College Graduates in the Labor Market: Geographic Mobility and Sorting into Firms and Occupations*

Chen Liu — NUS

Sibo Liu — HKPU

Yifan Zhang — CUHK

April 22, 2026

Abstract

This paper integrates data from three online labor market platforms—LinkedIn user profiles, Burning Glass job postings, and Glassdoor wages—to study how college graduates' job outcomes vary across universities of different rankings and locations, and analyze factors that determine the outcomes. Estimating a model of graduates' choices over firms, occupations, and locations, we find strong positive sorting: graduates from higher-ranked universities match to more cognitively intensive jobs and high-amenity cities. Distance reduces job-matching probabilities for most universities, whereas its impact is markedly attenuated for elite institutions. We also find a substantial geographic premium for universities in major cities.

Keywords: Geographic mobility, task complementarity, the geographic premium

JEL Codes: I23, J61, R23

*Chen Liu: Department of Economics, National University of Singapore, ecsliuc@nus.edu.sg; Sibo Liu: School of Accounting and Finance, Hong Kong Polytechnic University, sibo.liu@polyu.edu.hk; Yifan Zhang: Department of Economics, The Chinese University of Hong Kong, yifan.zhang@cuhk.edu.hk.

We thank participants at various seminars and conferences for helpful comments. Generous research support was received from the National University of Singapore and the Chinese University of Hong Kong. All remaining errors are our own.

1 Introduction

While it is well known that graduates from elite universities often receive stronger training and skill development that are sought by top employers (Dale and Krueger, 2002) and enjoy a substantial earnings premium (Chetty, Friedman, Saez, Turner and Yagan, 2020), many of the most prestigious U.S. universities are located far from the country's largest clusters of high-wage firms, desirable occupations, and major urban labor markets. The spatial divide between universities and employers suggests that graduates must be mobile to access elite firms and secure high-wage opportunities. As a result, differences in graduates' outcomes may reflect not only the quality of the university they attend, but also the geographic access that the university provides to labor-market opportunities. This creates a central question for understanding the earnings gap across colleges: to what extent are differences in graduates' early-career outcomes driven by university quality versus geography? More broadly, does geography amplify or dampen the labor-market advantage associated with elite universities?

Addressing these questions requires rich micro-level data. One would need to observe where the students studied, where they work, which firms employ them, which occupations they enter, and how those jobs differ in wages and task content. Most existing datasets do not jointly observe these dimensions. As a result, the literature has made substantial progress in understanding returns to college quality, worker sorting across firms, and geographic mobility, but we still know relatively little about how the sorting of college graduates into firms, occupations, and cities varies by university quality and location.

In this paper, we bring together three online labor market platforms—LinkedIn user profiles, Burning Glass Technologies (BGT), and Glassdoor—to construct a rich individual-level dataset. Using LinkedIn data, we created a snapshot of users working in the U.S. in 2018, whose profiles contain rich information on education and employment histories. This facilitates estimations that are consistent with other datasets, such as the 2018 American Community Survey. The BGT data contain job posting information with the tasks required by each employer and detailed job titles. Our Glassdoor data provide entry-level wages by employer and job title. We standardize information—including employer names, locations, and occupations—and construct a crosswalk to link these datasets. Using this newly assembled dataset, we estimate a unified framework where fresh college graduates sort into firms, occupations, and locations. We use the estimated model to study how job market outcomes differ across colleges with different rankings and locations, and account for multiple factors that contribute to the disparities in job outcomes across colleges.

Our sample includes more than 240 thousand fresh graduates who received their bachelor's degrees between 2016 and 2018, from 264 U.S. universities.[1] We show that our sample captures nearly 10% of all bachelor's degree recipients from these institutions, and that the key sample moments are strongly correlated with benchmark datasets such as the U.S. Census and the Integrated Postsecondary Education Data System (IPEDS). Compared to the commonly used matched employer-employee datasets, our dataset offers richer information: on the labor supply side, we observe the college each individual attended and their employer and occupation; on the demand side, we observe the tasks required by firm and occupation; and on the spatial dimension, we observe the geographic location of the college and the current location of work.

In our model, fresh college graduates choose a combination of a firm and an occupation, while recognizing the tasks performed, amenities of the city where the job is located, and mobility costs.[2] Our model implies a gravity-type equation of worker sorting into jobs, which regresses log employment shares on university and firm-occupation fixed effects. These fixed effects account for the availability of alternative employment options for each group, as well as firm-occupation wage efficiency units and local prices. We explain the residual variation using three sets of observable factors of interest.

The first is task complementarity, one of the most emphasized determinants of worker-to-firm sorting in theory (Eeckhout and Kircher, 2011). In our setting, it refers to the supermodular relationship between skills acquired at universities and the tasks demanded by firms. Positive sorting occurs if graduates from higher-ranked universities have a comparative advantage in working at more productive firms. Our rich dataset allows us to empirically measure university–job-specific productivity. Building on the literature that emphasizes the importance of education for skill acquisition (Hanushek and Kimko, 2000, Hanson and Liu, 2023) and the varying relevance of tasks across jobs (Acemoglu and Autor, 2011, Atalay et al., 2018), we model and estimate labor productivity using interactions between university rankings and four commonly used task requirements (cognitive, social, routine, and manual) obtained from BGT.

The second factor is the geographic separation between top universities and major employers, shaped by historical, economic, and institutional forces. Many of the

[1] We focus on these universities that are ranked in the World University Rankings (WUR). See Section 2.1 for sample construction in detail.

[2] We refer to a firm as the combination of a company name and the location of its establishment. Thus, establishments of the same company in different locations are treated as distinct firms in our analysis. For simplicity, we refer to a job as the combination of a firm and an occupation.

most prestigious universities were founded in the 18th and 19th centuries, often in small towns in the Northeast. The Morrill Land-Grant Act of 1862 granted federal land to states for the creation of public universities, intentionally situated outside major cities to better serve rural communities. Many of these universities later became flagship public universities in their respective states. However, economic activities and high-paying jobs shifted geographically due to technological innovation, westward territorial expansion, and population growth. Once concentrated in the Northeast during the colonial era, economic hubs eventually expanded into the Midwest during the industrial age, while the South and West rose dramatically in the late 20th century. We account for moving costs by incorporating interaction terms between university rankings and geographic distance. To account for potential nonlinear effects of distance and unobserved region-specific factors, we include indicators for whether a university and a job are located in the same city or the same state.

Third, city amenities may also influence job choice. If larger firms and high-paying jobs tend to locate in high-amenity cities, then amenities can contribute positively to attracting talents. To examine how graduates from universities of different rankings sort across cities based on amenities, we incorporate interaction terms between university rankings and city amenities.

Our estimates reveal several findings. First, we find strong evidence of positive sorting into cognitive tasks. Specifically, as cognitive task requirements become one standard deviation higher (the difference between working as a computer scientist between Google and Sanmina corporation, an American electronics manufacturing services), graduates from Top 20 universities have 0.200 higher in log points or about 20% more likely to match to the more cognitive-intensive job, relative to the benchmark group (outside Top 1000), holding other factors constant. The literature on worker-firm sorting often relies on the Abowd-Kramarz-Margolis (AKM) framework (Abowd et al., 1999, Card et al., 2013, Lopes de Melo, 2018, Song et al., 2019, Bonhomme et al., 2023). Complementing this literature, our findings provide empirical support for the theoretical framework of worker-firm matching developed by Eeckhout (2018). As our data incorporate rich task characteristics on the demand side, they allow us to analyze sorting without relying on wage variation among job switchers.

Our task variables are taken from Burning Glass Technologies (BGT), which were first used by Hershbein and Kahn (2018) to study labor demand adjustments during the Great Recession, and have been used to study the effect of technological change on earnings dynamics (Deming and Noray, 2020), and in explaining regional disparities in earnings (Atalay, Sotelo and Tannenbaum, 2024). Our dataset enables us to estimate task-based sorting at a much more granular level, capturing heterogeneity across both firms and occupations.[3]

Second, we find that geographic distance reduces the likelihood of a job match for most universities, but has no effect on the job choices of Top 20 universities. The distance effect for lower-ranked universities is sizable: for example, comparing graduates from Boston College and UC Davis—both ranked outside the Top 20 but within the Top 200 globally—the former is 0.209 log points higher or about 20.9% more likely to obtain a job in New York City than the latter. In contrast, no such difference exists between Harvard and Stanford graduates in terms of matching to the same job in New York City. The results hold when we additionally control for the interaction between these geographic variables and variation in the origin composition of student bodies, measured by the fraction of in-state enrollment.

Third, we find evidence of positive sorting on amenities: graduates from Top 20 universities, relative to the benchmark, are 0.068 higher in log points or about 6.8% more likely to choose a job in Seattle (99th percentile in amenities) than in Detroit (25th percentile in amenities). While prior studies have primarily examined spatial sorting and amenities between college and non-college workers (Diamond, 2016, Diamond and Gaubert, 2022), our findings complement this literature by highlighting positive sorting across graduates from different universities.

Because students are often willing to move long distances to attend elite private universities, they may be less locally attached and more geographically mobile upon graduation. In contrast, appropriations for public universities are often closely tied to in-state enrollment, resulting in a student body that is predominantly composed of in-state residents.[4] This raises the question of whether our results are driven by differences in the geographic composition of students across colleges.

To address this concern, we conduct two exercises. First, we re-estimate our model using multiple subsamples that exclude elite private universities or restrict the sample to individuals attending public universities. Second, we augment the baseline regression by interacting universities' in-state enrollment shares with geographic variables. Both exercises yield similar estimates, suggesting that our results are unlikely to be driven by the differences in the geographic composition of the student body.

[3]The Dictionary of Occupational Titles (DOT) and the Occupational Information Network (O*NET) both provide occupation-specific task content, which have been widely used to study worker sorting across occupations (Acemoglu and Autor, 2011, Yamaguchi, 2012, Deming, 2017).

[4]In Section 4.4, we show that even at highly ranked public universities, such as UC Berkeley, UCLA, and UIUC, more than 70% of students are in-state.

To provide insight into the relative importance of sorting into firms versus sorting into occupations, our model also yields an alternative gravity-type specification that relates the share of graduates from each university employed in a given U.S. firm—conditional on the same occupation—to factors emphasized above. This specification introduces additional controls for university–occupation fixed effects, further netting out the role of alternative employment options for each university–occupation pair and exploiting residual variation across firms within the same occupation. Comparing these estimates with those from the baseline, the evidence suggests that sorting along both dimensions—firms and occupations—is quantitatively important in shaping the overall positive sorting on cognitive tasks.

We show that our estimated results are not driven by potential confounding factors in four ways. First, we augment our model with college selectivity criteria, which proxy for students' average ability prior to entering college, rather than academic standards or the quality of education or institution reputation that are captured in college rankings. Second, we consider the university's field emphasis in education, the share of graduates in STEM majors. Third, we use different sources of university rankings or more disaggregated groups. Including these factors, our conclusions remain unchanged.

Fourth, we show that our results are not contaminated by university-to-firm network effects. Under common assumptions where networks are a function of historical university-to-job matching probabilities, we show that accounting for network effects does not alter the model specification but instead requires only a re-interpretation of the regression coefficients. Empirically, we also control for a network measure capturing university alumni exposure to specific firms and examine which sorting channel—task, amenity, or geography—the network primarily operates through. Our findings indicate that it operates mainly through the geographic channel, and we continue to find positive sorting on cognitive tasks or amenities.

Our estimated model informs the assignment of talents to jobs as a function of observable factors, providing a platform to quantify the contribution of each factor in shaping labor market outcomes. To this end, we conduct counterfactual exercises in which we shut down one factor at a time—task complementarity, amenities, and geography—and estimate the job-matching probabilities and earnings that graduates from each university would have experienced under each scenario.

We find that geographic factors have a more negative impact on graduates from higher-ranked colleges than on those from lower-ranked institutions—consistent with a spatial divide between top-ranked colleges and high-paying jobs. Across all colleges in our sample, geography narrows labor market disparities across colleges, especially by reducing the probability gap of matching to top-5%-paid jobs (by 6.3%), and also lowers the average earnings gap by 2.2%. In contrast, positive sorting on tasks and amenities—driven by the concentration of high-paying jobs in high-skill, high-amenity locations—widens earnings gap across colleges.

Finally, we estimate the geographic premium in annual earnings for each college. We find a substantial premium for institutions located in major cities, both in nominal and real terms, particularly in the Bay Area. Notably, this premium is sizable relative to the wage premium associated with university rankings. Relative to Midwestern and Southeastern cities that host major universities—such as Lafayette (LA), Ann Arbor (MI), and Bloomington (IN)—the Bay Area premium, in both nominal and real terms, amounts to roughly 20–40% of the premium associated with attending a Top 20 university.[5]

Our paper relates to the literature on labor market outcomes associated with attending elite colleges. Since information on the universities attended is rarely publicly available in large samples, Chetty, Friedman, Saez, Turner and Yagan (2020) link multiple administrative datasets to examine how attending elite universities shapes income segregation and inter-generational mobility. Chetty, Deming and Friedman (2023) show that attending an Ivy-Plus college instead of the average flagship public university triples their chances of working in a prestigious firm. Zimmerman (2019) shows that attending elite universities substantially increases mobility into top-paid jobs and raises income in Chile. Conzelmann et al. (2025) also use LinkedIn data to study how the geographic mobility of college graduates shapes intergenerational economic mobility and the social returns to public investment in higher education. Our analysis integrates and extensively processes data from three online labor market platforms, which provide rich information on firms, occupations, and job task requirements. Our paper complements early work by analyzing sorting into firms, occupations, and locations within a unified framework.

