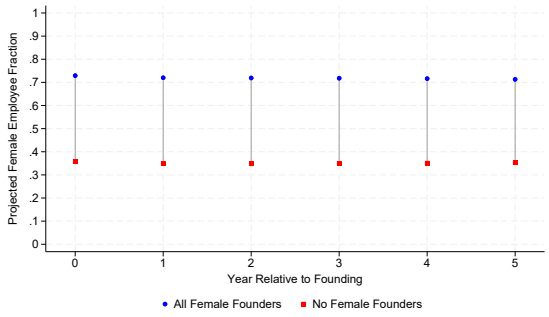
Number of Founders. As noted in the main text, a limitation of the LEHD data is that we cannot measure founder directly. Our baseline specification identifies founders as the top three earners in a new firm's first year. Here, we assess sensitivity to this definition.

Table A3 reports estimates from the balanced LEHD sample under three alternative founder definitions: one, two, and three top earners. For parsimony, we pool all six years together in the reported estimates instead of reporting estimates by age and demographic. Panel A restricts founders to the single highest earner, Panel B uses the top two, and Panel C (our default) uses the top three.

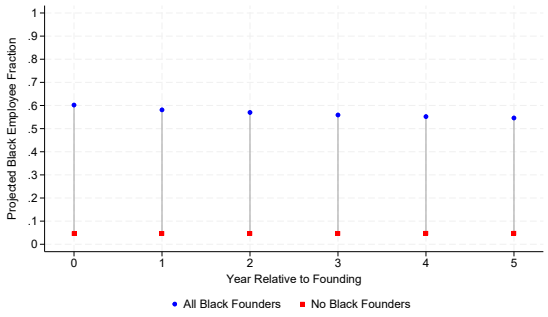
The estimates increase monotonically with the number of individuals classified as founders. For example, the female homophily coefficient rises from 0.162 (one founder) to 0.277 (two founders) to 0.364 (three founders). One leading explanation is that when we classify more individuals as founders, we are also including early managers, who may have a separate influence on hiring.

B.A.3 Geographic Control Level

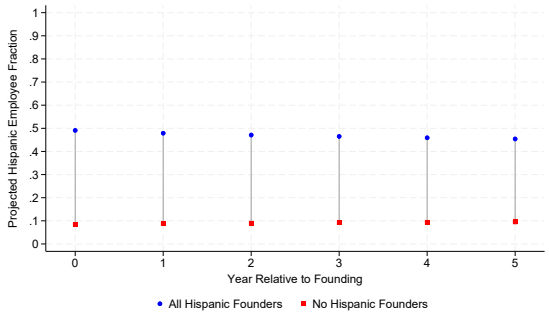
In the core results, we compare startups within two-digit by MSA by founding year cells. MSA may be too coarse to control for the labor market conditions facing new firms. Figure A3 reports a comparison of the estimates using MSA versus zip code as the geographic cell. We find slightly smaller estimates across the demographic variables when analyzing within zip codes, but the estimates remain quantitatively large.



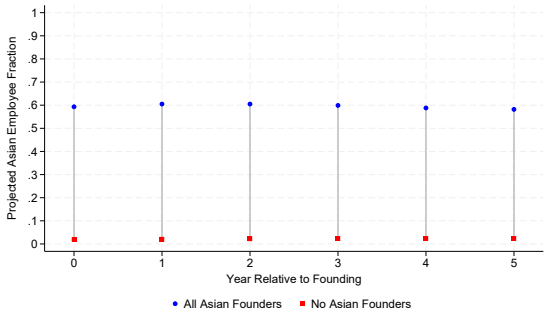
(a) Female



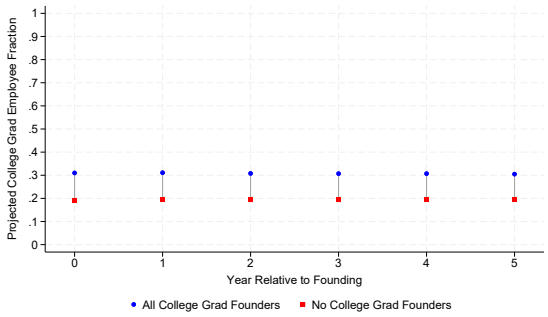
(b) Black



(c) Hispanic



(d) Asian



(e) College Educated

Figure A2: Projected Worker Fractions from Balanced Sample of New Firms

Notes: Each panel plots projected worker fractions based on Equation 1 for a different demographic variable. For each year relative to founding, we plot the estimated intercept (firms with no founders from the group) and the intercept plus $\hat{\beta}_t$ (firms whose founders are all from the group). Estimates come from the specification with two-digit industry-by-MSA-by-founding-year fixed effects. We omit age because the founder measure is not a fraction. All standard errors are clustered at the firm level and are small relative to the point estimates, so we have omitted confidence intervals for legibility. Source: authors' calculations based on U.S. Census data (FSRDC Project Number 2229, Disclosure Review Board (DRB) approval number: CBDRB-FY25-P2229-R11972).

Table A3: Entrepreneur-Worker Homophily (Robustness to Founder Number)

Panel A: One Founder

Employee Fraction:	Female	Black	Hispanic	Asian	Age	College
Female Founder Fraction	0.162***					
	(0.001)					
Black Founder Fraction		0.309***				
		(0.002)				
Hispanic Founder Fraction			0.194***			
			(0.001)			
Asian Founder Fraction				0.363***		
				(0.001)		
Average Founder Age					0.156***	
					(0.001)	
College Grad Founder Fraction						0.040***
						(0.000)
Constant	0.446***	0.060***	0.114***	0.034***	25.98***	0.214***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.029)	(0.000)
Observations	7222000	7222000	7222000	7222000	7222000	7222000

Panel B: Two Founders

Employee Fraction:	Female	Black	Hispanic	Asian	Age	College
Female Founder Fraction	0.277***					
	(0.001)					
Black Founder Fraction		0.434***				
		(0.002)				
Hispanic Founder Fraction			0.298***			
			(0.001)			
Asian Founder Fraction				0.484***		
				(0.001)		
Average Founder Age					0.251***	
					(0.001)	
College Grad Founder Fraction						0.078***
						(0.001)
Constant	0.396***	0.053***	0.102***	0.027***	22.39***	0.204***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.034)	(0.000)
Observations	7222000	7222000	7222000	7222000	7222000	7222000

Panel C: Three Founders (Default)

Employee Fraction:	Female	Black	Hispanic	Asian	Age	College
Female Founder Fraction	0.364***					
	(0.001)					
Black Founder Fraction		0.512***				
		(0.002)				
Hispanic Founder Fraction			0.373***			
			(0.001)			
Asian Founder Fraction				0.565***		
				(0.001)		
Average Founder Age					0.324***	
					(0.001)	
College Grad Founder Fraction						0.113***
						(0.001)
Constant	0.353***	0.047***	0.092***	0.023***	19.76***	0.195***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.035)	(0.000)
Observations	7222000	7222000	7222000	7222000	7222000	7222000

Notes: Regressions pooling all six years within the balanced LEHD sample. All specifications include Industry-MSA-founding year fixed effects. Standard errors in parentheses *$p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Source: authors' calculations based on U.S. Census data (FSRDC Project Number 2229, Disclosure Review Board (DRB) approval number: CBDRB-FY25-P2229-R11972).

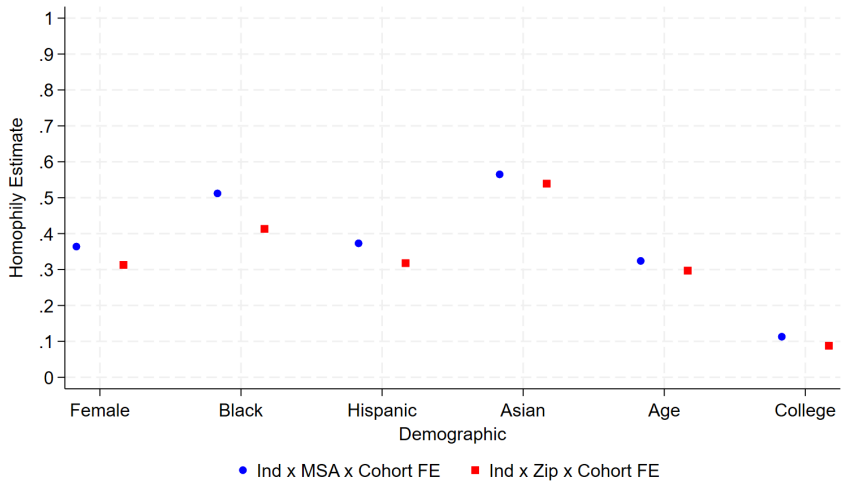


Figure A3: Comparison of Results at Different Levels of Geographic Controls

Notes: estimates of Equation 1 using the balanced LEHD sample. We compare estimates using two-digit industry-by-MSA-by-founding-year fixed effects versus two-digit industry-by-zip-code-by-founding-year fixed effects. Source: authors' calculations based on U.S. Census data (FSRDC Project Number 2229, Disclosure Review Board (DRB) approval numbers: CBDRB-FY25-P2229-R11972 and CBDRB-FY26-P2229-R12962).

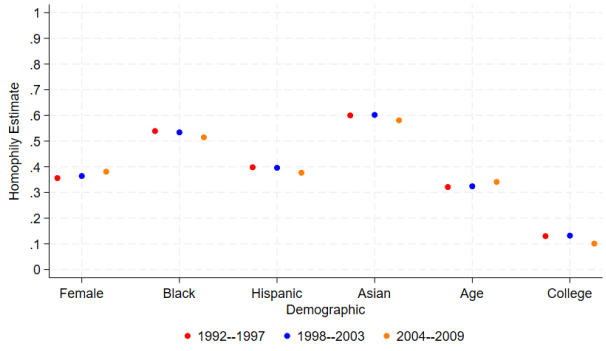


Figure A4: Homophily by Founding Year Cohorts

Notes: estimates of Equation 1 using the balanced LEHD sample and reporting results separately for each founding cohort group (1992–1997, 1998–2003, 2004–2009). All standard errors are small relative to the point estimates; confidence intervals are omitted for legibility. Source: authors' calculations based on U.S. Census data (FSRDC Project Number 2229, Disclosure Review Board (DRB) approval number: CBDRB-FY25-P2229-R11972).

B.A.4 Founding-Year Cohorts (LEHD)

B.B LEHD Weighting and High-Growth Results

Table A4 reports specifications with two layers of weighting. We first compute employee demographic shares using individual earnings as weights, so that higher-paid employees receive more weight. We then weight the regression by the square root of total employees, giving more influence to larger firms. The year-0 estimates are close to the corresponding unweighted year-0 baseline: 0.360 for female (versus 0.371), 0.544 for Black (versus 0.556), and 0.377 for Hispanic (versus 0.406).

We also compare firms in the top decile of realized employment growth from year 0 to year 5 with the remaining firms. Figure 2 in the main text reports the nearly identical homophily estimates across these groups.