The implications of labor mobility on aggregate productivity have received increasing attention in developing countries (Lagakos and Waugh, 2013, Tombe and Zhu, 2019). Pellegrina and Sotelo (2021) study how the migration of farmers to western Brazil shaped regional comparative advantage in agriculture. In the United States, geographic mobility has been studied in response to the Great Recession (Cadena and

[5]These premiums are relative to institutions ranked outside the Top 1000.

Kovak, 2016) and import competition (Greenland et al., 2019, Autor et al., 2025), and in determining the aggregate productivity (Albert and Monras, 2022). We complement the literature to analyze patterns of geographic mobility across universities in different locations.

The paper is organized as follows. Section 2 briefly summarizes the data and presents motivating facts. Section 3 describes the model. Section 4 discusses the estimation results, and Section 5 tests potential confounders. Section 6 estimates the geographic premium. Section 7 concludes.

2 Data and Facts

This section describes our data sources and presents motivating evidence on the sorting patterns of college graduates.

2.1 Data Sources

Our data come from multiple sources, briefly described below. The complete details are provided in Appendix A.

LinkedIn. Our first data source comes from Revelio Labs, which processes LinkedIn profiles containing detailed résumés of individuals from employment records. The dataset provides rich self-reported information on the firms and job titles individuals have held, the institutions they attended, the degree awarded, fields of study, and the start and end dates of each job and degree. Some users report multiple universities for their undergraduate studies, which might involve exchange programs or institutions where they completed minors. We define the primary university as the institution where the individual spent the longest time. We created a 2018 snapshot by restricting our sample to 52 million LinkedIn users working in the U.S. in 2018.

Because we focus on job matching among fresh college graduates, we restrict the sample to individuals who earned a bachelor's degree (as their highest degree) between 2016 and 2018, graduated from a U.S. institution, and are currently employed by a U.S. firm.[6] We focus on fresh graduates who first enter the job market, as they are likely to maintain accurate and up-to-date information on their LinkedIn profiles. Our objective is to measure individuals' first "primary" job immediately after graduation.[7] We exclude individuals currently working as interns or enrolled in master's or Ph.D.

[6] We include 2016 graduates to increase the sample size. Most LinkedIn users report the same job during 2016-2018.

[7] By first "primary" job, we exclude jobs or work experience that users report as internships.

programs.

Burning Glass Technologies. Our second data source is the universe of job postings, which measures task content by occupation and employer.[8] We use BGT job posting data from 2018. Following Spitz-Oener (2006) and Atalay, Phongthiengtham, Sotelo and Tannenbaum (2020), we apply text analysis to construct four commonly used task measures from job advertisements: cognitive, social, routine, and manual. For each task, we compute percentile rankings across all postings and then average them across postings that share the same firm name and occupation code.[9] All task measures are standardized to take values between 0 and 1. In Section 4, we further standardize each BGT task variable by its standard deviation to ease interpretation.

Glassdoor. We obtain firm-occupation-specific wage information from Glassdoor.[10] Glassdoor is an online platform where users voluntarily and anonymously report wages and review employers. The wage information is derived from raw salary data submitted by platform users and aggregated by Glassdoor's internal algorithms.[11] Specifically, workers are incentivized to contribute through a "give-to-get" policy, where users must share their own data to access information provided by others. This arrangement improves the reliability of the wage data, which we also validate later.

We obtain a snapshot of Glassdoor data collected between September and October 2024, which includes detailed wage information by firm, occupation, location, and years of experience. To integrate this dataset with our other sources, we match job titles to Standard Occupational Classification (OCCSOC) occupation codes, firm names to those in Burning Glass and LinkedIn, and job locations to commuting zones (hereafter, CZ).

We use entry-level wages (0–3 years of experience), which likely reflect the earnings of recent graduates. In our Glassdoor data, 16% of entry-level wages are reported as exact values, while 84% are reported as intervals. Importantly, since these reported intervals are generally narrow among entry-level wages, we choose the midpoint as the

[8]Notably, the BGT data are based on online job postings and are available beginning in 2007. For earlier periods, see Atalay, Phongthiengtham, Sotelo and Tannenbaum (2020), who measure job tasks using newspaper postings dating back to 1960.

[9]Because many postings list multiple possible locations with identical task descriptions, our BGT measures do not capture variation across locations within the same firm–occupation combination, in order to maintain consistency.

[10]In a recent study, Martellini et al. (2024) use an individual-level sample of Glassdoor data to estimate college quality and assess its role in explaining cross-country variation in entrepreneurship and innovation. By contrast, our data are accessed directly from the public Glassdoor platform and are less granular, containing wage information at the firm–occupation–experience level.

[11]Importantly, we do not use wage information derived from job postings, which contain incomplete and biased wage information, as documented by Batra et al. (2023).

wage for each job.[12] We deflate wages to 2018 dollars using the Consumer Price Index (CPI) from the Bureau of Labor Statistics.

Amenity Index. Our city amenity measure uses a single index taken from Diamond (2016). The measure is based on a collection of rich amenity variables from six different categories: the retail environment, transportation infrastructure, crime, environmental quality, school quality, and local skill demand. The single amenity index is calculated as the first component of the principal component analysis. We convert it into a percentile ranking, normalized to range from 0 to 1.

University Rankings. We group universities into groups based on the World University Rankings (WUR), which is widely recognized as a proxy-based ranking and has been recently used in Martellini et al. (2024). The ranking is based on factors such as academic reputation, employer reputation, faculty-student ratio, and citations per faculty, with more weight given to the latter two factors. We also use rankings from U.S. News as an alternative measure.

2.2 Data Processing

We extensively process data from the three online labor market platforms, as detailed in Appendix A.2. In brief, our work involves four main tasks. First, we standardize employer names across the three datasets. Second, for LinkedIn and BGT, the Standard Occupational Classification (OCCSOC) codes are directly available. We use a large language model (ChatGPT-4o) to map job titles in Glassdoor to OCCSOC code. In our analysis, we aggregate the detailed OCCSOC occupations into 22 broad categories based on their first two-digit codes. Third, for LinkedIn data, we standardize self-reported university names to align with institutional names in the World University Rankings and U.S. News. Finally, we process geographic information from LinkedIn profiles to construct commuting zone codes, identifying both the location of universities and their place of work.

Putting it all together, we restrict the sample to LinkedIn users who: (1) received their bachelor's degree (as their highest degree) between 2016 and 2018 from a U.S. institution; (2) were working in the United States in 2018; (3) have employer names and occupational titles that are clearly identified and can be matched to BGT data; (4) have identifiable employer geographic locations; and (5) their home institutions are

[12] For jobs with salaries reported in intervals, we calculate the interval range relative to the midpoint. For example, for an interval $[a, b]$, we compute $\frac{b-a}{(a+b)/2}$. On average, this ratio is 11%, with a maximum of 20%, indicating that reported entry-level wages fall within a relatively narrow range.

ranked in the WUR.[13] To improve estimation precision, we further restrict the sample to U.S. universities with at least 100 LinkedIn users who meet conditions (1)-(5).

The sample covers 264 universities, 25246 distinct firms, and a total of 244,632 LinkedIn users.[14]

Throughout the paper, we define a firm as the combination of a company name and the location of its establishment. A firm appears in our sample if we observe at least one LinkedIn user employed there. We group universities into four tiers using the WUR: the Top 20 globally ranked (Top 20), those ranked 21–200 globally (Top 21–200), those ranked 201–1000 globally (Top 201–1000), and all others (including those ranked outside the Top 1000). In robustness checks, we also use more detailed grouping.

According to IPEDS, the 264 universities in our study awarded 2.52 million bachelor's degrees (including international students) between 2016 and 2018, implying that our sample covers approximately 10% of this population. Because some graduates do not enter the labor force—for example, by pursuing further study, remaining unemployed, or leaving the U.S.—our sample represents more than 10% of those who do enter the U.S. labor force.

2.3 Sample Validation

Since the BGT data have been extensively used and validated in prior studies (Hershbein and Kahn, 2018, Atalay, Sotelo and Tannenbaum, 2024), we focus on validating our LinkedIn and Glassdoor samples. Below, we briefly describe the six validation exercises we performed, with full details provided in Appendix B.

We validate the LinkedIn data in three ways. First, we assess spatial representativeness by comparing each CZ's share of national college-graduate employment between LinkedIn and the American Community Survey (ACS). Second, we assess occupational representativeness by comparing employment shares across two-digit SOC occupations between LinkedIn and the ACS. In both cases, correlations exceed 0.9, with OLS regression slopes close to one and R^2 values above 0.9. Third, we evaluate the representativeness of graduating class sizes across U.S. universities by comparing

[13] WUR ranks the top 2,000 universities globally. Among them, 348 are U.S. universities.

[14] We begin with 52.66 million LinkedIn users who report working in the United States (with a listed employer) in 2018. Restricting the sample to individuals who obtained a bachelor's degree in the United States reduces the sample to 21.26 million users. We then restrict to users whose employers can be matched to Burning Glass Technologies (BGT) firm names and locations, reducing the sample to 13.75 million. Next, limiting the sample to individuals whose colleges are ranked in the World University Rankings (WUR) further reduces the sample to 3.57 million. Among these, restricting to individuals whose highest degree is a bachelor's degree yields 3.18 million observations. Of these, 427K individuals received their bachelor's degree between 2016 and 2018. Finally, restricting to universities with at least 100 LinkedIn users in the matched sample yields a final sample of 244K observations.

the national share of graduates by university between LinkedIn and IPEDS. These exercises show that LinkedIn data are broadly representative of the U.S. college-educated workforce and graduating class sizes.

We conduct extensive validation to show that Glassdoor wage data provide meaningful information for college graduates.[15] Specifically, we compare average annual wages from Glassdoor and the ACS across occupations, commuting zones (CZs), and CZ–occupation pairs. In all cases, we find strong correlations, indicating that Glassdoor wages capture meaningful variation across regions and occupations.

2.4 Empirical Facts

This section presents motivating facts on the spatial distribution of universities and jobs, as well as the sorting patterns of recent college graduates.

Fact 1. Top U.S. universities and high-paying jobs are not clustered in the same locations. To compare the geographic distribution of top U.S. college graduates (supply) with that of high-paying jobs (demand), Figure 1 plots the relationship between each city's share of top-paid U.S. jobs (y-axis) and its share of locally educated bachelor's degree recipients from top U.S. universities (x-axis), each expressed as a share of the national total. We use data from the 2018 American Community Survey (Ruggles et al., 2010) to define top-paid jobs as those that belong to the top 5% of the income distribution among all college-educated wage earners. To measure the supply, we use the World University Rankings (WUR) to define top U.S. universities as those ranked among the Top 20 globally or belonging to the Ivy League. The number of bachelor's degree recipients from each top university is obtained from the IPEDS. The dashed line in the figure represents the 45-degree line.

Most cities are clustered either near the y-axis or the x-axis. For clarity, we restrict the plot to cities that either account for more than 1% of the top U.S. jobs or host at least one top-20 university. On the one hand, cities such as Arlington (Virginia), Seattle, Dallas, Houston, and Atlanta lie near the y-axis. These cities host a sizable share of high-paying jobs, but produce few top college graduates. Similarly, New York City and Newark are located close to the y-axis, reflecting that they host a relatively large share of top jobs compared with the share of top graduates produced locally.

On the other hand, several cities near the x-axis—such as Ann Arbor (MI), Champaign (IL), and Ithaca (NY)—are home to prominent universities but offer relatively

[15] As a recent study documents that online job postings contain little wage information (Batra, Michaud and Mongey, 2023), we do not use wage information from BGT in our analysis.

fewer high-paying job opportunities. In addition, many cities fall between the 45-degree line and the x-axis, representing locations that produce a disproportionately large share of top college graduates while hosting a smaller share of top jobs. Similar patterns persist when top-paid jobs are defined using the 75th or 90th percentile of the income distribution (see Appendix Figure E.1).

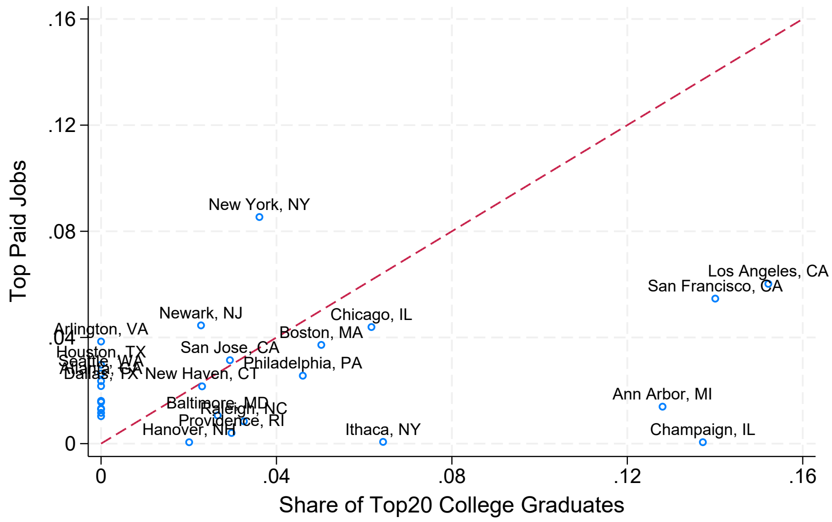

Figure 1: Spatial Distribution of Top College Graduates and Top Paid Jobs.

Notes: The y-axis shows a city's share of top jobs in the United States, where top jobs are defined as those in the top 5% of the income distribution among all college-educated wage earners. The x-axis shows the city's share of college graduates in the national total, focusing on graduates from universities ranked in the global Top 20 or belonging to the Ivy League.

This geographic separation between the "best jobs" and the "best talents" suggests that college graduates must move in order to access elite firms and better job opportunities. Since some facts on mobility have been previously documented in Conzelmann et al. (2025) while using a different sample, we briefly state them here and leave the full details in Appendix C.

Fresh graduates are highly mobile across cities, with the majority moving within the same state or to neighboring states. Specifically, Appendix Table C.1 shows 47.7% stay and work in the same CZ, and 70% of graduates are employed within the same state. In addition, graduates from higher-ranked universities are more mobile, with the most notable group being graduates from the Top 20 universities. For graduates from outside the Top 1,000 universities, 56.1% remain in the same CZ and 80.0% remain in the same state as their college, compared with 37.6% and 55.1%, respectively, for graduates from Top 20 universities.

Fact 2. High-wage cities retain a larger share of their locally educated graduates than do low-wage cities.

We regress CZ-level retention rates on indicators for college rank—Top 20, Top 21–200, and Top 201-1000—as well as the log of average weekly wages in the university's host city. CZ-level retention rates for each college are measured as the share of graduates who remain in the same CZ as their college to work. Universities outside the Top 1000 serve as the benchmark group. We use the IPEDS to compute the share of undergraduate students who are in-state enrollees. Controlling for this variable accounts for the origin composition of student bodies, thereby addressing potential home bias in location choice (Kennan and Walker, 2011). In addition, we include state fixed effects to absorb cross-state differences in employment opportunities for college graduates. Eight out of 264 universities that are the sole institution (in our sample) within their state are automatically excluded.

Table 1: Retention Rates by College Rankings

	(1) Retention Rate	(2) Retention Rate	(3) Distance	(4) Distance
Top20	-0.133**	-0.125***	286.407***	296.630***
	(0.058)	(0.039)	(35.801)	(45.969)
Top21-200	-0.080**	-0.057*	107.908***	64.172*
	(0.039)	(0.032)	(24.565)	(37.441)
Top201-1000	-0.073**	-0.029	40.627**	17.208
	(0.030)	(0.025)	(18.920)	(29.581)
Log wage	1.418***		-397.902***	
	(0.107)		(66.293)	
In-state Enrollment	0.412***	0.409***	-509.098***	-478.842***
	(0.056)	(0.042)	(35.068)	(49.090)
Observations	256	152	256	152
R^2	0.57	0.85	0.73	0.80

Notes: The dependent variable is within-CZ retention rate in Columns (1) and (2), and average distance traveled in Columns (3) and (4). Columns (1) and (3) control for the log of average wages in the university-hosting city and for state fixed effects. Columns (2) and (4) include CZ fixed effects.

Column (1) of Table 1 reports a coefficient of -0.133 (s.e. = 0.058) on the Top 20 dummy, indicating that graduates from Top 20 universities are 13.3 percentage points (ppts) less likely to remain in the same city compared with graduates from universities outside the Top 1000. In addition, the in-state enrollment share also matters for retention rates: a 10 ppts increase in the share of in-state students is associated with a 4.12 ppts higher retention rate.

The estimated wage coefficient is positive, at 1.418 (s.e. = 0.107), suggesting that higher wages increase the likelihood that cities retain locally educated college graduates. The magnitude of the wage effect on retention rate is large: comparing Boston (among the top 10 highest-paid U.S. cities) with Pittsburgh (PA), the former has a 34.0 ppts higher retention rate for its college graduates than the latter.[16] In Column (2), we replace the log wage of the university-hosting CZ and state fixed effects with CZ fixed effects. The CZ fixed effects capture all unobserved, CZ-specific factors that influence employment outcomes.[17] We find similar results under this specification.

In Columns (3) and (4), the outcome variable is replaced with the distance traveled between the college city and the job city. Compared to the benchmark group, the Top 20 university graduates travel an additional 286 miles on average to secure employment. Graduates from Top 21–200 universities travel more than 108 miles farther than the benchmark group. Column (4), again, reports similar estimates for rank dummies when including CZ fixed effects.

Fact 3. There is a mover premium: movers, compared with stayers, tend to relocate to access better employment opportunities and higher amenities. The mover premium is high for universities located in low-wage cities and low for universities located in high-wage cities.

Using individual-level data, we regress the cognitive, social tasks (BGT), and log wages (Glassdoor) on a mover dummy, which equals one if an individual's CZ of work differs from the CZ of his/her university, and zero otherwise. For all regressions, we control for university fixed effects, and we compare movers and stayers within the same university.

Panel A of Table 2, Columns (1) and (2), shows that movers tend to obtain jobs with task requirements 3.0 percentiles higher in cognitive and 2.2 percentiles higher in social task intensity, compared to stayers. Because occupations or jobs that are cognitively and socially intensive often pay higher wages (Deming, 2017, Atalay et al., 2024), we consider better jobs those with high cognitive or social task requirements. Column (3) shows that movers also tend to land in jobs that pay significantly higher, about 13.8%, relative to stayers. Column (4) shows that movers settle in cities with local amenities 7.3 percentiles higher than stayers. The evidence points to the existence of a compensating differential: migration is indeed costly, and college graduates tend to move farther to secure better employment opportunities or to access high-amenity

[16]The value is computed as $1.418\times(7.52-7.28)$, where 7.52 and 7.28 are the log weekly college wages for Boston and Pittsburgh (PA), respectively.

[17]Because we condition on CZ-level fixed effects, we compare universities of different rankings located within the same CZ, and universities that are the sole institution within their CZ are excluded.

locations.

Because task variables and wages vary across firms and occupations, we also report estimates that condition on university and occupation fixed effects. When comparing movers and stayers from the same university who enter the same occupation but different firms, the mover premium remains statistically significant (albeit attenuated) for social tasks and wages, while becoming imprecisely estimated for cognitive tasks; see Columns (5)–(7).

Table 2: Job Outcomes of Movers and Stayers

	(1) Cognitive	(2) Social	(3) Log Wages	(4) Amenity	(5) Cognitive	(6) Social	(7) Log Wages
Panel A: The Mover Premium							
Movers	0.030***	0.022***	0.138***	0.073***	0.008	0.011***	0.078***
	(0.010)	(0.007)	(0.021)	(0.006)	(0.007)	(0.003)	(0.014)
Occupation FE					✓	✓	✓
Observations	244,234	244,234	244,154	242,798	244,234	244,234	244,154
R^2	0.07	0.03	0.13	0.42	0.29	0.19	0.56
Panel B: The Mover Premium by University Hosting Cities							
Movers	0.032***	0.024***	0.147***	0.086***	0.009	0.013***	0.085***
	(0.011)	(0.008)	(0.023)	(0.008)	(0.007)	(0.004)	(0.015)
Mover $\times \ln Wage_g^{\text{Host}}$	-0.090**	-0.097***	-0.417***	-0.557***	-0.051**	-0.066***	-0.310***
	(0.036)	(0.018)	(0.064)	(0.018)	(0.025)	(0.015)	(0.054)
Occupation FE					✓	✓	✓
Observations	244,234	244,234	244,154	242,798	244,234	244,234	244,154
R^2	0.07	0.03	0.14	0.43	0.29	0.20	0.56

Notes: All regressions are estimated using individual LinkedIn users and include controls for university fixed effects. Columns (1) to (4) include university fixed effects. Columns (5) to (7) include university and occupation fixed effects. The dependent variable in column (3) is log wages, measured using Glassdoor data for the firm–occupation in which a LinkedIn user is employed. $\ln Wage_g^{\text{Host}}$ is the demeaned log weekly wage in university-hosting cities obtained from the ACS. Standard errors in parentheses are clustered at the firm-occupation level.

Mover premium depends on job opportunities available in university-hosting cities. We augment the model by including an interaction between the mover dummy and $\ln Wage_g^{\text{Host}}$, the demeaned log weekly wage in university-hosting cities obtained from the ACS.[18] We find negative and precisely estimated interaction coefficients that are both sizable and economically meaningful across all columns. For example, hosting cities with a 0.1 log-point lower average log wage are associated with a 0.9-percentile higher mover premium for cognitive tasks (Column 1), a 0.97-percentile higher premium for social tasks (Column 2), and a 4.17% higher wage premium (Column 3).

Motivated by these documented patterns, we next estimate a model of how gradu-

[18]We demean the log wage to ease interpretation of the main effect: in Panel B, the coefficient on the mover dummy captures the mover premium for cities at the sample mean of the log wage.

ates sort into firms, cities, and occupations, allowing a comprehensive assessment of the relative importance of multiple factors in shaping labor market sorting.

3 The Model

We present a model where college graduates decide which firm (indexed by f) and occupation (indexed by o) to pursue, taking into account wages, search costs, the amenities of the city (c) where the job is located, and individual tastes. We group workers by their university (indexed by g). Again, a firm refers to the combination of an employer name and the location of its establishment. Since the firm index f and the university index g already nest information in the location information for jobs and schools, respectively, we omit the city subscript c when its exclusion does not create confusion.

3.1 Preference

For individuals who attend university g, their utility of living and working in city c, firm f and occupation o is

$$U_{fo}^{i} = \left(\frac{C_{fo}^{g}}{1-\kappa}\right)^{1-\kappa}\left(\frac{H_{fo}^{g}}{\kappa}\right)^{\kappa} a_c^g \tau_{fo}^g \varepsilon_{fo}^i. \tag{1}$$

In Equation (1), C_{fo}^g represents consumption of tradable goods, and H_{fo}^g represents consumption of non-tradable goods. κ denotes the expenditure share allocated to non-tradables. a_c^g captures the utility obtained from the local amenity in city c felt by group g. τ_{fo}^g is the cost incurred during job searching, specific to university g graduates to firm f and occupation o. To ease interpretation, we model search cost as an iceberg cost (or take-home utility) with a higher value indicating lower costs. ε_{fo}^i is idiosyncratic preferences for job $\{f, o\}$ which allows workers to have heterogeneous preferences over the work environments of different potential employers and occupations.[19]

Individuals face the following budget constraints $P_c C_{fo}^g + R_c H_{fo}^g \leqslant W_{fo}^g$, where P_c is the price for tradables and R_c is the rent for non-tradables. W_{fo}^g is the wage that g-workers would earn from job $\{f, o\}$. Utility optimization implies that the demand is

$$C_{fo}^g = (1-\kappa)\frac{W_{fo}^g}{P_c}, \qquad H_{fo}^g = \kappa\frac{W_{fo}^g}{R_c}. \tag{2}$$

[19] In our model, students from the same university differ only in idiosyncratic tastes, ε_{fo}^i. We analyze sorting across universities of different rankings, while remaining silent on heterogeneity within each university.

We can then obtain the indirect utility as

$$V^i_{fo} = \frac{W^g_{fo}}{p_c} \tau^g_{fo} a^g_c \varepsilon^i_{fo}, \tag{3}$$

where $p_c = P_c^{1-\kappa} R_c^{\kappa}$ is the price index in city c.

3.2 Labor Allocation

For tractability, we assume ε^i_{fo} is drawn from i.i.d. *Fréchet* distribution with shape parameter θ and scale parameter 1. In equilibrium, the fraction of g-graduates who choose firm f and o can be expressed as

$$\Pi^g_{fo} = \frac{\left(W^g_{fo} a^g_c \tau^g_{fo} / p_c\right)^\theta}{\sum_{f'o'} \left(W^g_{f'o'} a^g_{c'} \tau^g_{f'o'} / p_{c'}\right)^\theta}. \tag{4}$$

By applying the Law of Conditional Probability, we can also derive the fraction of g-graduates who choose firm f, conditional on choosing occupation o as

$$\Pi^g_{f|o} = \frac{\left(W^g_{fo} a^g_c \tau^g_{fo} / p_c\right)^\theta}{\sum_{f'} \left(W^g_{f'o'} a^g_{c'} \tau^g_{f'o} / p_{c'}\right)^\theta}. \tag{5}$$

3.3 Estimating Equations

We assume that the wage has two components as follows:

$$W^g_{fo} = \omega_{fo} \times T^g_{fo}. \tag{6}$$

The first is a firm-occupation-specific component, ω_{fo}, capturing wage per efficiency unit paid at f-o, common across groups. The second is T^g_{fo}, capturing labor productivity (or efficiency units) for g-group if working at f and o. We also consider search cost τ^g_{fo} consist of two parts as follows

$$\tau^g_{fo} = \tau^{g,\text{Geo}}_c \times \tau^{g,\text{UBV}}_{fo}. \tag{7}$$

Here, $\tau^{g,\text{Geo}}_c$ is the component related to moving costs, which we will measure as a function of geographic variables. $\tau^{g,\text{UBV}}_{fo}$ is the unobserved component, highlighted using superscript UBV.

Sorting into Jobs. We can use Equations (4), (6), and (7) to derive a log-linear estimation equation for the determination of the log share of g-group who choose firm f and occupation o,

$$\ln \Pi^g_{fo} = \theta \ln T^g_{fo} + \theta \ln a^g_c + \theta \ln \tau^{g,\text{Geo}}_c + \lambda_g + \lambda_{fo} + \theta \ln \tau^{g,\text{UBV}}_{fo}. \tag{8}$$

The term $\lambda_g = -\ln \sum_{f'o'} \left(W^g_{f'o'} a^g_{f'} \tau^g_{f'o'}/p_{c'}\right)^\theta$ is group fixed effects that capture overall employment opportunities for g-workers. This term will also absorb any reputation effects that are specific to university g and common across firms and occupations.

The term $\lambda_{fo} = \theta \ln \omega_{fo} - \theta \ln p_c$ is the firm-occupation-specific fixed effects. It absorbs the wage per efficiency unit. Since index f nests city c, it also absorbs the local price index, p_c. $\theta \ln \tau^{g,\text{UBV}}_{fo}$ will be treated as the structural residual in the estimation.