B.C Additional Crunchbase Estimates

B.C.1 Projected Worker Fractions

Figure A5 presents the core Crunchbase estimates in terms of projected worker fractions, constructed in the same way as Figure A2. The projected female employee share at firms with no female founders rises from 0.31 in year 0 to 0.34 in year 5, while the gap relative to firms with all female founders remains stable. The projected shares for Black and Hispanic employees change by less than 0.01.

B.C.2 Weighting by Firm Size

Weighting by Firm Size. Tables A5 and A6 present estimates from the balanced Crunchbase sample under two alternative weighting schemes: square root of employees and total employees, respectively. The estimates are somewhat smaller under employee weighting (e.g., 0.138 versus 0.174 for female in year 0 with linear weights), consistent with slightly lower homophily among the largest firms. However, the estimates under linear weighting increase as firms age, suggesting that growing firms with different founder backgrounds maintain compositional differences over time.

Table A4: Entrepreneur-Worker Homophily Twice-Weighted (Balanced Sample, LEHD)

Panel A: Female

Year relative to founding:	0	1	2	3	4	5
Female Founder Fraction	0.360***	0.359***	0.363***	0.360***	0.357***	0.353***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Constant	0.352***	0.334***	0.327***	0.326***	0.326***	0.328***
	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
N	825000	825000	825000	825000	825000	825000

Panel B: Black

Year relative to founding:	0	1	2	3	4	5
Black Founder Fraction	0.544***	0.517***	0.506***	0.495***	0.486***	0.478***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Constant	0.056***	0.054***	0.053***	0.053***	0.054***	0.054***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
N	825000	825000	825000	825000	825000	825000

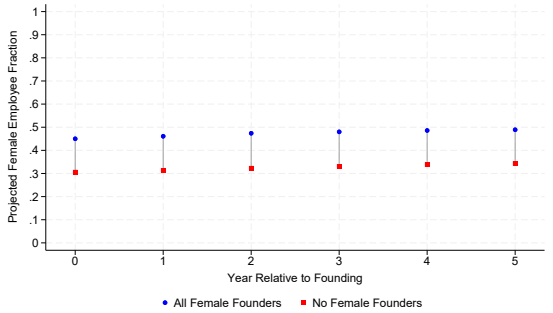
Panel C: Hispanic

Year relative to founding:	0	1	2	3	4	5
Hispanic Founder Fraction	0.377***	0.361***	0.352***	0.344***	0.337***	0.331***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Constant	0.094***	0.095***	0.095***	0.097***	0.099***	0.100***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
N	825000	825000	825000	825000	825000	825000

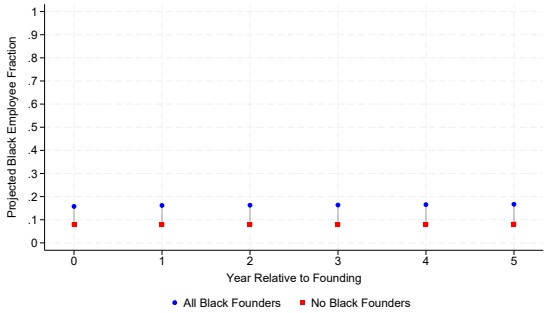
Panel D: College-Educated

Year relative to founding:	0	1	2	3	4	5
College-Educated	0.133***	0.136***	0.137***	0.137***	0.136***	0.134***
Founder Fraction	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Constant	0.184***	0.194***	0.197***	0.199***	0.201***	0.203***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
N	825000	825000	825000	825000	825000	825000

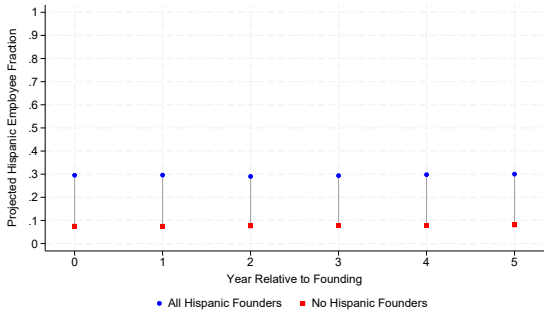
Notes: All specifications include Industry-MSA-founding year fixed effects. Standard errors in parentheses *$p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Source: authors' calculations based on U.S. Census data (FSRDC Project Number 2229, Disclosure Review Board (DRB) approval number: CBDRB-FY25-P2229-R11972).



(a) Female



(b) Black



(c) Hispanic

Figure A5: Projected Worker Fractions from Balanced Sample of New Firms (Crunchbase)

Notes: Each panel plots projected worker fractions based on Equation 1 for a different demographic variable, using the balanced Crunchbase sample. For each year relative to founding, we plot the estimated intercept (firms with no founders from the group) and the intercept plus $\hat{\beta}_t$ (firms whose founders are all from the group). Estimates come from the specification with industry-by-state-by-founding-year fixed effects. Source: Crunchbase merged to LinkedIn (Revelio Labs).

Table A5: Entrepreneur-Worker Homophily Weighting by Square Root of Employees (Balanced Sample, Crunchbase)

Panel A: Female

Year relative to founding:	0	1	2	3	4	5
Female Founder Fraction	0.164***	0.173***	0.176***	0.175***	0.171***	0.166***
	(0.00561)	(0.00492)	(0.00471)	(0.00469)	(0.00475)	(0.00478)
Constant	0.333***	0.334***	0.341***	0.349***	0.356***	0.363***
	(0.00188)	(0.00157)	(0.00145)	(0.00139)	(0.00138)	(0.00139)
N	44545	44545	44545	44545	44545	44545

Panel B: Black

Year relative to founding:	0	1	2	3	4	5
Black Founder Fraction	0.0985***	0.104***	0.107***	0.109***	0.108***	0.109***
	(0.00681)	(0.00646)	(0.00621)	(0.00621)	(0.00614)	(0.00643)
Constant	0.0793***	0.0783***	0.0784***	0.0791***	0.0799***	0.0807***
	(0.000675)	(0.000581)	(0.000538)	(0.000522)	(0.000513)	(0.000526)
N	44545	44545	44545	44545	44545	44545

Panel C: Hispanic

Year relative to founding:	0	1	2	3	4	5
Hispanic Founder Fraction	0.226***	0.231***	0.229***	0.229***	0.227***	0.229***
	(0.0102)	(0.00909)	(0.00881)	(0.00868)	(0.00846)	(0.00849)
Constant	0.0783***	0.0775***	0.0783***	0.0797***	0.0809***	0.0822***
	(0.000915)	(0.000770)	(0.000713)	(0.000690)	(0.000681)	(0.000683)
N	44545	44545	44545	44545	44545	44545

Notes: Estimates of Equation 1 using the balanced Crunchbase sample, reported separately by year relative to founding, weighting observations by the square root of firm employment and including industry-by-state-by-founding-year fixed effects. Standard errors, clustered at the firm level, in parentheses. *$p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Source: Crunchbase merged to LinkedIn (Revelio Labs).

Table A6: Entrepreneur-Worker Homophily Weighting by Employees (Balanced Sample, Crunchbase)

Panel A: Female

Year relative to founding:	0	1	2	3	4	5
Female Founder Fraction	0.138***	0.147***	0.153***	0.157***	0.159***	0.154***
	(0.0285)	(0.0245)	(0.0220)	(0.0205)	(0.0200)	(0.0193)
Constant	0.369***	0.367***	0.369***	0.373***	0.377***	0.382***
	(0.0138)	(0.0116)	(0.0101)	(0.00888)	(0.00812)	(0.00781)
N	44545	44545	44545	44545	44545	44545

Panel B: Black

Year relative to founding:	0	1	2	3	4	5
Black Founder Fraction	0.107***	0.112***	0.114***	0.116***	0.118***	0.123***
	(0.0238)	(0.0225)	(0.0214)	(0.0206)	(0.0204)	(0.0210)
Constant	0.0738***	0.0739***	0.0745***	0.0758***	0.0761***	0.0762***
	(0.00422)	(0.00392)	(0.00363)	(0.00328)	(0.00320)	(0.00330)
N	44545	44545	44545	44545	44545	44545

Panel C: Hispanic

Year relative to founding:	0	1	2	3	4	5
Hispanic Founder Fraction	0.130***	0.143***	0.153***	0.156***	0.154***	0.158***
	(0.0388)	(0.0316)	(0.0280)	(0.0249)	(0.0223)	(0.0208)
Constant	0.0875***	0.0866***	0.0856***	0.0856***	0.0858***	0.0861***
	(0.00393)	(0.00327)	(0.00277)	(0.00238)	(0.00213)	(0.00199)
N	44545	44545	44545	44545	44545	44545

Notes: Estimates of Equation 1 using the balanced Crunchbase sample, reported separately by year relative to founding, weighting observations by total firm employment and including industry-by-state-by-founding-year fixed effects. Standard errors, clustered at the firm level, in parentheses. $^{*}p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Source: Crunchbase merged to LinkedIn (Revelio Labs).

Table A7: Entrepreneur-Worker Homophily by Firm Employment Growth (Crunchbase)

Panel A: Female

	Top Decile	Other
Female Founder Fraction	0.130***	0.144***
	(0.0106)	(0.00386)
Constant	0.353***	0.335***
	(0.00238)	(0.00116)
N	45424	404191

Panel B: Black

	Top Decile	Other
Black Founder Fraction	0.101***	0.0826***
	(0.0267)	(0.00490)
Constant	0.0793***	0.0813***
	(0.00171)	(0.000471)
N	45425	404727

Panel C: Hispanic

	Top Decile	Other
Hispanic Founder Fraction	0.179***	0.223***
	(0.0210)	(0.00715)
Constant	0.0812***	0.0794***
	(0.00148)	(0.000591)
N	45425	404727

Notes: estimates of Equation 1 using the balanced Crunchbase sample and pooling across years 0–5 and including industry-by-state-by-founding-year fixed effects. Standard errors in parentheses and are clustered at the firm level *$p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

B.C.3 Employment Growth and Firm Size

Employment Growth. Table A7 splits the Crunchbase sample by employment growth, comparing the fastest-growing decile to the remaining firms. The estimates are similar across both groups.

Firm Size. Table A8 presents estimates by firm size quintile. We find slightly smaller coefficients for the top quintile but estimates that remain quantitatively large across all size groups.

B.C.4 Founding-Year Cohorts

Cohort Trends. Figure A6 reports homophily estimates by detailed founding year in the Crunchbase sample. We find fairly stable estimates with a slight upwards trend.