Sorting across firms conditional on occupation. Equation (5) can be used to derive a log-linear estimating equation for the determination of the log share of g-group who choose firm f, conditional on occupation o,

$$\ln \Pi^g_{f|o} = \theta \ln T^g_{fo} + \theta \ln a^g_c + \theta \ln \tau^{g,\text{Geo}}_c + \lambda_{go} + \lambda_{fo} + \theta \ln \tau^{g,\text{UBV}}_{fo}. \tag{9}$$

The only difference from Equation (8) is $\lambda_{go} = -\theta \ln \sum_{f'} \left(W^g_{f'o'} a^g_{f'} \tau^g_{f'o}/p_{c'}\right)$, which is a group-occupation fixed effect absorbing the average employment opportunities conditional on a specific group and an occupation. Note that this term will also absorb any reputation effects that are specific to university g and occupation o, but common across firms.

3.4 Parameterization and Testable Implications

To estimate the model, we impose parametric assumptions on $\ln T^g_{fo}$, $\ln a^g_c$, and $\ln \tau^g_c$.

Labor Productivity (Task Complementarity). T^g_{fo} is specific to each university, firm, and occupation, and is high-dimensional. Our unique dataset enables us to reduce this dimensionality. Specifically, we assume university-job-specific labor productivity as follows:

$$\ln T^g_{fo} = \sum_{j=1}^{J} \sum_{k=1}^{K} \beta^{\text{BGT}}_{jk} X^g_j Y^{\text{BGT}}_{fok}, \tag{10}$$

where j indexes for the four university tier groups. $X^g = \{X^g_1, ..., X^g_4\}$ is therefore a set of binary variables representing these groups. $Y^{\text{BGT}}_{fo} = \{Y^{\text{BGT}}_{fo1}, ..., Y^{\text{BGT}}_{foK}\}$ represents the BGT task requirements that are specific to firm f and occupation o. k indexes for the type of tasks, which include cognitive, social, routine, and manual tasks.

Taking cognitive tasks as an example, $\beta^{\text{BGT}}_{j,\text{cog}} \times Y^{\text{BGT}}_{fo,\text{cog}}$ can be considered as the marginal productivity for tier-j graduates of performing cognitive tasks in firm f and occupation o. Suppose g_1 is the group of better-ranked universities, and g_4 is the group of lowest-ranked universities. When $\beta^{\text{BGT}}_{1,\text{cog}} > \beta^{\text{BGT}}_{2,\text{cog}} > \beta^{\text{BGT}}_{3,\text{cog}} > \beta^{\text{BGT}}_{4,\text{cog}} > 0$, it then implies that labor productivity increases as cognitive task requirements are higher, and increases more for graduates from better-ranked universities. The specification in Equation (10) is consistent with a micro-founded task complementarity (or supermodularity) framework, which gives graduates from better-ranked universities a comparative advantage in matching with higher-quality firms and generates positive assortative matching (Becker, 1973, Eeckhout, 2018).[20]

Positive sorting in task space has been formulated in theory (Costinot and Vogel, 2010), and extended to the probabilistic version that satisfies the Monotone Likelihood Ratio Property (Costinot and Vogel, 2015). In our setting, positive sorting in cognitive-intensive jobs means graduates from better-ranked universities are more likely to work in cognitive-intensive jobs

$$\frac{\partial \ln \Pi^{g_1}_{fo}}{\partial Y^{\text{BGT}}_{fo,\text{cog}}} > \frac{\partial \ln \Pi^{g_2}_{fo}}{\partial Y^{\text{BGT}}_{fo,\text{cog}}} > \frac{\partial \ln \Pi^{g_3}_{fo}}{\partial Y^{\text{BGT}}_{fo,\text{cog}}} > \frac{\partial \ln \Pi^{g_4}_{fo}}{\partial Y^{\text{BGT}}_{fo,\text{cog}}}. \tag{11}$$

Amenity. We measure the group-specific utility derived from local amenities as

$$\ln a^g_c = \sum_{j=1}^{J} \beta^{\text{Amen}}_j X^g_j Y^{\text{Amen}}_c, \tag{12}$$

where Y^{Amen}_c is the measure of local amenity. $\beta^{\text{Amen}}_j X^g_j$ captures the marginal utility to amenity for tier-j graduates. $\beta^{\text{Amen}}_1 > \beta^{\text{Amen}}_2 > \beta^{\text{Amen}}_3 > \beta^{\text{Amen}}_4$ implies that graduates from better-ranked universities value local amenities more. Similarly, positive sorting in amenity means graduates from better-ranked universities are more likely to work in high-amenity cities

$$\frac{\partial \ln \Pi^{g_1}_{fo}}{\partial Y^{\text{Amen}}_c} > \frac{\partial \ln \Pi^{g_2}_{fo}}{\partial Y^{\text{Amen}}_c} > \frac{\partial \ln \Pi^{g_3}_{fo}}{\partial Y^{\text{Amen}}_c} > \frac{\partial \ln \Pi^{g_4}_{fo}}{\partial Y^{\text{Amen}}_c}. \tag{13}$$

Mobility costs. We measure group-city-specific mobility costs using the interaction

[20] Since we include interaction terms for four types of tasks, our estimation will reveal task complementarity for each of the four dimensions.

between X^g and geographic variables, Y^{Geo}_{fgk}. Specifically, we assume

$$\ln \tau_c^{g,\text{Geo}} = \sum_{j=1}^{J}\sum_{k=1}^{K} \beta^{\text{Geo}}_{jk} X^g_j Y^{\text{Geo}}_{fgk}, \tag{14}$$

where Y^{Geo}_{fgk} include three variables, the logarithm of the geographic distance between university and firm, a dummy variable CZ_{fg} that equals one if university g and the firm f are in the same city; and a dummy $State_{fg}$ that equals one if the university and firm are in the same state. These dummies can pick up any non-linear features between costs and log distance. Similarly, $\beta^{\text{Geo}}_{jk} X^g_j$ captures the marginal utility of geographic barriers for tier-j graduates. These interactions pick up potential differential impacts of geography on job choice across groups. For example, positive sorting in distance means graduates from better-ranked universities are more likely to match to distant jobs

$$\frac{\partial \ln \Pi^{g_1}_{fo}}{\partial Y^{\text{Geo}}_{fg,\text{dis}}} > \frac{\partial \ln \Pi^{g_2}_{fo}}{\partial Y^{\text{Geo}}_{fg,\text{dis}}} > \frac{\partial \ln \Pi^{g_3}_{fo}}{\partial Y^{\text{Geo}}_{fg,\text{dis}}} > \frac{\partial \ln \Pi^{g_4}_{fo}}{\partial Y^{\text{Geo}}_{fg,\text{dis}}}. \tag{15}$$

Now we can rewrite Equation (8) in estimable interaction terms as follows

$$\begin{aligned}\ln \Pi^g_{fo} = &\sum_{j=1}^{J}\sum_{k=1}^{K} \widetilde{\beta}^{\text{BGT}}_{jk} X^g_j Y^{\text{BGT}}_{fok} + \sum_{j=1}^{J} \widetilde{\beta}^{\text{Amen}}_{j} X^g_j Y^{\text{Amen}}_{c} + \sum_{j=1}^{J}\sum_{k=1}^{K} \widetilde{\beta}^{\text{Geo}}_{jk} X^g_j Y^{\text{Geo}}_{fgk} \\ &+ \lambda_g + \lambda_{fo} + \theta \ln \tau^{g,\text{UBV}}_{fo}.\end{aligned} \tag{16}$$

Two points are worth mentioning. First, we denote $\widetilde{\beta}_{jk} = \theta\beta_{jk}$. Our estimation only recovers the reduced-form parameter $\tilde{\beta}_{jk}$, and does not separately identify θ and β_{jk}.

Second, because $\frac{\partial \ln \Pi^{g_j}_{fo}}{\partial Y^{\text{BGT}}_{fo,\text{cog}}} = \widetilde{\beta}^{\text{BGT}}_{j,\text{cog}}$, $\frac{\partial \ln \Pi^{g_j}_{fo}}{\partial Y^{\text{Amen}}_{c}} = \widetilde{\beta}^{\text{Amen}}_{j}$, and $\frac{\partial \ln \Pi^{g_j}_{fo}}{\partial Y^{\text{Geo}}_{jk}} = \widetilde{\beta}^{\text{Geo}}_{jk}$, this offers testable implications on sorting that we emphasize in the proposition below.

Proposition 1 (Testable Implications of Positive Sorting).

1. ***Cognitive tasks:*** $\widetilde{\beta}^{BGT}_{1,cog} > \widetilde{\beta}^{BGT}_{2,cog} > \widetilde{\beta}^{BGT}_{3,cog} > \widetilde{\beta}^{BGT}_{4,cog}$.

2. ***Amenities:*** $\widetilde{\beta}^{Amen}_{1} > \widetilde{\beta}^{Amen}_{2} > \widetilde{\beta}^{Amen}_{3} > \widetilde{\beta}^{Amen}_{4}$.

3. ***Distance:*** $\widetilde{\beta}^{Geo}_{1,dis} > \widetilde{\beta}^{Geo}_{2,dis} > \widetilde{\beta}^{Geo}_{3,dis} > \widetilde{\beta}^{Geo}_{4,dis}$.

4 The Estimation Results

We group universities into four tiers as in Section 2, covering 18 U.S. universities that are ranked among the top 20 globally or in the Ivy League (Top 20), 44 universities are ranked between 21 and 200 globally (Top 21-200), which are typically the best universities within each state, 115 U.S. universities are ranked between Top 201 and 1000 (Top 201-1000), and 87 universities from Top 1000 and 2000 (outside Top 1000).[21] Our estimation includes three dummy variables for Top 20, Top 21-200, and Top 201-1000, with universities outside Top 1000 considered as the benchmark group.

Our sample for estimation is formed by a combination of 264 universities, 25246 firms, and 22 occupations that have non-zero matches.

4.1 Task Complementarity

We begin by estimating Equation (16) as our baseline model. We first estimate an OLS regression for the interaction coefficients between our four BGT task variables (cognitive, social, routine, and manual) with rank dummies (Top 20, Top 21-200, and Top 201-1000). To ease interpretation, we further standardize each BGT task variable by its standard deviation.[22]

We control for the group fixed effects, λ_g, and the firm-occupation fixed effects, λ_{fo}. Since the inclusion of firm-occupation fixed effects absorbs the main effects of BGT task variables, only the interaction effects are included in the regression. Column (1) in Table 3 reports the coefficients from the OLS regression of $\ln \Pi_{fo}^{g}$ on the interaction coefficients. We find positive and precisely estimated interaction coefficients for cognitive tasks, suggesting positive sorting of university ranking into cognitive tasks. Specifically, the cognitive-Top 20 coefficient equals 0.200 (s.e. = 0.043). The cognitive-Top 21-200 coefficient remains precisely estimated but decreases to 0.074 (s.e. = 0.020), and further decreases to 0.030 (s.e. = 0.015) for cognitive-Top 1000. The monotonic decline in the coefficients provides evidence of positive sorting into cognitive tasks.

The estimates indicate that, holding other factors constant, a standard deviation increase in cognitive task requirements raises the likelihood of working as a computer scientist at Google rather than at Sanmina Corporation (an American electronics manufacturing services provider), graduates from Top 20 universities are 0.200 log points or about 20% more likely to match to the cognitive-intensive job (25.8% of a standard deviation of $\ln \Pi_{fo}^{g}$), relative to the benchmark group (outside Top 1000).[23] Relative to the benchmark group, we also find a 0.074 log points higher or 7.4% more likely for

[21]Top 20 universities include Harvard, MIT, Stanford, Columbia, Princeton, UC Berkeley, Penn, Chicago, Yale, Cornell, Northwestern, UCLA, Michigan, Johns Hopkins, UIUC, Duke, Dartmouth, and Brown.

[22]The variables in Table 2 are expressed as percentiles, ranging from 0 to 1. In this section, the BGT task variable is further standardized to have a standard deviation of one.

[23]The outcome variable, $\ln \Pi_{fo}^{g}$, has a standard deviation of 0.775. The 25.8% is computed as the ratio between 0.200 and 0.775.

graduates from Top 21-200, and a 0.030 log points higher or 3% more likely for Top 1000 to match the cognitive-intensive job.

Additionally, we observe small, negative, but mostly imprecisely estimated coefficients for interactions with routine and manual tasks, indicating little systematic sorting into these tasks across university rankings among college graduates. The interaction terms between social tasks and Top 20 is 0.057, although statistically insignificant.

4.2 Local Amenities

We augment our model to estimate the effects of the city's amenities in determining job matching. To ease interpretation, we further standardize the amenity to have a unit standard deviation. The amenity index is city-specific and does not vary across firms within the same city. We include interaction terms between the amenity and rank dummies. Again, the main effect of amenities is absorbed by the firm–occupation fixed effect.

Column (2) in Table 3 reports a sizable interaction coefficient equal to 0.092 (s.e. = 0.036) for amenity-Top 20. For example, consider Detroit, MI, which sits at the 25th percentile in terms of amenities among all U.S. cities, and Seattle, which ranks at the 99th percentile in terms of amenities. Holding job characteristics constant, graduates from Top 20 universities are 6.8% more likely to choose a job in Seattle than in Detroit, relative to graduates from universities outside Top 1000.[24] These results also indicate positive sorting on amenities among graduates from Top 20 universities. Relative to the benchmark group, we find no pattern in differential sorting on amenity for universities in Top 21-200, and a small negative coefficient for universities in Top 201-1000.

Our estimated results for positive sorting into cognitive tasks remain robust with the inclusion of these amenity interactions, and the estimates change slightly.