Table A8: Entrepreneur-Worker Homophily by Size (Crunchbase)

Panel A: Female

	Q1	Q2	Q3	Q4	Q5
Female Founder Fraction	0.150***	0.163***	0.152***	0.136***	0.110***
	(0.0104)	(0.00610)	(0.00504)	(0.00461)	(0.00521)
Constant	0.292***	0.315***	0.336***	0.356***	0.389***
	(0.00295)	(0.00176)	(0.00142)	(0.00125)	(0.00135)
N	94433	91629	82180	88180	93193

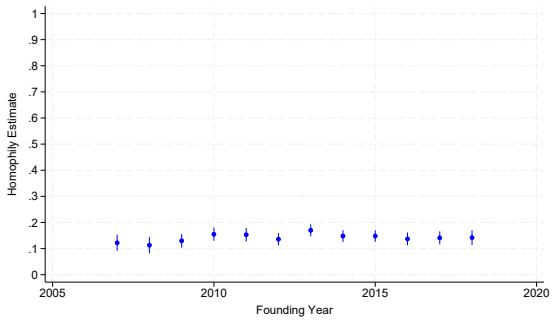
Panel B: Black

	Q1	Q2	Q3	Q4	Q5
Black Founder Fraction	0.0964***	0.0805***	0.0881***	0.0597***	0.0838***
	(0.0115)	(0.00689)	(0.00733)	(0.00570)	(0.0119)
Constant	0.0813***	0.0805***	0.0800***	0.0811***	0.0834***
	(0.00125)	(0.000680)	(0.000593)	(0.000502)	(0.000802)
N	94957	91642	82180	88180	93193

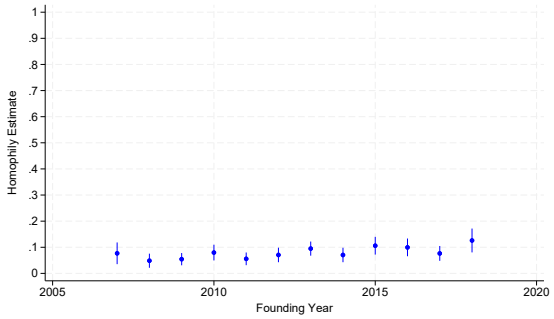
Panel C: Hispanic

	Q1	Q2	Q3	Q4	Q5
Hispanic Founder Fraction	0.197***	0.210***	0.240***	0.244***	0.199***
	(0.0140)	(0.00957)	(0.0106)	(0.0110)	(0.0136)
Constant	0.0750***	0.0772***	0.0791***	0.0804***	0.0869***
	(0.00139)	(0.000855)	(0.000812)	(0.000795)	(0.000931)
N	94957	91642	82180	88180	93193

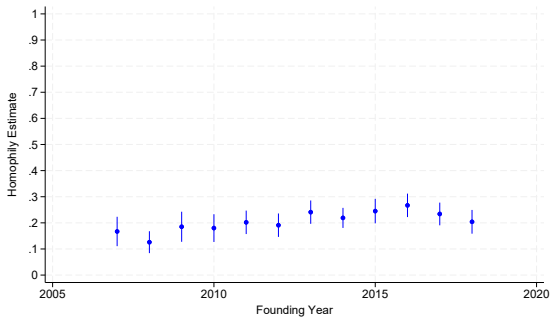
Notes: estimates of Equation 1 using the balanced Crunchbase sample and pooling across years 0–5 and including industry-by-state-by-founding-year fixed effects. Q5 represents the set of largest firm-years. Standard errors in parentheses and are clustered at the firm level: $^*p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.



(a) Female



(b) Black



(c) Hispanic

Figure A6: Estimates by Founding Year (Crunchbase)

Notes: estimates of Equation 1 using the balanced Crunchbase sample. Standard errors are clustered at the firm level and the figure reports 95 percent confidence intervals.

B.D Reconciling the LEHD and Crunchbase Estimates

As noted in the main text, the homophily estimates from the Crunchbase/LinkedIn sample are smaller than those from the LEHD, and the gap is highly uneven across demographics. We provide additional evidence to unpack this gap and reach the following conclusions:

B.D.1 Pooled Crunchbase Estimates

Before comparing founder definitions, we put the two samples on the same footing. The Census founder-number robustness (Table A3) uses industry-by-MSA-by-founding-year fixed effects, so we use comparable controls for the Crunchbase sample.

Table A9 reports the pooled Crunchbase homophily under no fixed effects, industry-by-founding-year fixed effects, and industry-by-state-by-founding-year fixed effects. Adding the detailed controls lowers the estimates from 0.177 to 0.147 (female), 0.107 to 0.083 (Black), and 0.240 to 0.218 (Hispanic). All comparisons below use the detailed specification on both sides.

Table A9: Crunchbase Homophily under Alternative Fixed Effects

Employee fraction	No FE	Industry×founding year	Industry×state×founding year
Female	0.177	0.150	0.147
Black	0.107	0.104	0.083
Hispanic	0.240	0.237	0.218

Notes: Pooled homophily (years 0–5) in the balanced Crunchbase sample under three fixed-effects specifications, estimated on the panel expanded through the Crunchbase–Revelio crosswalk (as in the headline estimates); robust standard errors. "Cohort" is the founding year. Source: Crunchbase merged to LinkedIn (Revelio Labs).

B.D.2 Founder Counts

In the LEHD, we do not observe job titles and define founders as the three top earners in the firm's first year. In Crunchbase we observe titles and define founders from them, so the *number* of founders varies: in the balanced sample, 43.7% of firms have one title-identified founder, 29.3% two, 14.3% three, and 12.7% four or more (mean 2.3, median 2; Table A10, Panel A).

As noted above, the LEHD homophily coefficient rises monotonically with the number of people labeled founders (Table A3). We assess whether this gradient appears in the Crunchbase sample in two ways. First, we estimate homophily separately by number of founders in the Crunchbase sample. Table A10, Panel B reports the estimates. We generally find increasing estimates as the number of founders increase. Second, to be more in line with the Census methodology, we count founders based on seniority in the first year of a firm.[1] Holding the firm sample and the employee comparison group fixed and widening only the founder group (top-one, top-two, top-three most senior at founding), the Crunchbase gradient is 0.115 / 0.161 / 0.208 (female), 0.055 / 0.096 / 0.139 (Black), and 0.168 / 0.263 / 0.340 (Hispanic) (Table A11). The Hispanic estimates are closest in magnitude to the LEHD estimates. Female and Black increase as we include more founders, but are consistently smaller than the corresponding LEHD estimates.

We quantify the founder-definition channel by reweighting the per-count LEHD estimates by the Crunchbase founder-count distribution, $\beta^{\text{LEHD,reweighted}} = \sum_N p_N \, \beta^{\text{LEHD}}_{\min(N,3)}$ (Table A12). Reweighting lowers the implied LEHD female coefficient from 0.364 to 0.250, against a Crunchbase estimate of 0.147—closing about half of the female gap. For Hispanic it falls from 0.373 to 0.273, against 0.218, closing about 65% of the gap. For Black it falls only from 0.512 to 0.400, against 0.083: it closes about a quarter of the gap and leaves a large residual. Re-defining founders as the three most-senior people at founding tells the same story, raising the Crunchbase estimates from the single-founder level toward 0.210 / 0.138 / 0.347 for female / Black / Hispanic—into the LEHD range for Hispanic, partway for female, and still far below for Black.

The estimate rises with the number of people counted as founders in both data sets. Two non-exclusive mechanisms can generate this. The first potential mechanism is noise in founder measurement (e.g., noise in demographic classification, founders missing from LinkedIn). A firm's early members each partly reflect a latent demographic propensity of the founding team, so averaging more of them yields a less noisy proxy. This would make the higher-count estimate closer to the truth. A second mechanism is that the second and third

[1]Because there is no earnings data, we use the seniority from Crunchbase and then break ties based on: i) whether the person's role contains "founder"; and ii) start date at the firm.

individuals are capturing early hires (esp. managers) rather than owners. For example, if two male founders hire a female manager, then the manager would be measured as a "founder." If the manager ends up hiring more female workers, then we would be misattributing this as founder-employee homophily.

Table A10: Number of Founders in the Balanced Crunchbase Sample

Panel A: Distribution of the number of title-identified founders

	1	2	3	4+	Mean (Median)
Share of firms	0.437	0.293	0.143	0.127	2.27 (2)

Panel B: Homophily by realized founder count, detailed FE (firms grouped by count)

Employee fraction	1 founder	2 founders	3 founders	4+ founders
Female	0.116	0.159	0.201	0.200
Black	0.065	0.095	0.136	0.164
Hispanic	0.177	0.250	0.286	0.359

Notes: Balanced Crunchbase sample (44,545 firms). Panel A: distribution of the number of founders identified from LinkedIn job titles. Panel B: each cell is the coefficient from a separate regression of the employee demographic share on the founder demographic share, pooled over years 0–5 with industry-by-state-by-founding-year fixed effects, for firms grouped by their realized founder count. Because firms are grouped by count, sample composition changes across columns (cf. Table A11, which holds the sample fixed). Source: Crunchbase merged to LinkedIn (Revelio Labs).

B.D.3 Geographic Coverage of the Crunchbase Sample

Geographic coverage is another potential source of the difference between the LEHD and Crunchbase estimates. As shown above, our LEHD sample only covers thirty states, and does not include California. In contrast, our Crunchbase–LinkedIn sample covers all states, and about one third of the firms are headquartered in California. If homophily is generally weak in California for all startups, then it could account for some of the discrepancy between the LEHD and Crunchbase estimates.

To shed light on this, we restrict the balanced Crunchbase–LinkedIn sample to the same set of states as in our LEHD sample and re-estimate homophily. Figure A7 compares the estimates from this restricted sample to those derived from the full sample. At founding, the baseline coefficients change from 0.174 to 0.194 for female, 0.104 to 0.103 for Black, and 0.243 to 0.197 for Hispanic homophily. In the detailed fixed-effects specification, the

Table A11: Homophily by Founder Definition, Fixed Sample with Detailed FE (Crunchbase vs. LEHD)

Employee fraction	1 founder	2 founders	3 founders
Female			
Crunchbase (top-N senior)	0.115	0.161	0.208
LEHD (top-N earners)	0.162	0.277	0.364
Black			
Crunchbase (top-N senior)	0.055	0.096	0.139
LEHD (top-N earners)	0.309	0.434	0.512
Hispanic			
Crunchbase (top-N senior)	0.168	0.263	0.340
LEHD (top-N earners)	0.194	0.298	0.373

Notes: The Crunchbase rows hold the firm sample and the employee comparison group fixed (employees = everyone below the top three senior people at founding) and vary only the founder group (top-1/2/3 by seniority), pooled over years 0–5 with industry-by-state-by-founding-year fixed effects; the sample is the 31,582 balanced firms with ≥ 4 documented profiles at founding. The LEHD rows reproduce Table A3 (founders = top-N earners, industry-by-MSA-by-founding-year fixed effects). Sources: Crunchbase/LinkedIn (Revelio Labs) and authors' calculations on U.S. Census data.

Table A12: Reweighting the LEHD Estimate by the Crunchbase Founder-Count Distribution

Employee fraction	LEHD (3 founders)	LEHD (reweighted)	Crunchbase (detailed FE)	% of gap closed
Female	0.364	0.250	0.147	52
Black	0.512	0.400	0.083	26
Hispanic	0.373	0.273	0.218	65

Notes: "LEHD (reweighted)" applies the Crunchbase founder-count shares (Table A10, Panel A) to the per-count LEHD coefficients (Table A3), capping counts ≥ 4 at the three-founder coefficient. "% of gap closed" = (LEHD default − LEHD reweighted) / (LEHD default − Crunchbase). Both the LEHD and Crunchbase estimates pool years 0–5 with detailed fixed effects. Sources: Crunchbase/LinkedIn (Revelio Labs) and authors' Census calculations.