4.3 Geographic Factors

Next, we estimate the extent to which geographic proximity determines job matching and sorting. We augment our model by incorporating the interaction of ranking dummies with three geographic variables: the logarithm of the geographic distance between a university and the firm location, a dummy variable CZ_{fg} that equals one if university g and firm f are in the same city; a dummy $State_{fg}$ that equals one if university g and firm f are in the same state.

Column (3) of Table 3 reports the coefficient estimates of OLS regression with ge-

[24]We estimate the value as $0.092 \times (0.99 - 0.25) = 6.8\%$, using our preferred estimates in Column (2).

ography variables included. The coefficients for log distance and its interaction terms are identified from out-of-state movers. We see a negative and statistically significant coefficient for -0.077 (s.e. = 0.011) for log distance, indicating that a log point increase in geographic distance reduces job matching probability by a 0.077 log point for those graduating from outside the Top 1000 universities. The interaction terms between log distance and tier dummies pick up any differential effects of how geographic distance affects job matching relative to the baseline group. Importantly, we find a sizable and positive coefficient for the interaction of log distance and Top 20, equal to 0.071 (s.e. = 0.024). The interaction coefficients with Top 21-200 and Top 1000 are small and statistically insignificant.

These interaction coefficients show that, among movers, distance has no effect on job matching for Top 20 graduates but has notable effects for graduates from all other tiers. To interpret this using a concrete example, it means that graduates from Harvard and Stanford (both Top 20) are equally likely to secure a job in New York City, despite Boston's geographic proximity to New York. In contrast, distance matters significantly for graduates outside the Top 20 universities. For example, those who studied in Boston (e.g., Boston College) are 20.9% more likely to obtain a job in New York City than those who studied in the Bay Area (e.g., UC Davis).[25]

Unsurprisingly, we find positive, sizable, and statistically significant coefficients for CZ_{fg} and $State_{fg}$: these coefficients pick up the fact that college graduates are disproportionately likely to retain in university city or state.[26] The estimates indicate that for universities outside top 1000, their graduates are 0.44 log points or about 44.0% more likely to find a job in the same city as their alma mater, relative to other cities, holding other factors constant; and 0.073 log points or about 7.3% more likely to find a job within the state relative to other states.

Column (3) also reports positive and precisely estimated interaction effects for CZ $\times$ Top 21–200, CZ $\times$ Top 201-1000, State $\times$ Top 20, and State $\times$ Top 21-200. These positive interaction coefficients could capture two potential forces. The first is related to unobserved economic factors: how graduates from each tier of universities capitalize on job opportunities within the CZ or state of their university's location, relative to the benchmark group. The second is preferences, which capture the sorting of graduates across universities by unobserved local amenities.

[25] We estimate the value as $-0.077 \times (\ln(2557) - \ln(171)) = -0.209$, where the distances between the two city pairs are 2,557 and 171 miles.

[26] Note that although only 47.8% of college graduates stay in their university city after graduation, the share is much higher than the city's employment share in the national total.

4.4 Robustness

In-State Enrollment Shares. Among the 264 universities in our sample, 176 are public, and 88 are private, with 81% of individuals graduating from public institutions. On average, in-state enrollment is substantially lower at private universities (32.5%) than at public universities (76.1%). The high in-state shares at public universities likely reflect two factors. First, state legislatures aim to provide higher education opportunities for in-state residents, which creates incentives for public universities to admit more local students, as funding is often closely tied to in-state enrollment (Bound et al., 2020). Second, in-state students typically benefit from substantially lower tuition, making public universities more affordable for local residents.

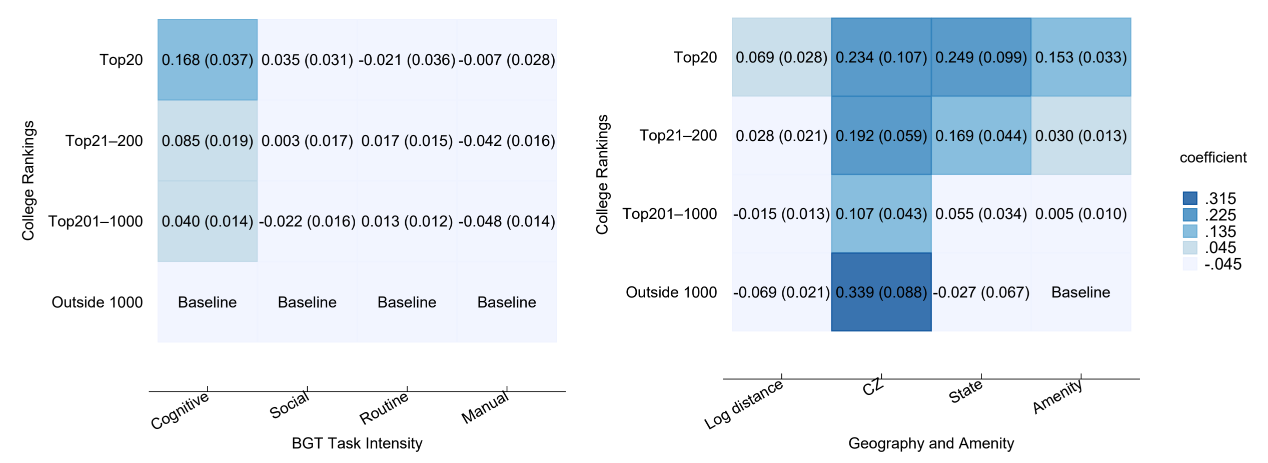

Figure 2: Summary for All Estimated Interaction Coefficients of Equation (16)

Notes: The coefficients are based on column (4) of Table 3. Darker blue colors indicate positive and statistically significant estimates; small or statistically insignificant estimates are plotted in light blue. Reported values are coefficients with standard errors in parentheses.

Because students are often willing to move long distances to attend elite private colleges, the geographic composition of student bodies differs even more markedly between elite public and private institutions. Consistent with this, IPEDS data show that in-state enrollment shares are low at elite private institutions such as Columbia, MIT, and Yale, all of which have in-state shares below 10%. In contrast, among public universities with similarly high rankings, the majority of students are in-state: for example, in-state shares are 71.5% at UC Berkeley, 62.3% at the University of Michigan, 70% at UCLA, and 77.3% at UIUC.

Given the importance of home bias in location choice emphasized in the migration literature, we conduct two exercises to check whether our results are driven by variation in the share of in-state students in the student body.

The first exercise examines the extent to which our estimates change when excluding private or elite institutions that tend to admit students from across the world. Specifically, Column (1) of Table 4 reports estimates of Equation (16) using a subsample that excludes groups g corresponding to private universities ranked in the global top 20 or belonging to the Ivy League. This restriction removes elite private institutions whose students are particularly willing to travel long distances for high-quality education and may therefore be less locally attached than those at other institutions.

Column (2) of Table 4 reports estimates using a subsample that includes only public universities. As discussed above, public universities enroll a majority of in-state students, including even highly ranked public institutions. Focusing on this sample, therefore, mitigates concerns that differences in students' states of origin drive mobility patterns. Across both samples, we find that the estimates and qualitative conclusions are similar to the baseline results.

In the second exercise, we estimate the baseline regression using the full sample, while augmenting the model with three additional interaction terms between three geographic variables (log distance, CZ dummy, and state dummy) and the share of in-state student enrollment. Column (4) of Table 3 reports the estimates. The three newly-added interaction terms appear to be imprecisely estimated. Conditional on these interactions, we find modest changes for other coefficients.

To summarize the coefficient estimates with the full set of controls, Figure 2 displays the university ranking-related interaction coefficients based on the preferred specification reported in Column (4). Darker blue colors indicate positive and statistically significant estimates; small or statistically insignificant estimates are plotted in light blue. Reported values are coefficients with standard errors in parentheses.

The previously highlighted interaction coefficients change only modestly but remain statistically significant. For example, we estimate a cognitive–Top 20 coefficient of 0.168 (s.e. = 0.037), an amenity–Top 20 coefficient of 0.153 (s.e. = 0.033), a distance–Top 20 coefficient of 0.069 (s.e. = 0.028), and a distance coefficient of –0.069 (s.e. = 0.021).

Family Background. Students enrolled in public and private universities differ systematically in family income background (Chetty et al., 2020), and such differences may shape labor market outcomes. To account for this possibility, we augment the model with four interaction terms between the BGT task measures (cognitive, social, routine, and manual) and a public-university dummy. Because our regression already controls for university fixed effects and ranking-related interactions, these public–task interaction terms capture differential sorting between public and private universities within the same ranking tier—likely reflecting differences in family background. Column (5) shows that none of these additional controls matter. Importantly, the coefficient estimates remain nearly identical with or without these controls (columns 4 vs. 5). In a recent study, Chetty et al. (2023) find that the factors giving children from high-income families an admissions advantage are uncorrelated with post-college outcomes. Our findings are in line with theirs.

4.5 Positive Sorting into Firms or Occupations?

Unlike the DOT or O*NET data, the BGT task requirements vary across both firms and occupations, thus the positive sorting in cognitive tasks that we estimate may reflect two distinct channels: (1) firm sorting—graduates from higher-ranked universities tend to work in firms with greater cognitive task requirements; and (2) occupational sorting—graduates from higher-ranked universities tend to enter more cognitively intensive occupations. This section examines the relative importance of these two channels in driving the observed positive sorting.

Previous literature has analyzed sorting into occupations using O*NET data (Deming, 2017, Burstein et al., 2019) and sorting into firms using the AKM (1999) approach and employer–employee matched data (Card et al., 2013, Lopes de Melo, 2018, Song et al., 2019, Bonhomme et al., 2023). Our approach estimates sorting into both firms and occupations within a unified framework, allowing us to shed light on the relative strength of firm versus occupation dimensions.

We rewrite Equation (9) in an estimable form as

$$\begin{aligned}\ln \Pi^{g}_{f|o} = & \sum_{j=1}^{J}\sum_{k=1}^{K}\widetilde{\beta}^{\text{BGT}}_{jk}X^{g}_{j}Y^{\text{BGT}}_{fok} + \sum_{j=1}^{J}\widetilde{\beta}^{\text{Amen}}_{j}X^{g}_{j}Y^{\text{Amen}}_{c} + \sum_{j=1}^{J}\sum_{k=1}^{K}\widetilde{\beta}^{\text{Geo}}_{jk}X^{g}_{j}Y^{\text{Geo}}_{fgk} \\ & + \lambda_{go} + \lambda_{fo} + \theta \ln \tau^{g,\text{UBV}}_{fo}.\end{aligned} \tag{17}$$

The outcome variable is the fraction of g-graduates who choose f conditional on occupation o. The estimation requires controlling for university-occupation and firm-occupation fixed effects. Since the university-occupation fixed effects absorb the aggregate occupational employment opportunities specific to a university, the coefficients for the rank-task interaction terms now characterize the strength of sorting across firms, holding the same occupation.

For comparison, Figure 3 displays the university ranking–related interaction coefficients. The complete sets of coefficients are reported in Table F.1 with different specifications. Because university-occupation pairs that are only observed in one firm would not contribute to identification, the sample size differs slightly from the baseline. The estimates are notably smaller. As cognitive task requirements increase by one standard deviation, graduates from Top 20 universities have a 0.087 log points higher or 8.7% more likely to work in a cognitive-intensive firm relative to the benchmark group. Compared to Figure 2, the estimates suggest that sorting into firms accounts for about $0.087/0.168 = 51\%$ of the overall sorting into jobs (firms and occupations combined). The interaction coefficients for Cognitive $\times$ Top 21–200 and Cognitive $\times$ Top 201–1000 also decline, falling to roughly half of the corresponding estimates in Figure 2. Positive sorting into both firms and occupations is quantitatively important in shaping the overall positive sorting on cognitive tasks.[27]

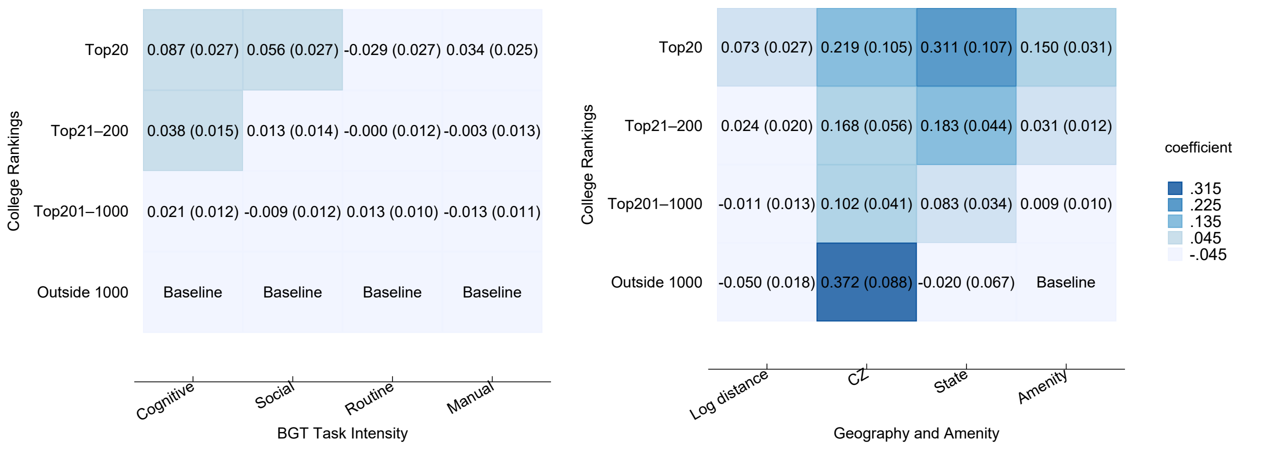

Figure 3: Summary for All Estimated interaction Coefficients of Equation (17)

Notes: The coefficients are based on column (4) of Table F.1. Darker blue colors indicate positive and statistically significant estimates; small or statistically insignificant estimates are plotted in light blue. Reported values are coefficients with standard errors in parentheses.