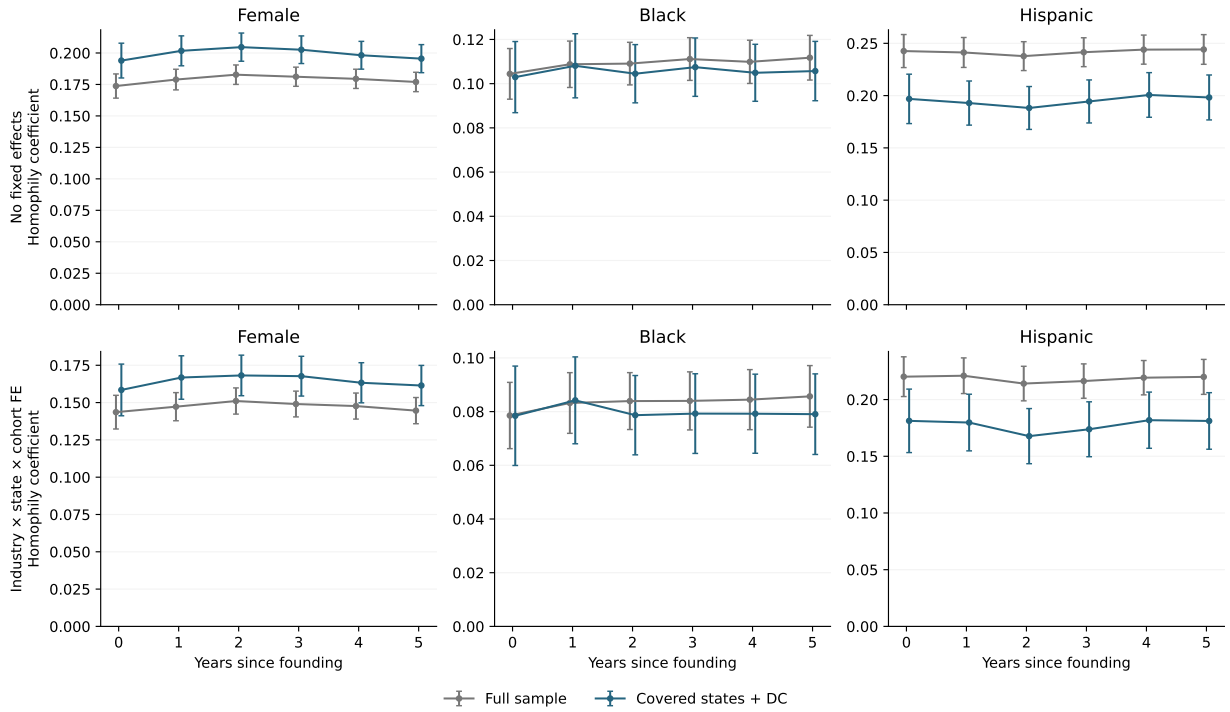


Figure A7: Crunchbase--LinkedIn Estimates Restricted to LEHD-Covered States

Notes: Comparisons of homophily estimates using the full sample of Crunchbase–LinkedIn startups and startups located in the LEHD states for which we have access. Intervals are 95 percent confidence intervals using heteroskedasticity-robust standard errors, as in the underlying core code.

corresponding changes are 0.144 to 0.158, 0.079 to 0.078, and 0.220 to 0.181.

The evidence suggests that geographic differences in homophily are unlikely to be accounting for the discrepancy between the core LEHD estimates and the Crunchbase–LinkedIn estimates. Of course, without LEHD data from the other states, we cannot provide the corresponding comparison.

B.E Persistence in Unbalanced Panels

The balanced panel ensures that our estimates are not confounded by differential firm exit, but it restricts the sample to relatively successful firms. Here, we relax the survival requirement and report estimates from unbalanced panels in both the LEHD and Crunchbase samples, which also allow us to track homophily over longer horizons.

LEHD Unbalanced Panel. Figure A8 reports estimates from the unbalanced LEHD sample, extending to 9 years after founding. The initial effects are very similar to the

balanced panel, and although they attenuate gradually they remain quantitatively large even at year 9 (the point estimates are reported in the main text). Because the sample is unbalanced, the decay reflects both within-firm convergence and differential attrition across firm types.

Crunchbase Unbalanced Panel. Figure A9 reports estimates from the unbalanced Crunchbase sample, covering years 0 through 20. The LinkedIn data allow us to track homophily patterns over a substantially longer horizon than the LEHD. The initial estimates are very similar to the balanced sample, and the effects persist strongly out to 20 years after founding (the point estimates are reported in the main text).

B.F Worker and Industry Heterogeneity

Industry Subsamples. Table A13 reports homophily estimates across broad Crunchbase industry groups (Software, Hardware, Manufacturing, Health Care). The results show consistent positive homophily across industries, though the magnitudes vary by industry and demographic group, with some estimates imprecisely estimated in smaller industries.

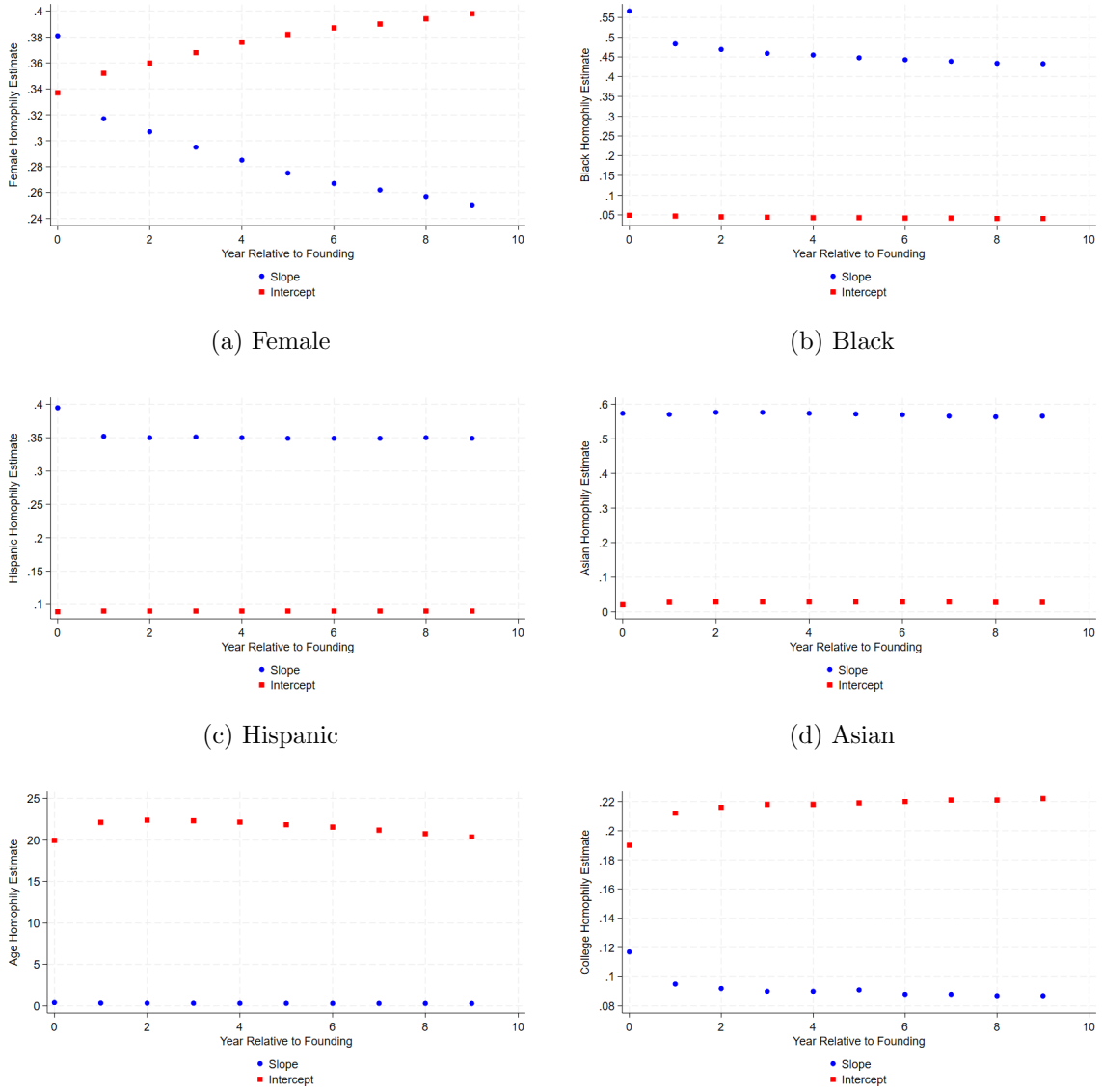
B.G Founder versus Initial Employees – Crunchbase Sample

Here, we report the horserace estimates using the Crunchbase--LinkedIn sample, where job titles allow more direct identification of founders. In the main text, we report corresponding estimates using the Census LEHD sample (Figure 5). Consistent with the results from the other sample, we find that founder effects persist over time, whereas the predictive coefficient on initial employee composition declines as firms age (Figure A11).

B.H Hiring versus Retention

Next, we present evidence comparing homophily estimates when we use all employees versus just new hires. Figure A12 shows that, for each demographic, the homophily coefficient for new hires decreases significantly more than for the overall homophily coefficient, albeit still remaining quantitatively large over time. This suggests that retention also plays a large role in generating the persistent homophily over time.

We also report the analogous comparison using the Crunchbase sample. Figure A13 shows



(a) Female

(b) Black

(c) Hispanic

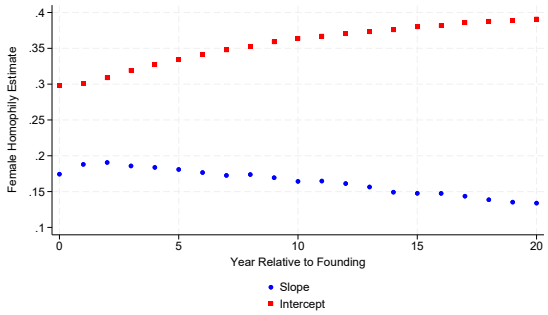
(d) Asian

(e) Age

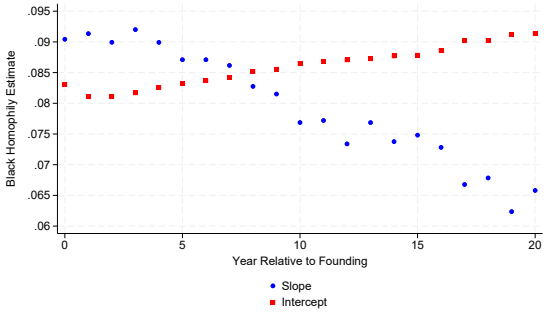
(f) College Educated

Figure A8: Estimates from Unbalanced Sample of New Firms

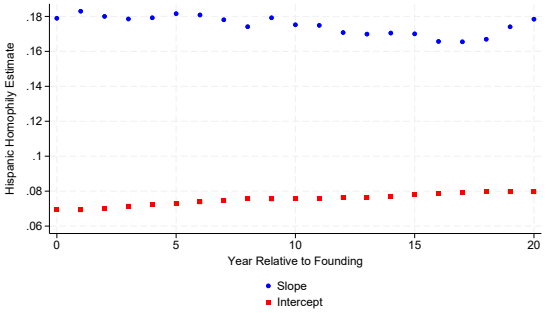
Notes: estimates of Equation 1 using the unbalanced LEHD sample. All standard errors are small relative to the point estimates; confidence intervals are omitted for legibility. Source: authors' calculations based on U.S. Census data (FSRDC Project Number 2229, Disclosure Review Board (DRB) approval number: CBDRB-FY25-P2229-R11972).



(a) Female



(b) Black



(c) Hispanic

Figure A9: Entrepreneur-Worker Homophily (Unbalanced Sample, Crunchbase)

Notes: estimates of Equation 1 using the unbalanced Crunchbase sample, covering years 0 through 20 relative to founding. Standard errors are clustered at the firm level and the figure reports 95 percent confidence intervals.

Table A13: Entrepreneur-Worker Homophily by Industry (Crunchbase)

Panel A: Female

	Software	Hardware	Manufacturing	Health Care
Female Founder Fraction	0.0932***	0.111***	0.0615	0.135***
	(0.00888)	(0.0242)	(0.0423)	(0.00991)
Constant	0.296***	0.261***	0.267***	0.408***
	(0.00236)	(0.00587)	(0.00946)	(0.00361)
N	81133	11932	4638	45130

Panel B: Black

	Software	Hardware	Manufacturing	Health Care
Black Founder Fraction	0.0928***	0.123**	0.0512	0.0839***
	(0.0112)	(0.0386)	(0.0377)	(0.0156)
Constant	0.0704***	0.0679***	0.0771***	0.0836***
	(0.000972)	(0.00265)	(0.00370)	(0.00139)
N	81271	11948	4643	45185

Panel C: Hispanic

	Software	Hardware	Manufacturing	Health Care
Hispanic Founder Fraction	0.263***	0.265***	0.229**	0.138***
	(0.0162)	(0.0503)	(0.0806)	(0.0190)
Constant	0.0710***	0.0768***	0.0718***	0.0786***
	(0.00129)	(0.00343)	(0.00553)	(0.00150)
N	81271	11948	4643	45185

Notes: estimates of Equation 1 using the balanced Crunchbase sample and pooling across years 0–5 and including industry-by-state-by-founding-year fixed effects. Standard errors in parentheses and are clustered at the firm level *$p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A14: Entrepreneur-Worker Homophily by Seniority (Crunchbase)

Panel A: Female

	Senior	Junior
Female Founder Fraction	0.142***	0.140***
	(0.00424)	(0.00393)
Constant	0.285***	0.385***
	(0.00124)	(0.00122)
N	407,430	407,025

Panel B: Black

	Senior	Junior
Black Founder Fraction	0.0805***	0.0785***
	(0.00500)	(0.00535)
Constant	0.0755***	0.0854***
	(0.000492)	(0.000508)
N	412,125	411,932

Panel C: Hispanic

	Senior	Junior
Hispanic Founder Fraction	0.216***	0.222***
	(0.00744)	(0.00736)
Constant	0.0688***	0.0887***
	(0.000613)	(0.000643)
N	412,125	411,932

Notes: estimates of Equation 1 using the balanced Crunchbase sample and pooling across years 0–5 and including industry-by-state-by-founding-year fixed effects. Standard errors in parentheses and are clustered at the firm level *$p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A15: Entrepreneur-Worker Homophily by Occupation (Crunchbase)

Panel A: Female

	Admin	Engineer	Finance	Marketing	Operations	Sales	Scientist
Female Founder Fraction	0.0806***	0.0771***	0.0813***	0.123***	0.120***	0.140***	0.134***
	(0.00638)	(0.00433)	(0.00765)	(0.00496)	(0.00666)	(0.00506)	(0.00906)
Constant	0.567***	0.205***	0.426***	0.479***	0.298***	0.338***	0.424***
	(0.00209)	(0.00118)	(0.00228)	(0.00165)	(0.00202)	(0.00151)	(0.00291)
N	227352	323327	192435	305620	248262	346563	134058

Panel B: Black

	Admin	Engineer	Finance	Marketing	Operations	Sales	Scientist
Black Founder Fraction	0.0861***	0.0831***	0.0723***	0.0753***	0.0867***	0.0681***	0.0688***
	(0.00851)	(0.00675)	(0.00962)	(0.00641)	(0.00847)	(0.00584)	(0.0112)
Constant	0.0918***	0.0714***	0.0830***	0.0853***	0.0809***	0.0839***	0.0886***
	(0.000870)	(0.000602)	(0.000918)	(0.000633)	(0.000858)	(0.000595)	(0.00112)
N	237805	332012	200471	313814	258641	353947	142167

Panel C: Hispanic

	Admin	Engineer	Finance	Marketing	Operations	Sales	Scientist
Hispanic Founder Fraction	0.224***	0.243***	0.221***	0.224***	0.200***	0.196***	0.182***
	(0.0112)	(0.00915)	(0.0122)	(0.00867)	(0.0109)	(0.00786)	(0.0137)
Constant	0.0892***	0.0795***	0.0793***	0.0842***	0.0780***	0.0776***	0.0798***
	(0.00101)	(0.000783)	(0.00107)	(0.000783)	(0.000999)	(0.000701)	(0.00125)
N	237805	332012	200471	313814	258641	353947	142167

Notes: estimates of Equation 1 using the balanced Crunchbase sample and pooling across years 0–5 and including industry-by-state-by-founding-year fixed effects. Standard errors in parentheses and are clustered at the firm level *$p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A16: Distribution of Homophily Estimates at the Industry Level

Panel A – All Industries

Percentile	Homophily Coefficient					
	Female	Black	Hispanic	Asian	Age	College
5	-0.2173	-0.1285	-0.2085	-0.02077	-0.4312	-0.2179
10	-0.04522	-0.02846	-0.0146	0.05343	-0.09821	-0.09125
25	0.1197	0.1895	0.2113	0.2836	0.1554	0.02385
50	0.2489	0.3725	0.3435	0.4926	0.2879	0.09186
75	0.3795	0.5019	0.4591	0.6697	0.4144	0.1921
90	0.5328	0.6409	0.6461	0.8196	0.624	0.3288
95	0.7178	0.7694	0.8454	0.9538	0.9075	0.4969

Panel B – Top Quartile (Number of New Firms)

Percentile	Homophily Coefficient					
	Female	Black	Hispanic	Asian	Age	College
5	0.06349	0.1399	0.1431	0.07074	0.1191	-0.008247
10	0.09168	0.2145	0.1845	0.1721	0.1575	0.01526
25	0.1496	0.3159	0.2659	0.3397	0.223	0.04129
50	0.241	0.4145	0.3418	0.4773	0.291	0.07855
75	0.3385	0.5005	0.4042	0.5935	0.3658	0.1285
90	0.4109	0.5665	0.4816	0.6965	0.4196	0.1915
95	0.4549	0.6407	0.5357	0.7443	0.4579	0.2274

Notes: Distribution of homophily estimates using the balanced LEHD sample, with coefficients estimated industry-by-industry (NAICS six-digit). All specifications include MSA-founding year fixed effects. Source: authors' calculations based on U.S. Census data (FSRDC Project Number 2229, DRB approval number: CBDRB-FY26-P2229-R12962).

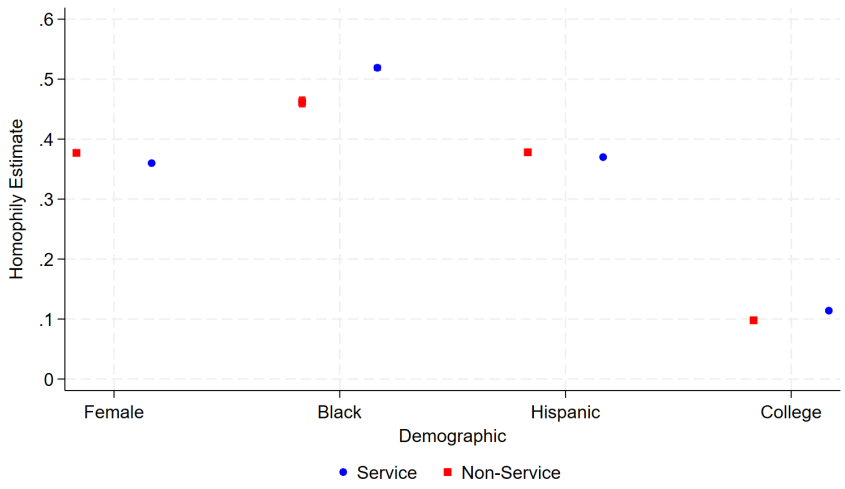


Figure A10: Homophily by Industry Sector (LEHD)

Notes: Pooled homophily coefficients from the balanced LEHD sample (Equation 1), separately for service industries (two-digit NAICS 51 and above) and non-service industries (49 and below). All specifications include industry-by-MSA-by-founding-year fixed effects. Bars show 95 percent confidence intervals. Source: authors' calculations based on U.S. Census data (FSRDC Project Number 2229, DRB approval numbers: CBDRB-FY25-P2229-R11972 and CBDRB-FY26-P2229-R12962).

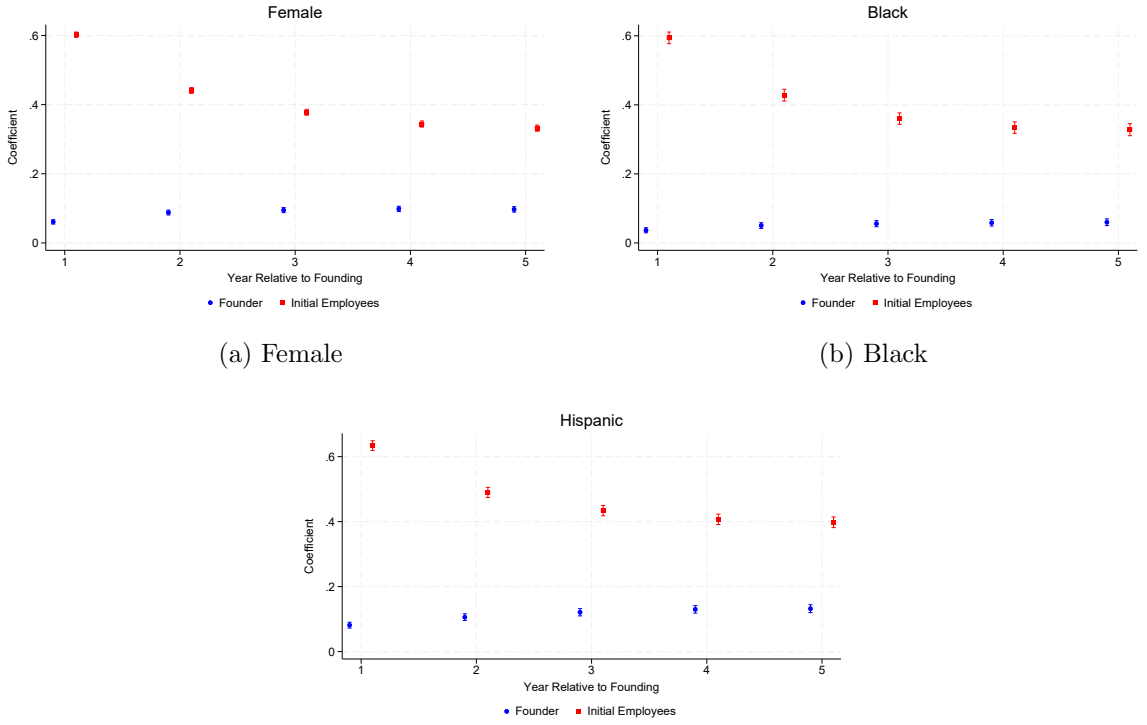
the overall homophily estimates alongside new-hire-specific estimates across years relative to founding. The patterns are broadly consistent with the LEHD results: new-hire homophily declines more over time than overall homophily, reinforcing the role of retention in generating persistent compositional differences.