5 Sorting on Other Mechanisms

This section tests several potential confounding mechanisms that might influence our findings.

[27] In addition, the interaction coefficients related to geography and amenities are similar between the estimates of the two figures. The results are expected, since spatial variation is captured in our definition of a firm rather than in occupation.

5.1 Admission Selectivity

The sorting we have found based on university ranking can be a byproduct of two primary factors: the effect of attending a specific university and the selection—students admitted to better universities may, on average, have higher pre-college ability and therefore better labor-market outcomes regardless of the university they attend. This section examines the relative importance of the two factors.

To this end, we use Barron's Profiles of American universities (2017 edition), which classifies universities into six tiers based on SAT/ACT scores and other admission requirements such as high school transcripts and class rank (Barron's Educational Series, 2017) (see Table F.4). Unlike the WUR or U.S. News rankings, these six categories assess the selectivity of college admissions and serve as proxies for students' ability prior to entering college, rather than providing ratings based on academic standards, educational quality, or institutional reputation.

Importantly, the selectivity criterion does not align with the WUR rankings. For example, UC Berkeley (ranked 8th in WUR), Tulane University (407th in WUR), and Davidson College (1866th in WUR) are all classified among the most selective universities. To increase precision in the estimation, we aggregate the admission criteria into four groups: most selective, highly and very selective, selective, and less and non-selective.

Column (2) of Table 5 augments the baseline model with interaction between the selectivity dummies and Y_{fok}, where Y_{fok} includes the three sets of variables (BGT tasks, amenities, and geographical variables). In comparison, Column (1) reports the baseline estimates. As shown in Column (2), there is no evidence of labor-market sorting by college selectivity once university-ranking interactions are included. Importantly, the ranking-related interaction coefficients remain significant and change modestly relative to the baseline.

We find little evidence of positive sorting into cognitive jobs due to differential selection in college admissions. Rather, the evidence suggests that systematic differences associated with universities' training, reputation, and educational quality—captured by the WUR—are crucial in shaping graduates' career outcomes.

5.2 STEM vs. Non-STEM Majors

Graduates who majored in fields such as STEM and quantitative disciplines are known to be more likely to obtain high-paid jobs (Deming and Noray, 2020). To examine whether our results are driven by variation in major composition across institutions, we augment the model with interaction terms between the share of graduates in STEM majors and Y_{fok} (BGT tasks, amenities, and geographical variables). The share of STEM majors for each university is calculated from our LinkedIn data, where STEM majors include one of the following broad fields: Science, Technology, Engineering, or Mathematics.

In Column (3) of Table 5, we find a sizable and positive coefficient for the cognitive–STEM interaction, equal to 0.223 (s.e. = 0.068). These results suggest that universities with a stronger STEM orientation tend to place their graduates in more cognitively intensive but less socially intensive jobs. Interestingly, we also find positive and precisely estimated coefficients for the distance-STEM and amenity-STEM interactions, suggesting that STEM graduates are more geographically mobile than non-STEM graduates and are more likely to relocate to cities with better amenities.

Importantly, while the estimates underscore the role of major composition in shaping job outcomes, the positive sorting on cognitive-ranking related interaction coefficients are similar (some fall modestly) and remain statistically significant when major composition is considered.

5.3 University-to-Firm Network

The literature has documented that migration networks not only improve the likelihood of job matching but also enhance job outcomes through referrals (Munshi, 2003). Our dataset allows us to measure the university-to-firm network. This section provides evidence, from two perspectives, that our estimated results hold even when university-to-firm network effects are taken into account.

A Theoretical Re-Interpretation. First, we show that by incorporating network effects into the model under a certain structure, one can derive an estimating equation isomorphic to our baseline Equation (16). Specifically, suppose that unobserved job search costs are related to the university-to-job (firm and occupation) matching pattern in the previous period. In particular, we assume that $\ln \Pi^{g}_{fot}$ follows an $AR(1)$ process driven by the unobserved search cost component, that is,

$$\ln \tau^{g,\mathrm{UBV}}_{fo,t} = \rho \ln \Pi^{g}_{fo,t-1} + \varepsilon_{fo,t}, \quad \rho \in (0,1), \tag{18}$$

where $\Pi^{g}_{fo,t-1}$ is the share of g-graduates who work in firm f and occupation o in the past or that of earlier graduates (alumni). The parameter ρ captures the extent to which the job matching of alumni influences current job search costs. The term $\varepsilon_{fo,t}$ represents white noise—factors unrelated to $\ln \Pi^g_{fo,t-1}$ but affecting job search costs $\ln \tau^{g,\text{UBV}}_{fo,t}$.

In the data, there is a strong persistence pattern of university-to-firm job matches (see Appendix Figure E.2). The pattern is similar to the well-documented literature on the immigrant enclave, where studies have documented that for U.S. immigrants from a particular origin country, their location settlement (Card, 2001) and occupational choices (Hanson and Liu, 2016) tend to persist over time.

Assuming T^g_{fo}, a^g_c, and $\ln \tau^g_c$ are all time invariant, we can substitute Equation (18) into (16), recursively, to obtain

$$\ln \Pi^g_{fo,t} = \lambda_g + \lambda_{fo} + \widetilde{\theta} \ln T^g_{fo} + \widetilde{\theta} \ln a^g_c + \widetilde{\theta} \ln \tau^{g,\text{Geo}}_c + \theta \sum_{m=0}^{t} \rho^m \varepsilon_{fo,t-m}, \tag{19}$$

where $\widetilde{\theta} = \frac{\theta}{1-\rho}$ is a function of the job-matching elasticity to productivity, θ, and the path dependence of university-specific job matching, $1-\rho$. Equation (19) is therefore isomorphic to our baseline equation, differing only in the interpretation of the reduced-form parameter and the structural residuals. Put simply, as long as idiosyncratic shocks to job matching at any time t, $\varepsilon_{fo,t-m}$, are independent of the observed controls, the OLS estimate of the reduced-form parameter is unlikely to suffer from the omitted variable bias. The main difference lies in the interpretation of the coefficients.

Empirical Results. The second piece of evidence we provide is empirical. Using the LinkedIn dataset, we can control for firm exposure to alumni from the university g as

$$\text{Network}^g_f = \frac{L^{g,\text{alumni}}_f}{L_f}, \tag{20}$$

capturing the share of firm f's current employees that graduated from g before 2014 (we refer to as alumni), $L^{g,\text{alumni}}_f$, relative to L_f, the current employment size of firm f. Both L_f and $L^{g,\text{alumni}}_f$ are estimated from the LinkedIn data.

Since unobserved factors that lead to a job match of alumni might also promote current graduates, we do not claim any causal estimates for the impact of alumni networks on job matching. The empirical exercise below aims to shed light on which of the sorting channels we have estimated so far—task, amenity, or geography—the network primarily operates through.

Column (4) of Table 5 augments the model with the network measure and shows that the university-to-firm network is a strong predictor of employment outcomes, with a coefficient of 1.669 (s.e. = 0.048), indicating that a one–standard deviation increase in network exposure is associated with a 1.669 log points higher of graduates working at that firm.

Despite the sizable network coefficient, the cognitive- and amenity-related interaction coefficients remain statistically significant, though somewhat reduced in magnitude. For example, the coefficient on Cognitive × Top 20 declines from 0.148 to 0.12, while the coefficient on Cognitive × Top 21–200 decreases from 0.072 to 0.048.

With this network control, the largest changes occur in the interaction terms involving geographic variables. This pattern suggests that the university-to-firm network operates primarily through geographic channels.

5.4 Other Robustness Checks

Using Disaggregated Groupings. Our baseline estimation divides the universities into four groups. We show that the results remain similar when using more disaggregated groupings. Specifically, we split the Top 21–200 category into two subgroups: universities ranked 21–100 (labeled Top 21–100) and those ranked 101–200 (labeled Top 101–200). Likewise, we split the Top 201–1000 category into two subgroups: universities ranked 201–500 (labeled Top 201–500) and those ranked 501–1000 (labeled Top 501–1000). Table F.2 reports the estimated coefficients, showing that most interaction terms are similar across the subgroups within each broader category. These findings suggest that using broader groupings, as in the baseline estimation, does not obscure substantial heterogeneity in the sorting pattern.

U.S. News Rankings. While our baseline estimates rely on the WUR rankings, we find similar results when using the U.S. News rankings. Unlike WUR, which covers the Top 2000 universities globally, U.S. News provides rankings only for the Top 1000. For U.S. universities that appear in both sources, the correlation between the two ranking measures is 0.862. In the estimation, we classify U.S. universities into three groups: Top 20, Top 21–200, and Top 201–1000 (the benchmark group). As reported in Table F.3, estimates based on U.S. News rankings are similar in magnitude and lead to the same conclusions.

6 Labor Market Outcomes

Our estimation has recovered the assignment of talent to jobs as a function of university ranking, task contents, amenities, and geography. In this section, we use the estimated model to (1) compare the relative importance of these factors in shaping the labor market outcomes of recent graduates from each college; and (2) examine whether these factors widen or narrow disparities in outcomes across colleges. We focus on two key outcomes. Section 6.2 examines the probability that graduates match to jobs in the top 5% of the wage distribution, highlighting access to jobs in the upper tail of the wage distribution. Section 6.3 examines average earnings among college graduates.

6.1 Counterfactual Job Matching Probabilities

We first use the estimated model to compute the benchmark probability as

$$\Pi_{fo}^{g,\text{Benchmark}} = \frac{\exp\left(\lambda_{fo}\right)\left(T_{fo}^{g} a_{c}^{g} \tau_{c}^{g,\text{Geo}} \tau_{fo}^{g,\text{UBV}}\right)^{\theta}}{\sum_{f'o'} \exp\left(\lambda_{f'o'}\right)\left(T_{f'o'}^{g} a_{c'}^{g} \tau_{c'}^{g,\text{Geo}} \tau_{f'o'}^{g,\text{UBV}}\right)^{\theta}}, \tag{21}$$

where each component is estimated based on our preferred specification, reported in Column (4) of Table 3. Specifically, λ_{fo} is the estimated firm-occupation fixed effects, and $\exp\left(\lambda_{fo}\right)$ measures $(\omega_{fo}/p_c)^{\theta}$. T_{fo}^{g}, a_c^g, and $\tau_c^{g,\text{Geo}}$ are estimated as

$$\begin{aligned} T_{fo}^{g} &= \exp\left(\sum_{j=1}^{J}\sum_{k=1}^{K} \hat{\beta}_{jk}^{\text{BGT}} X_j^g Y_{fok}^{\text{BGT}}\right), \quad a_c^g = \exp\left(\sum_{j=1}^{J} \hat{\beta}_{j}^{\text{Amen}} X_j^g Y_{c}^{\text{Amen}}\right), \\ \tau_c^{g,\text{Geo}} &= \exp\left(\sum_{j=1}^{J}\sum_{k=1}^{K} \hat{\beta}_{jk}^{\text{Geo}} X_j^g Y_{fk}^{\text{Geo}} + \sum_{j=1}^{J}\sum_{k=1}^{K} \hat{\gamma}_{jk}^{\text{Geo}} \text{In-state}_j^g Y_{fk}^{\text{Geo}}\right). \end{aligned} \tag{22}$$

$\hat{\beta}_{jk}^{\text{BGT}}$, $\hat{\beta}_{j}^{\text{Amen}}$, $\hat{\beta}_{jk}^{\text{Geo}}$, and $\hat{\gamma}_{jk}^{\text{Geo}}$ denote the estimated coefficients. In-state$_j^g$ is the fraction of in-state student enrollment.

In computing T_{fo}^{g}, a_c^g, and $\tau_c^{g,\text{Geo}}$, we rely on coefficients that are statistically significant at the 95% confidence level, setting statistically insignificant parameters to zero. $\tau_{fo}^{g,\text{UBV}}$ is obtained as the exponent of the regression residuals. Since the estimated model has an R^2 of 0.75—implying that unobserved components account for 25% of the variation in outcomes—we thus decide to incorporate the regression residual (through $\tau_{fo}^{g,\text{UBV}}$) when computing the benchmark probability.

Note that, because we compute $\Pi_{fo}^{g,\text{Benchmark}}$ using statistically significant coefficients, the predicted probabilities do not match exactly the observed allocations but almost perfectly. Appendix Figure E.3 shows that the predicted allocations fit the observed ones very well: a simple OLS regression of observed shares on predicted shares yields a coefficient of 0.95 and an R^2 of 0.99.

We then compute three counterfactual probabilities: when each force (geography, task complementarity, or amenity) is absent. Specifically, when turning off geographic factors, we set $\hat{\beta}_j^{\text{Geo}} = \hat{\gamma}_j^{\text{Geo}} = 0$ (which implies $\tau_c^{g,\text{Geo}} = 1$) and estimate the counterfactual probability as

$$\Pi_{fo}^{g,\text{Geo}} = \frac{\exp\left(\lambda_{fo}\right)\left(T_{fo}^{g} a_c^g \tau_{fo}^{g,\text{UBV}}\right)^{\theta}}{\sum_{f'o'} \exp\left(\lambda_{f'o'}\right)\left(T_{f'o'}^{g} a_{c'}^g \tau_{f'o'}^{g,\text{UBV}}\right)^{\theta}}. \tag{23}$$

$\Pi_{fo}^{g,\text{Geo}}$ is determined by other factors such as task complementarity (T_{fo}^g), sorting in amenity a_c^g, the unobserved factor $\tau_{fo}^{g,\text{UBV}}$, and firm-occupation-specific factor λ_{fo}. The counterfactual probabilities $\Pi_{fo}^{g,\text{BGT}}$ and $\Pi_{fo}^{g,\text{Amen}}$ are estimated similarly, while setting $\hat{\beta}_j^{\text{BGT}} = 0$ or $\hat{\beta}_j^{\text{Amen}} = 0$ (which implies $T_{fo}^g = 1$ or $a_c^g = 1$), respectively.