B.I Production Structure and Additional Controls

A potential concern with our core results is that founders from different backgrounds may create firms with systematically different production structures, and that these differences are driving the observed employee composition patterns. We address this concern using data from the Economic Census.

Founder Characteristics and Production Structure. Table A17 reports associations between founder demographics and four measures of firm production structure: average wage, labor share, import intensity, and export intensity. These are the primary measures available to us in our Census project. These regressions use firm-level data from the 2007 and 2012 Economic Census, restricted to new firms within 5 years of founding.

We find meaningful differences in average wages: female-founded firms pay $15.3k less



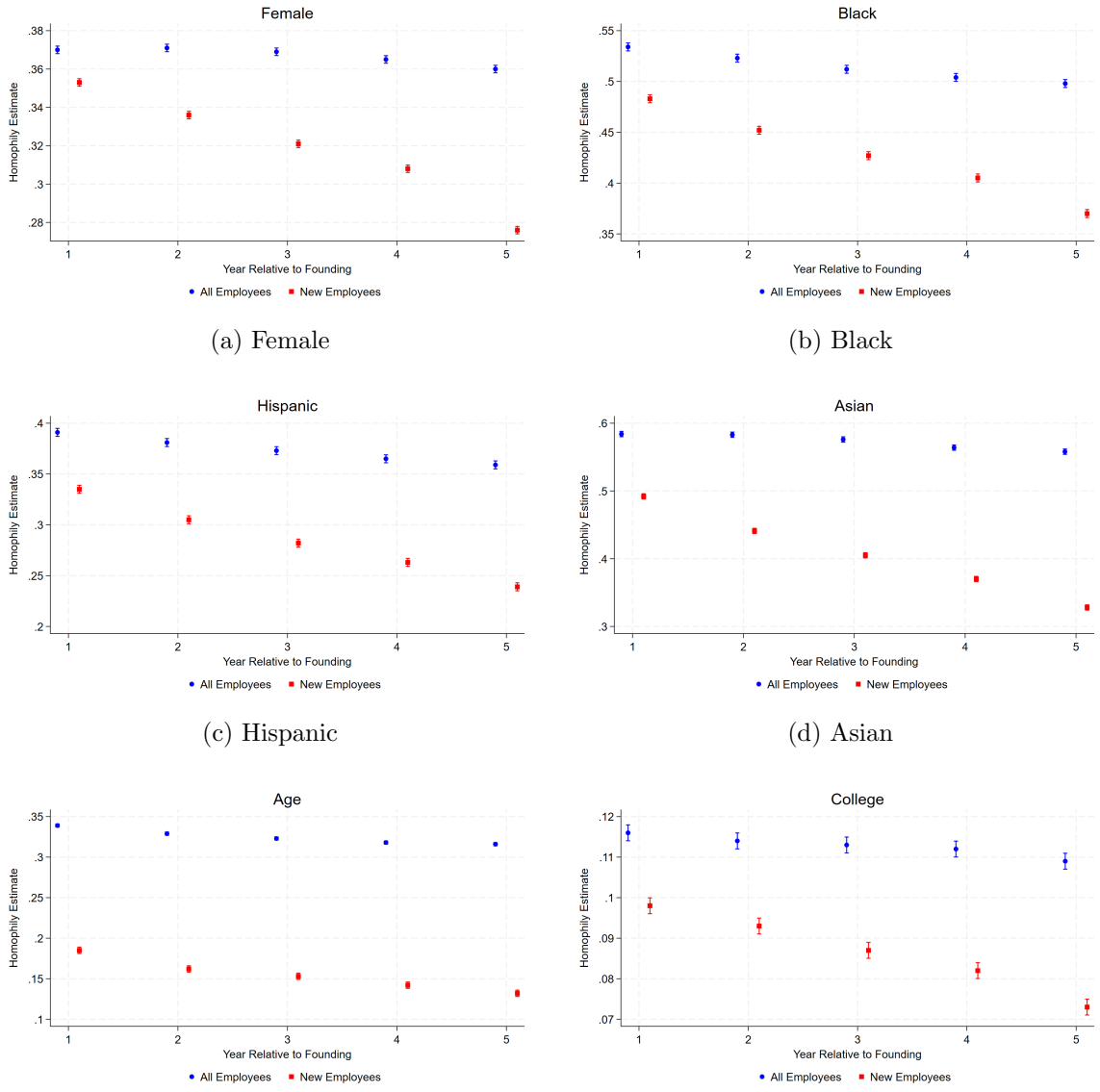
(a) Female

(b) Black

(c) Hispanic

Figure A11: Founder versus Initial Employee Demographics Horserace (Crunchbase)

Notes: Each panel plots coefficients from a regression of employee demographic share on both founder demographic share and initial (year 0) non-founder employee demographic share, estimated separately for years 1 through 5 relative to founding. Year 0 is excluded as the dependent variable because the initial employee share variable is exactly the same as the year 0 outcome. All specifications include industry-by-state-by-founding-year fixed effects. Standard errors are clustered at the firm level. The shaded areas represent 95% confidence intervals. Source: Crunchbase merged to LinkedIn (Revelio Labs).



(a) Female

(b) Black

(c) Hispanic

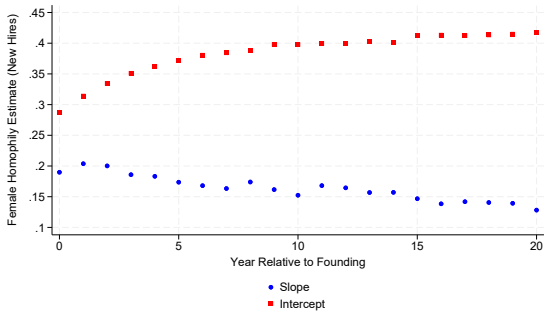
(d) Asian

(e) Age

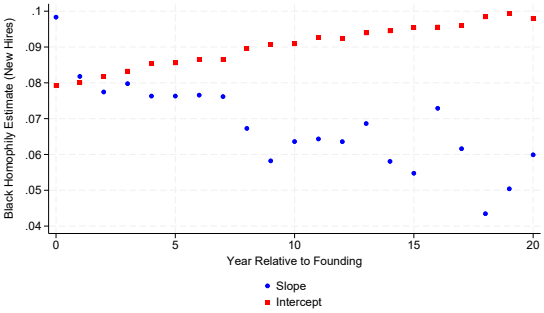
(f) College Educated

Figure A12: Comparison of Overall Homophily Estimates versus New Hire Homophily Estimates

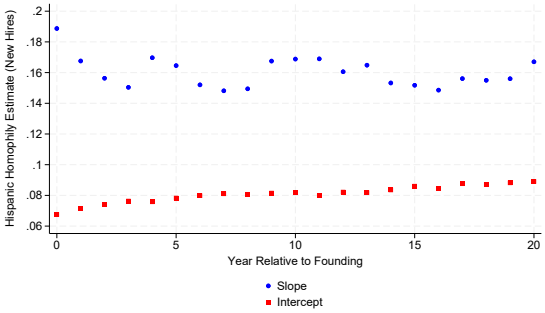
Notes: estimates of Equation 1 using the balanced LEHD sample. The two outcomes that are compared are overall demographic composition of workers and the demographic composition of new hires (the coefficients represent the difference across firms with different founder backgrounds). All standard errors are small relative to the point estimates; confidence intervals are omitted for legibility. Source: authors' calculations based on U.S. Census data (FSRDC Project Number 2229, Disclosure Review Board (DRB) approval numbers: CBDRB-FY25-P2229-R11972 and CBDRB-FY26-P2229-R12962).



(a) Female



(b) Black



(c) Hispanic

Figure A13: New Hire Homophily Estimates (Crunchbase)

Notes: estimates of Equation 1 using the unbalanced Crunchbase sample, focusing on the demographic composition of new hires. Standard errors are clustered at the firm level and the figure reports 95 percent confidence intervals.

Table A17: Association between Founder Characteristics and Firm Production Structure

	Average Wage ($000s)	Labor Share	Import Intensity	Export Intensity
Female Founder Fraction	-15.27***	-0.020*	-0.001	-0.001**
	(0.220)	(0.009)	(0.000)	(0.000)
Black Founder Fraction	-11.63***	0.021***	-0.002***	-0.000
	(0.371)	(0.005)	(0.000)	(0.001)
Hispanic Founder Fraction	-6.467***	-0.020***	-0.004***	-0.001*
	(0.277)	(0.004)	(0.000)	(0.001)
Asian Founder Fraction	-0.406	-0.035***	0.006***	0.002***
	(0.402)	(0.007)	(0.001)	(0.001)
Average Founder Age	0.282***	0.001***	0.000*	0.000
	(0.008)	(0.000)	(0.000)	(0.000)
College Grad Founder Fraction	19.76***	0.051***	0.006***	0.003***
	(0.300)	(0.011)	(0.000)	(0.001)
Observations	337,000	337,000	337,000	337,000

Notes: Each cell reports the coefficient from a separate firm-level regression of the column outcome on the row founder characteristic, based on characteristics measured during the first Economic Census available for the firm (2007 or 2012). Average wage is in thousands of dollars. All specifications include industry-by-MSA-by-founding-year fixed effects. Standard errors in parentheses. *$p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Source: authors' calculations based on U.S. Census data (FSRDC Project Number 2229, Disclosure Review Board (DRB) approval number: CBDRB-FY25-P2229-R11972).

on average, Black-founded firms $11.6k less, and Hispanic-founded firms $6.5k less, while college-educated founders pay $19.8k more. Differences in labor share, import intensity, and export intensity are statistically significant in some cases but economically small.

Homophily Controlling for Production Characteristics. Given these observed differences, we want to test whether they absorb some of the homophily estimates. Table A18 presents homophily estimates with and without controls for the production structure variables described above. The sample is restricted to firms founded between 2002 and 2012 that can be matched to the Economic Census, which is why the observation count differs from the main balanced sample.

Across all demographic groups except college education, the changes in homophily coefficients after including production structure controls are quantitatively small. For example, the female coefficient moves from 0.385 to 0.382, and the Black coefficient from 0.519 to 0.517. An exception is the college-educated coefficient. There is a more meaningful decline after adding controls (0.092 to 0.079). This suggests that observable differences in production structure can explain some of the education-based homophily.

Table A18: Homophily Estimates Controlling for Production Characteristics

Employee Fraction:	Female		Black		Hispanic		Asian		Age		College	
Female Found. Fr.	0.385***	0.382***										
	(0.001)	(0.001)										
Black Found. Fr.			0.519***	0.517***								
			(0.002)	(0.002)								
Hispanic Found. Fr.					0.358***	0.357***						
					(0.002)	(0.002)						
Asian Founder Fr.							0.564***	0.564***				
							(0.002)	(0.002)				
Founder Age									0.347***	0.343***		
									(0.001)	(0.002)		
College Found. Fr.											0.092***	0.079***
											(0.001)	(0.003)
Constant	0.385***	0.391***	0.049***	0.053***	0.094***	0.096***	0.026***	0.026***	19.66***	19.47***	0.201***	0.187***
	(0.001)	(0.002)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.054)	(0.070)	(0.000)	(0.003)
Controls	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y
Observations	2021000	2021000	2021000	2021000	2021000	2021000	2021000	2021000	2021000	2021000	2021000	2021000

Notes: Standard errors in parentheses $^{*}p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A19: AKM Decomposition Regressions (No Controls)

	Firm Fixed-Effect Differential				
Female Founder Fraction	0.197***				
	(0.002)				
Black Founder Fraction		0.157***			
		(0.004)			
Hispanic Founder Fraction			0.101***		
			(0.002)		
Asian Founder Fraction				0.091***	
				(0.003)	
College Grad Founder Fraction					0.093***
					(0.003)
Observations	1,794,000	798,000	1,822,000	1,116,000	1,125,000

Notes: Standard errors in parentheses and are clustered at the firm level. *$p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Source: authors' calculations based on U.S. Census data (FSRDC Project Number 2229, Disclosure Review Board (DRB) approval number: CBDRB-FY25-P2229-R11972).

B.J AKM Robustness Checks and Additional Results

First, Table A19 presents the corresponding estimates when we compute firm fixed effects without any additional controls (the X's in the first-stage AKM wage equation). We find qualitatively similar but larger effects relative to the estimates presented in the main text. This suggests that there are significant observable differences across individuals hired by entrepreneurs from different backgrounds.

Next, we present regressions of the raw demographic fixed effect versus founder characteristics in Table A20. The slope captures the difference from having more founders from the same demographic background, and the intercept reflects how the full sample of startups compare to the broader set of firms. We generally find that female and minority founded firms pay less than other startups. College-educated founders appear to pay college-educated individuals more than both startups and the broader sample of firms.

Third, we implement the approach in Bonhomme et al. (2019) and Lerner et al. (2024) to test for the importance of complementarities between worker fixed effects and firms in our setting. The two models we compare are simplified linear and interacted versions of the core AKM approach:

$$Y_{it} = \alpha_i + \psi_{g(j,t)} + \varepsilon_{it}$$

Table A20: AKM Decomposition Regressions, Raw Firm Effects

	Female FE	Black FE	Hispanic FE	Asian FE	College FE
Female Founder Fraction	-0.256***				
	(0.002)				
Black Founder Fraction		-0.315***			
		(0.005)			
Hispanic Founder Fraction			-0.189***		
			(0.003)		
Asian Founder Fraction				-0.098***	
				(0.004)	
College Grad Founder Fraction					0.596***
					(0.003)
Constant	-0.155***	-0.137***	-0.684***	-0.142***	-0.341***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
Observations	1,775,000	791,000	1,802,000	1,104,000	584,000

Notes: The dependent variable is the raw firm fixed effect for the indicated demographic group $\hat{\psi}_{jsd}$. Firm fixed effects are estimated separately by state and demographic group, controlling for age, age squared, and firm tenure. Standard errors in parentheses and are clustered at the firm level. *$p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Source: authors' calculations based on U.S. Census data (FSRDC Project Number 2229, Disclosure Review Board (DRB) approval number: CBDRB-FY26-P2229-R12962).

$$Y_{it} = \sum_{k=1}^{K} \beta_k \, \alpha_i \, \mathbf{1}\{g(j,t) = k\} + \sum_{k=2}^{K} \psi_k \, \mathbf{1}\{g(j,t) = k\} + \varepsilon_{it}$$

Essentially, the first is a simplified version of AKM where firms are clustered into K groups indexed by g (instead of firm-specific fixed effects). The second is a model that allows for flexible complementarities (β_k) that are specific to each group. The specific steps for the implementation are listed below (done state by state due to computational limitations):

1. Clustering

2. We first purge the wage data of state-year mean effects to make the data comparable across years.

3. Using a set of "stayers" pooled across all years in the sample, we summarize each firm based on 21 quantiles (p1, p5, p10, p15,..., p95, p99).

4. We then cluster the firms into 10 clusters via K-means (500 iterations, 17 random starts)

5. Creating the estimation sample

6. We focus on a set of movers who change firms for the first time, and have earnings data in periods $t-1$ and $t+1$.

7. We then randomly sample non-movers with the same data coverage in the same years (each control has a pseudo "move year")

8. Estimating the models

9. Using the estimation sample, we then start by estimating the linear model.

10. We then estimate the interacted model:

11. The worker covariates α_i are constructed using the estimates from the linear model.

12. We then instrument for the interacted covariates using a LIML instrumental variables approach. The instruments are constructed based on pre-move log earnings interacted with the group indicators.

Table A21, Panel A reports statistics related to the fit of the two models. We average the stats across all the states for which we have data access. The models have similar fit statistics, which suggests that complementarities are not playing a large role in our setting. To focus on new firms in our sample, we also compare the average β_k (based on the clusters into which each firm is sorted) for new firms and the other firms in the overall LEHD sample. We find that new firms exhibit higher complementarities with worker abilities, a difference that is statistically significant but not quantitatively large.

Finally, we validate the group-specific firm fixed effects estimates using descriptive event studies, following Card et al. (2013). Each firm in our sample is classified into quintiles for each worker demographic (female, Black, etc.), with quintile 5 representing firms with the largest estimated effects. Because of disclosure limits, we focus on workers who move to quintile 5 firms in each demographic group. Figure 6 presents the results. We see relatively flat trends in the two years before the move, and larger jumps for movers who move from low quintile firms. This provides some support for the reliability of the firm-demographic fixed effects approach.

Table A21: Non-linear versus linear models

Panel A: Model Fit Statistics

	R^2	$\frac{Var(\hat{\alpha})}{Var(y)}$	$\frac{Var(\hat{\psi})}{Var(y)}$	$\frac{2 \cdot Cov(\hat{\alpha},\hat{\psi})}{Var(y)}$	$\frac{Var(\epsilon)}{Var(y)}$	$Corr(\hat{\alpha},\hat{\psi})$
Interacted Model	0.8063	0.5466	0.09191	0.1723	0.1891	0.3867
Linear Model	0.8106	0.5425	0.09687	0.1713	0.1894	0.3758

Panel B: Comparison of Match Coefficients

	Estimated Slope for Firm's Cluster
New Firm	0.021***
	(0.000)
Constant	0.985***
	(0.000)
Observations	5,594,000

Notes: Panel A follows the methodology in Bonhomme et al. (2019) to compare linear versus interacted models, using ten clusters of firms for each state. Panel B compares the group-level estimated slope for new firms in our sample versus the broader sample of firms. Standard errors in parentheses and are clustered at the firm level. *$p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Source: authors' calculations based on U.S. Census data (FSRDC Project Number 2229, Disclosure Review Board (DRB) approval number: CBDRB-FY26-P2229-R12962).

C Extra Model Details

C.A Model Derivations

We present the firm's markdown problem and derive closed-form expressions for factor demands, wages, and profits, the entrepreneurial entry threshold, and the equilibrium conditions.

C.A.1 Firm's Problem and Factor Demands

An entrepreneur of type g with ability z takes the market shifters (D_m, D_f) as given and faces the inverse labor supply $w_{gh} = (\ell_{gh}/D_h)^{1/\varepsilon}$ for each worker type. The firm chooses $\{\ell_{gm}, \ell_{gf}\}$ to maximize profit,

$$\pi_g(z) = z\,\ell_g^{\rho} - \sum_h w_{gh}\,\ell_{gh} = z\,\ell_g^{\rho} - \sum_h D_h^{-1/\varepsilon}\,\ell_{gh}^{1+1/\varepsilon}.$$

The first-order conditions equate the marginal revenue product to the perceived marginal cost of labor, yielding the constant markdown

$$w_{gh} = \frac{\varepsilon}{1+\varepsilon}\,MRP_{gh}, \qquad MRP_{gh} = z\rho\,\ell_g^{\rho-1} A_{gh}^{1/\gamma} \ell_{gh}^{-1/\gamma} \ell_g^{1/\gamma}.$$

Defining the combined elasticity $b = \gamma\varepsilon/(\gamma+\varepsilon)$, $q = b(\rho - 1 + 1/\gamma)$, $M_{gh} = A_{gh}^{b/\gamma} D_h^{b/\varepsilon}$, $K_g = (z\rho/(1+1/\varepsilon))^b$, and $S_g = \sum_h A_{gh}^{1/\gamma} M_{gh}^{(\gamma-1)/\gamma}$, the first-order conditions imply the following equation

$$\ell_g(z) = \left[K_g\, S_g^{\gamma/(\gamma-1)}\right]^{\frac{1}{1-q}}, \qquad \ell_{gh}(z) = K_g\, M_{gh}\, \ell_g(z)^q, \qquad w_{gh}(z) = \left(\ell_{gh}(z)/D_h\right)^{1/\varepsilon}.$$

Because the CES labor bundle is homogeneous of degree one in labor inputs, Euler's theorem gives $\sum_h MRP_{gh}\, \ell_{gh} = \rho\, z\, \ell_g^\rho$. Using $MRP_{gh} = (1+1/\varepsilon)\, w_{gh}$, the total wage bill is $\sum_h w_{gh}\, \ell_{gh} = \frac{\rho}{1+1/\varepsilon}\, z\, \ell_g^\rho$, so variable profit has the closed form

$$\pi_g(z) = \left(1 - \frac{\rho}{1+1/\varepsilon}\right) z\, \ell_g(z)^\rho = \left(1 - \frac{\rho\varepsilon}{1+\varepsilon}\right) z\, \ell_g(z)^\rho.$$

If $\rho \in (0,1)$ and $\varepsilon > 0$, q is mechanically less than 1. Then, if we assume that each founder type can make productive use of at least one type of worker ($A_{gh} > 0$ for at least one worker type in each founder row), $\ell_g(z)$ and $\pi_g(z)$ are strictly increasing in z. Because profit is strictly increasing in z, whereas wages are fixed, this implies a single entry cutoff z_g^*.