6.2 On the Probability of Matching to High-Paid Jobs

We first quantify the extent to which geography, task complementarity, and amenity determine the likelihood of matching to high-paid jobs. We define high-paid jobs as those in the top 5% of the Glassdoor entry-level wage distribution. For university g, we first compute

$$\text{ProbTop}_g^{\text{Geo}} = \sum_{fo}\left(\Pi_{fo}^{g,\text{Benchmark}} - \Pi_{fo}^{g,\text{Geo}}\right) \times \mathbb{1}_{fo}^{\text{Top5\%}}. \tag{24}$$

Here, $\mathbb{1}_{fo}^{\text{Top5\%}}$ is an indicator that equals one if the salary of a job (firm and occupation) is in the top 5%. $\Pi_{fo}^{g,\text{Benchmark}}$ is defined in Equation (21) and $\Pi_{fo}^{g,\text{Geo}}$ is defined in Equation (23). $\text{ProbTop}_g^{\text{Geo}}$ measures the difference in the probability of matching to top 5% paid job between the observed and the counterfactual equilibrium (when geographic factors are absent). We refer to $\text{ProbTop}_g^{\text{Geo}}$ as the effect of geography on matching to top 5% paid jobs.

Similarly, we compute $\text{ProbTop}_g^{\text{BGT}}$, the effect of task complementarity on the probability of matching to top 5% paid jobs, by replacing $\Pi_{fo}^{g,\text{Geo}}$ with $\Pi_{fo}^{g,\text{BGT}}$ in Equation (24); and compute the effect of amenity, $\text{ProbTop}_g^{\text{Amen}}$, by using $\Pi_{fo}^{g,\text{Amen}}$.

We estimate $\text{ProbTop}_g^{\text{Geo}}$, $\text{ProbTop}_g^{\text{BGT}}$, and $\text{ProbTop}_g^{\text{Amen}}$ for each g. Two points are worth emphasizing. First, we find that geography has a larger impact on determining the matching probability than the other two factors. Geographic factors tend to reduce the likelihood of matching to high-paid jobs, whereas task and amenity increase this probability. Averaging across colleges, geographic factors reduce the probability by 0.62 ppt, whereas the other two increase the likelihood modestly by 0.10 and 0.11 ppt, respectively. Moreover, the effect of geographic forces also varies more systematically across colleges, ranging from -6.2 to 1.9 ppt, with a standard deviation of 1.01 ppt. The effect of task complementarity and amenities varies modestly: the former ranges from -0.25 to 1.43 ppt with a standard deviation of 0.198 ppt, and the latter ranges from

-0.04 to 2.80 ppt, with a standard deviation of 0.354 ppt.

Second, we assess whether each factor narrows or widens the disparities in outcomes across colleges. Figure 4 plots, for each college, the benchmark probability of matching to top 5%-paid jobs (x-axis) against $\text{ProbTop}_g^{\text{Geo}}$ (left panel), $\text{ProbTop}_g^{\text{BGT}}$ (middle panel), and $\text{ProbTop}_g^{\text{Amen}}$ (right panel) on the y-axis. The left panel shows a negative relationship: for colleges, whose graduates are 10 ppts more likely to match to Top 5% high-paid jobs, geographic factors reduce their likelihood of matching to these high-paid jobs by 0.633 ppts.[28] The negative relationship reflects the fact that many highly ranked universities are located in relatively low-wage cities, suggesting that the spatial divide between top-ranked colleges and high-paying jobs, on average, *narrows* the probability of matching to high-paid jobs across colleges.

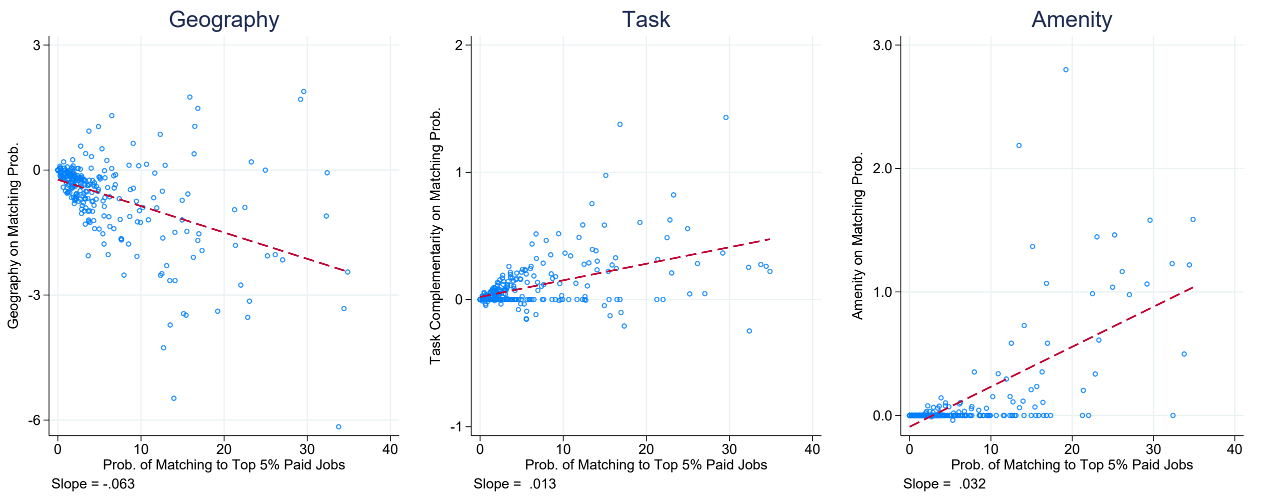

Figure 4: The Probability of Matching to Top 5% of Paid Jobs across Colleges

Notes: The x-axis represents the probability of matching to top 5% paid jobs ($\Pi_{fo}^{g,\text{Benchmark}} \times \mathbb{1}_{fo}^{\text{Top5\%}} \times 100$). The y-axis is $\text{ProbTop}_g^{\text{Geo}}$, $\text{ProbTop}_g^{\text{BGT}}$, and $\text{ProbTop}_g^{\text{Amen}}$, respectively (all $\times$ 100). Each blue circle denotes a specific college. A linear regression estimates a coefficient of -0.0633 (s.e.=0.0077) on the left, 0.013 (s.e.=0.0015) on the middle, and 0.0324 (s.e.=0.0023) on the right.

In contrast, because high-paid jobs are often those that have high cognitive task requirements or are located in high-amenity cities, we find positive associations in the middle and right panels. Specifically, for colleges whose graduates are 10 ppts more likely to match to Top 5% high-paid jobs, task complementarity increases the probability of matching to high-paid jobs by $0.013 \times 10 = 0.13$ ppts, and amenity increases the probability by $0.0324 \times 10 = 0.324$ ppts.[29]

[28] A linear regression estimates a coefficient of -0.0633 (s.e.=0.0077). The value 0.633 ppt is obtained as 0.0633×10.

[29] The estimates are 0.013 (s.e.=0.0015) on the middle, and 0.0324 (s.e.=0.0023) on the right.

In summary, we find that geographic frictions narrow the probability of accessing paid jobs by 6.3% across colleges, whereas task complementarity and amenity widen the probability by 1.3% and 3.2%, respectively.

6.3 On the Annual Earnings

Next, we quantify how each factor shapes the average earnings of fresh graduates. To this end, we first compute

$$\text{Wage}_g^{\text{Geo}} = \sum_{fo} (\Pi_{fo}^{g,\text{Benchmark}} - \Pi_{fo}^{g,\text{Geo}}) \times \text{wage}_{fo}, \tag{25}$$

wage_{fo} denotes the Glassdoor wage posted for firm f and occupation o.[30]

$\Pi_{fo}^{g,\text{Benchmark}}$ and $\Pi_{fo}^{g,\text{Geo}}$ are the matching probabilities given in Equations (21) and (23), respectively. $\text{Wage}_g^{\text{Geo}}$ measures the differences in average earnings of college g graduates between the observed and the counterfactual equilibria (when geographic factors are absent). We refer to $\text{Wage}_g^{\text{Geo}}$ as the effects of geography on earnings. Note that, by taking wage_{fo} unchanged, our approach is partial equilibrium, and the estimated wage effects from geography are driven entirely by changes in the matching probability.

Similarly, we compute $\text{Wage}_g^{\text{BGT}}$ and $\text{Wage}_g^{\text{Amen}}$, the effect of task complementarity or amenity on earnings, respectively. In line with the probability measure, we find geography has a larger negative impact on determining average earnings, whereas the other factors have small positive impacts: averaging across colleges, the former reduces average earnings by \$1,263; in contrast, task complementarity increases earnings by \$344, and amenity increases earnings by \$142.

Analogous to Figure 4, Figure 5 (left panel) plots $\text{Wage}_g^{\text{Geo}}$, on the y-axis against the average earnings of graduates by each college on the x-axis. Geographic frictions benefit lower-earning colleges more than higher-earning colleges, suggesting that they narrow earnings gaps across colleges. The fitted line has a slope of -0.022 with a standard error of 0.0068. Put simply, for a \$10,000 difference in starting salaries across colleges, geographic factors reduce the earnings gap by \$220.

The middle panel of Figure 5 replaces the y-axis with $\text{Wage}_g^{\text{BGT}}$, while the right panel uses $\text{Wage}_g^{\text{Amen}}$. We find positive slopes of 0.006 and 0.021, respectively. This implies that for every \$10,000 difference in starting salaries across colleges, task complemen-

[30]Here, Wage_{fo} corresponds to the firm–occupation-specific wage component, which does not vary across individuals. By using Wage_{fo}, our analysis abstracts from idiosyncratic wage variation within firm–occupation cells.

tarity widens the earnings gap by \$60, while amenities widen the gap by \$210.[31]

Putting these results together, we interpret geographic factors as narrowing earnings gaps across colleges by 2.2%. In contrast, sorting on tasks and amenities widens earnings gaps by 0.6% and 2.1%, respectively.

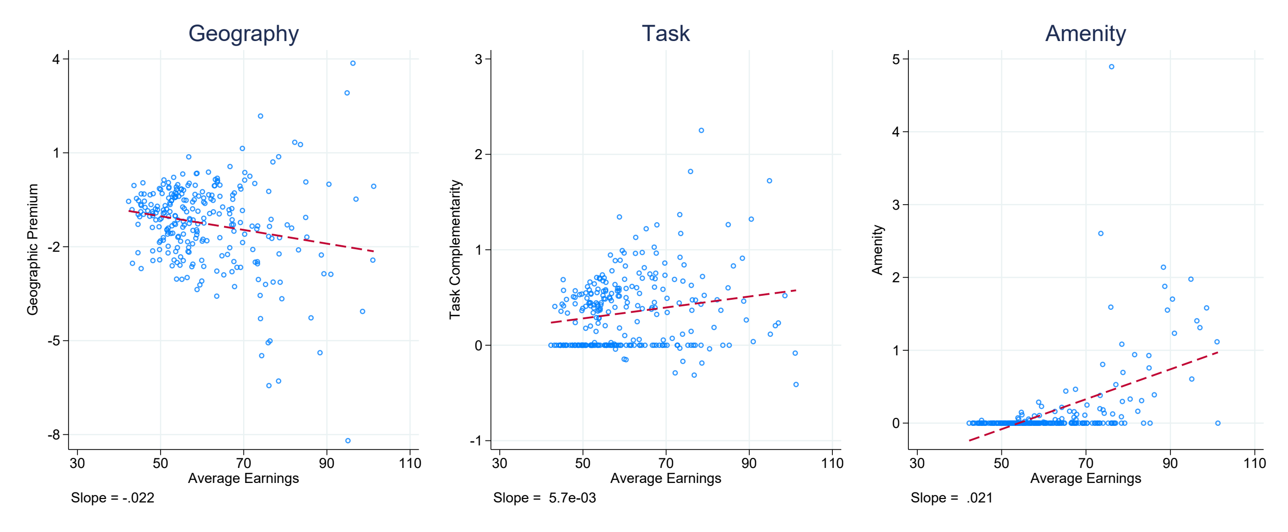

Figure 5: The Effects and the Average Earnings across Colleges

Notes: The x-axis represents average earnings of graduates from a college g. The y-axis is $\text{Wage}_g^{\text{Geo}}$, $\text{Wage}_g^{\text{BGT}}$, and $\text{Wage}_g^{\text{Amen}}$, respectively. All variables are measured in thousands. Each blue circle denotes a specific college. A linear regression estimates a coefficient of -0.022 (s.e.=0.007) on the left, 0.0057 (s.e.=0.002) on the middle, and 0.021 (s.e.=0.002) on the right.

6.4 The Geographic Premium

The Geographic Premium by Colleges. We next examine the effect of geography on annual earnings by university, $\text{Wage}_g^{\text{Geo}}$, estimated using Equation (25). The effects on matching probability to high-paid jobs are presented in Appendix D. Figure 6 (left panel) plots the average annual earnings (blue circles) and the wage premium associated with geography (red triangles), which we refer to as the geographic premium. We report results for 30 universities: the fifteen with the highest premiums and the fifteen with the lowest (most negative) premiums.