C.A.2 Worker Value and Entry Threshold

The value of supplying labor for a type-h worker is the employment-weighted average wage actually paid to type h,

$$W_h = \frac{\sum_g N_g \phi_g \int_{z_g^*}^\infty w_{gh}(z)\, \ell_{gh}(z)\, dF(z)}{\sum_g N_g \phi_g \int_{z_g^*}^\infty \ell_{gh}(z)\, dF(z)}.$$

The worker value W_h is a monetary-income object: idiosyncratic workplace tastes generate the firm-level labor supply $\ell_{gh} = D_h\, w_{gh}^\varepsilon$ conditional on working, but are assumed not to enter the entrepreneurship-versus-work decision, which depends only on income. An individual of type g with access becomes an entrepreneur if and only if $\pi_g(z) \geq W_g$. Since $\pi_g(z)$ is increasing in z, this defines a cutoff z_g^*; individuals with $z \geq z_g^*$ and access (ϕ_g) become entrepreneurs, and all others supply labor as workers.

C.A.3 Equilibrium

An equilibrium is a set of aggregates (D_m, D_f, W_m, W_f) such that (i) each worker market clears,

$$\sum_{g\in\{m,f\}} N_g\, \phi_g \int_{z_g^*}^\infty \ell_{gh}(z)\, dF(z) = N_h\left(1 - \phi_h\left[1 - F(z_h^*)\right]\right), \quad h \in \{m, f\},$$

where the left-hand side is aggregate demand for type-h workers and the right-hand side is the supply (population net of type-h entrepreneurs), and (ii) the worker values (W_m, W_f) equal the employment-weighted average wages implied by the firm-level wages above.

C.B Details on Model Computation

We solve for the equilibrium under any given parameter guess by a damped fixed-point iteration on the four aggregates (D_m, D_f, W_m, W_f). Because the entry margin is a step function of the worker values, this is more stable than a Jacobian-based solver. Starting from the competitive solution, we iterate:

1. Given the current (D_m, D_f, W_m, W_f), compute each potential entrepreneur's profit $\pi_g(z)$ from the closed form above and the entry decision $\pi_g(z) \geq W_g$; then compute every active firm's factor demands and wages.

2. Update the supply shifters toward market clearing (aggregate demand for type h is increasing in D_h) and update the worker values toward the implied employment-weighted average wages.

3. Iterate until the maximum (fractional) excess demand and worker-value discrepancy fall below 10^{-7}, then compute the seven moments.

The results in Table 4 simulate a population of ten million majority individuals per group, with N_f/N_m times as many minority individuals. The robustness exercise in Appendix C.D, which documents patterns rather than headline levels, uses a population of one million majority individuals per group for the core joint founder-definition and parameter grid. Each cell recalibrates the five free parameters before repeating the counterfactual.

We then wrap an fminsearch call around this routine to find the parameter values that minimize the (scale-free) distance between model and data moments. The optimization uses MATLAB's Nelder-Mead simplex algorithm. The firm labor-supply elasticity ε is restricted to a bounded range and started near the value implied by the AKM differential $\beta_1 = [\log(A_{ff}) - \log(A_{fm}) - \log(A_{mf}) + \log(A_{mm})]/(\gamma + \varepsilon)$.

The parameter σ_z captures the variation in entrepreneurial ability. Within each run, we fix the draws ($\zeta \sim N(0,1)$) and then rescale them based on the guess for σ_z.

Table A22: Calibration Targets and Fitted Values

Moment	Statistic	Female	Black	Hispanic	Non-College
Minority worker share at minority-founded firms	Target	0.717	0.559	0.465	0.805
	Fitted	0.727	0.592	0.468	0.818
Minority worker share at majority-founded firms	Target	0.353	0.047	0.092	0.692
	Fitted	0.345	0.050	0.092	0.658
Minority founder share	Target	0.408	0.054	0.105	0.719
	Fitted	0.410	0.055	0.105	0.725
Minority-to-majority wage ratio	Target	0.800	0.750	0.730	0.600
	Fitted	0.806	0.732	0.731	0.597
AKM firm-effect differential	Target	0.129	0.068	0.057	0.049
	Fitted	0.127	0.068	0.057	0.048
Mean of log firm employment	Target	2.303	2.303	2.303	2.303
	Fitted	2.592	2.434	2.522	2.565
SD of log firm employment	Target	1.461	1.461	1.461	1.461
	Fitted	1.246	1.277	1.334	1.275
Group size ratio	Input	1.000	0.081	0.151	3.42

Notes: The employee-share targets are based on the pooled estimates from the balanced LEHD sample in Panel C of Table A3. The founder-share target and group size ratio are based on the summary statistics for founders and workers in Table 1. The wage-ratio targets are approximate values based on various government reports from the sample period. The AKM firm-effect differential is from Table 3. The mean and standard deviation targets for log firm employment are implied, under lognormality, by the firm-size summary statistics of the analysis sample (Table 1, Panel B). Fitted-value source: the completed calibration with rho = 0.725 and gamma = 4.

Table A22 summarizes the input data moments for all of the different runs.

C.C Local Sensitivity of the Calibrated Parameters

Following the approach in Andrews et al. (2017), we estimate how small changes in the empirical moments would change the calibrated parameters. Specifically, we vary each moment near the baseline value, and then recalibrate all five parameters to assess the joint response.

Let θ denote the five parameters to be calibrated, d the seven empirical moments, and $m(\theta)$ the value of the moments at parameter values θ. Let $J = \partial m(\widehat{\theta})/\partial\theta'$, which captures how the moment values change near the calibrated value $\widehat{\theta}$. Because we are minimizing the sum of squared proportional differences (rather than the square of absolute differences), there are implied weights in the objective function $W = \text{diag}(d_1^{-2}, \ldots, d_7^{-2})$. We keep these weights fixed throughout the calibration.

The matrix J then allows us to approximate how the parameters would adjust to a small change in the empirical targets:

$$\Delta\theta \simeq S\Delta d, \qquad S = (J'WJ)^{-1}J'W.$$

where Δd is the change in the targets and $\Delta\theta$ is the approximate change in the parameter estimates. For the results, we report the approximate percentage change in a parameter (relative to the baseline calibrated value) following a 1% increase in one moment. Figure A14 reports the results for all four calibrations. Repeating the calculation with step sizes of 0.5% and 2%, we find generally similar sensitivity patterns.

C.D Robustness of the Model Calibration

As noted in the main text, we assess the robustness of our counterfactual predictions to variation in the homophily moment, the returns-to-scale parameter ρ, and the worker-substitution elasticity γ.

For this exercise, we take the Cartesian product of the one-founder and three-founder Census homophily moments, $\rho \in \{0.65, 0.70, 0.75, 0.80\}$, and $\gamma \in \{2, 3.5, 5\}$. For each combination of input parameters and moments, we recalibrate the remaining free parameters, and then compute the access friction counterfactual.

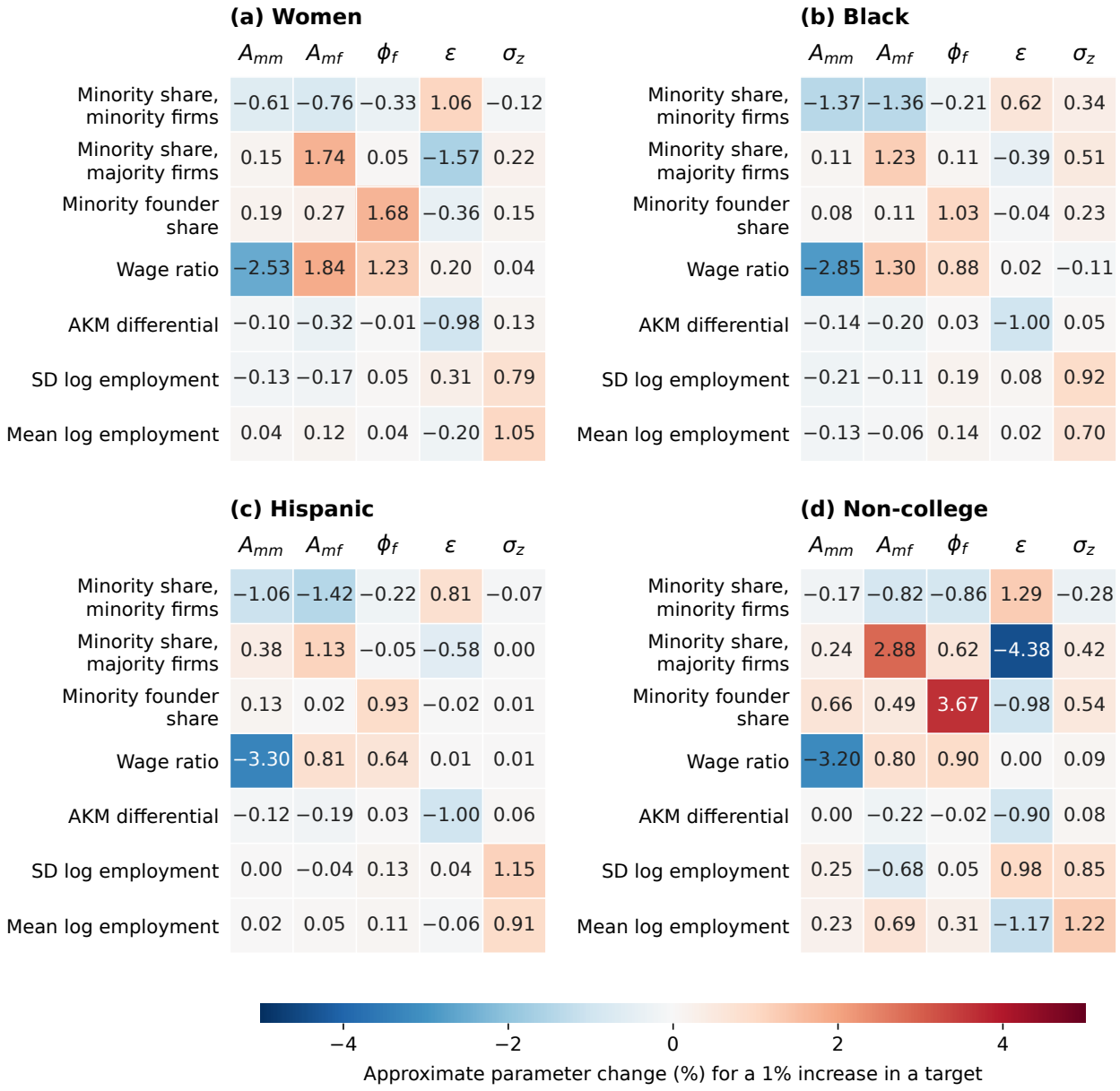


Figure A14: Local sensitivity of the calibrated parameters to the estimation moments

Notes: Each row represents the joint responses of the parameters to a 1 percent change in that row's moment. All panels use the same color scale, centered at zero.