Universities differ systematically in the average entry salary earned by their fresh graduates. Importantly, there is systematic variation in geographic premium. The largest ones are observed for Columbia University and UC Berkeley, amounting to annual salary premiums of \$3.9K and \$2.9K, respectively. Other universities in the Bay Area or New York City, such as UC Santa Cruz (\$2.2K), Santa Clara (\$1.3K), and

[31] These values are computed as $0.006 \times 10 = 0.06$ thousand, and $0.022 \times 10 = 0.22$ thousand.

NYU ($1.3K), also rank among the highest.

By contrast, the largest geographic penalty is observed for Carnegie Mellon University (-$8.2K). This reflects the university's specialization in training STEM graduates while being located farther from major technology hubs such as San Jose and Seattle, where many large U.S. tech firms are concentrated and wages are among the highest. Despite this geographic disadvantage, Carnegie Mellon graduates earn an average annual salary of $96.3K—higher than most universities displayed. For similar reasons, we also find large location penalties for Brown (–$6.4K), Georgia Tech (–$6.3K), Emory University (–$5.5K), and the University of Pennsylvania (–$5.4K).

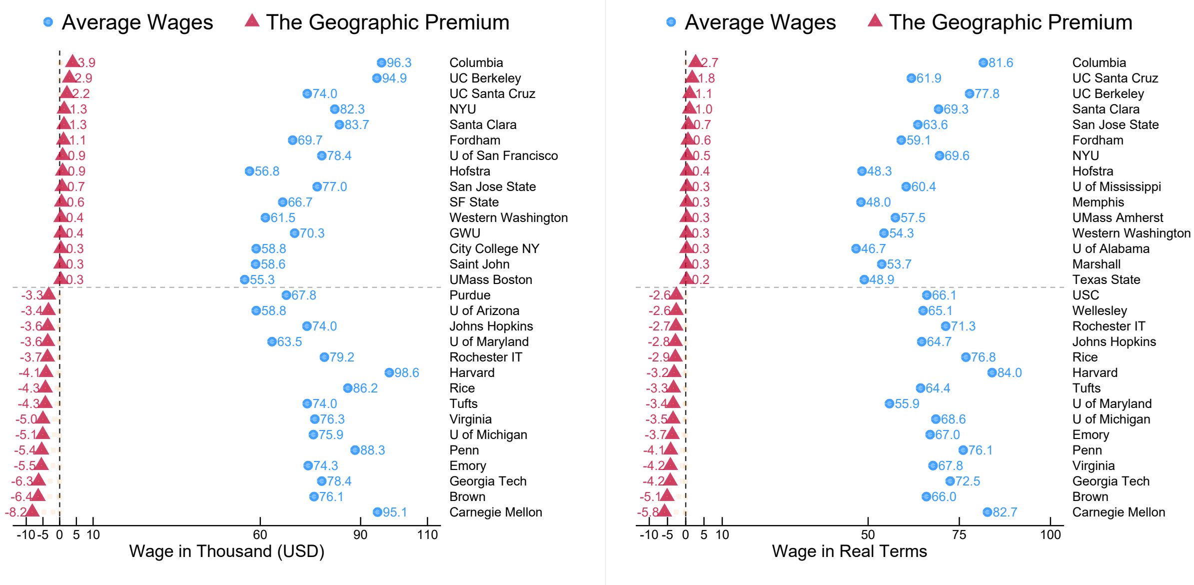

Figure 6: Geographic Premium by Colleges in Nominal (left) and in Real Terms (right)

Notes: The left panel shows the average nominal earnings in thousands. The blue circles plot the average wage earned by fresh graduates from each university. The red triangle plots the value of $\text{Wage}_g^{\text{Geo}}$, capturing the effect of geography on average earnings. The right panel shows the average earnings in thousands in real terms. We display 30 universities: the fifteen with the highest geographic premium and the fifteen with the lowest (most negative).

The extent to which geographic factors affect the earnings of college graduates depends on several elements: the geographic clustering of U.S. industries and firms, the skills imparted by universities that give graduates a comparative advantage in certain industries or jobs, and the geographic barriers that limit graduate mobility. To analyze mechanisms that drive variation, Appendix Figure E.4 plots the geographic premium (y-axis) against the mover premium (x-axis) across universities. Again, we measure the mover premium as the difference in average wages between movers and stayers among fresh graduates (2016–2018), which can be estimated directly from our sample. A negative mover premium implies that, on average, local stayers earn higher wages than those who migrate.

We find a strong negative correlation: a simple OLS regression yields a coefficient of -0.08 (s.e. = 0.01) and an R^2 of 0.52. It appears that a single variable—the mover premium—explains more than half of the variation in the geographic premium, underscoring the central role of university location and graduates' geographic mobility in shaping job outcomes.

The geographic premiums that we have discussed so far are expressed in nominal terms. However, cities differ substantially in their cost of living (Moretti, 2013, Albert and Monras, 2022), and part of the wages firms offer likely reflects local living costs, especially in expensive metropolitan areas. To account for this, we use the U.S. Bureau of Economic Analysis (BEA) Regional Price Parities (RPPs) for 2018. The RPPs provide relative price levels across MSAs, covering major expenditure categories such as housing, transportation, and food, and are constructed using data from the Bureau of Labor Statistics' Consumer Price Index, housing price data, and other regional sources. In a recent study, Diamond and Moretti (2024) demonstrate that the BEA index is strongly correlated with an alternative price index constructed using detailed consumption data.

We normalize the average value of the RPPs across all cities to one.[32] We deflate Glassdoor wages using each city's RPP and re-estimate Equation (25). Figure 6 (right panel) shows that adjusting for regional price variation reduces disparities in premiums across universities to some extent; however, sizable premiums persist for universities located in large metropolitan areas such as San Jose, San Francisco, and New York.

The Bay-Area Premium. In Figure 6, universities in the Bay Area, although spanning a wide range of rankings, occupy the top of the list.[33] Motivated by this pattern, we further estimate the wage premium associated with studying in the Bay Area, aggregating over Ω_{Bay}, which includes all universities in San Francisco and San Jose. Specifically, we estimate the geographic premium for the Bay Area as

$$\frac{1}{N_{\text{Bay}}} \sum_{g \in \Omega_{\text{Bay}}} \sum_{fo} (\Pi_{fo}^{g,\text{Benchmark}} - \Pi_{fo}^{g,\text{Geo}}) \times \text{wage}_{fo}, \tag{26}$$

[32] In 2018, the normalized RPP was as high as 1.26 in San Francisco and 1.20 in New York City, compared to 1.00 in Pittsburgh (PA), and as low as 0.88 in Jackson (TN).

[33] For example, UC Berkeley (Top 20), UC Santa Cruz (Top 201–1000), Santa Clara (outside the Top 1000), the University of San Francisco (outside the Top 1000), San Jose State (outside the Top 1000), and San Francisco State (outside the Top 1000) all appear among the top institutions in Figure D.1.

where N_{Bay} denotes the number of universities located in the Bay Area in our sample. The term measures the average geographic premium of attending a university in the Bay Area (averaged across all local universities). Similarly, we can use Equation (26) to estimate the geographic premium for any U.S. city c. Taking the difference allows us to measure how geography differentially affects the entry-level salaries of graduates from the Bay Area relative to those from city c (referred to as the Bay Area premium).

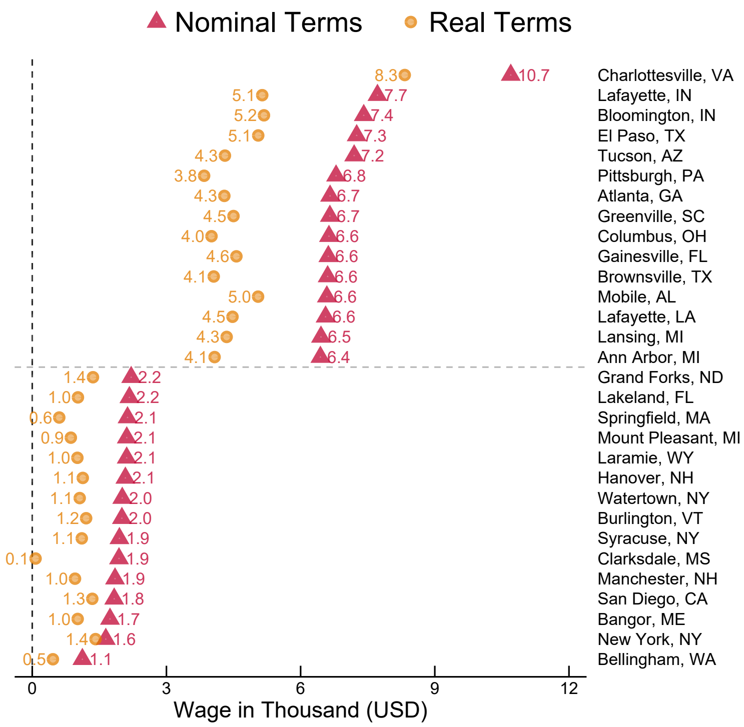

Figure 7: The Bay Area Premium Relative to a Given City in Nominal (red triangles) and Real Terms (orange circles)

Notes: The x-axis represents the average wage in thousands. The red triangle plots the value of the Bay Area premium in nominal terms. The orange circles plot the Bay Area Premium in real terms. We display 30 cities: the fifteen with the highest Bay Area Premium and the fifteen with the lowest.

Figure 7 displays the Bay Area premium relative to a given city (as labeled). Again, we report results for 30 cities: the fifteen with the highest premiums (above the dashed line) and the fifteen with the lowest (below the dashed line). The largest geographic disadvantage is observed in Charlottesville, VA, with an average nominal wage penalty of \$10.7K. Midwest or southwest cities that host major state universities—such as Lafayette (LA), Bloomington (IN), El Paso (TX), Tucson (AZ), Pittsburgh (PA), Atlanta (GA), Greenville (SC), Columbus (OH), Ann Arbor (MI), and Lansing (MI)—also appear on this list. Unsurprisingly, cities such as New York, Bellingham (WA), and San Diego appear at the bottom of Figure 7. Notably, New York City has a value of \$1.6K, implying that the NYC wage premium is \$1.6K lower than the Bay Area premium.

The orange circles represent real terms adjusted based on BEA RPPs. We see that the Bay Area premium persists, though to a lesser extent, after accounting for the high cost of living in the Bay Area.

To place the magnitude of the Bay Area premium in context, we compare it with the wage premium associated with university ranking tiers. In nominal terms, attending a Top 20 university carries a premium of \$25K relative to graduates from institutions outside the Top 1000.[34] Because Top 20 universities are more likely to be located in less expensive cities than institutions outside the Top 1000, we find a real wage premium of \$18K for graduates from Top 20 universities.

Using universities in the Midwest or Southwest as the benchmark, the Bay Area premium ranges from \$6.4K to \$10.7K in nominal terms and from \$4.1K to \$8.3K in real terms. This corresponds to at least 40% of the nominal premium and more than 20% of the real premium associated with Top 20 universities.

7 Conclusions

This paper constructs a rich individual-level dataset by combining detailed information from LinkedIn profiles, job postings from Burning Glass Technologies, and Glassdoor data to study the geographic mobility and job search behavior of recent college graduates. We examine how these outcomes vary across universities of different rankings and locations.

Our data provide detailed information on the college attended, job tasks at the firm–occupation level, and both the location of the college and the current workplace. These features are not available in commonly used matched employer–employee datasets. This allows us to study sorting along task, amenity, and spatial dimensions within a unified framework, distinct from the commonly estimated AKM approach.

Estimating a model of graduates' sorting into firms and occupations, we find evidence of positive sorting: graduates from higher-ranked universities are more likely to match into cognitively intensive jobs and to locate in high-amenity cities. At the same time, geographic distance significantly reduces job-matching probabilities for most universities, while its effect is substantially attenuated for elite institutions. Overall, geo-

[34] Graduates from Top 20 universities earn an average salary of \$81K, compared with \$56K for those from outside the Top 1000.

graphic frictions tend to narrow labor market disparities across colleges, whereas sorting on tasks and amenities widens earnings gaps across colleges.

The retention rate for local graduates is notably low for high-ranking universities and for cities that lack high-paying jobs, raising important questions about how cities can retain their college-educated talent. Investing in higher education is one of the major public expenditures in the U.S. and has long been viewed as a key driver of long-term economic growth. Policies aimed at attracting productive firms may be more effective in fostering growth and development. Such policies may include offering tax incentives to firms, improving infrastructure, and fostering a business-friendly regulatory environment. We hope to explore these questions further in future research.

Table 3: Estimates of Equation (16)

	(1)	(2)	(3)	(4)	(5)
Cognitive × Top20	0.200***	0.190***	0.167***	0.168***	0.170***
	(0.043)	(0.042)	(0.037)	(0.037)	(0.037)
Cognitive × Top21-200	0.074***	0.075***	0.086***	0.085***	0.083***
	(0.020)	(0.021)	(0.019)	(0.019)	(0.019)
Cognitive × Top201-1000	0.030*	0.032**	0.040***	0.040***	0.039***
	(0.015)	(0.015)	(0.014)	(0.014)	(0.014)
Social × Top20	0.057	0.049	0.036	0.035	0.033
	(0.037)	(0.037)	(0.031)	(0.031)	(0.032)
Social × Top21-200	0.005	0.004	0.003	0.003	0.004
	(0.020)	(0.020)	(0.017)	(0.017)	(0.018)
Social × Top201-1000	-0.030*	-0.031*	-0.022	-0.022	-0.021
	(0.018)	(0.018)	(0.016)	(0.016)	(0.016)
Routine × Top20	-0.057	-0.057	-0.023	-0.021	-0.022
	(0.046)	(0.045)	(0.036)	(0.036)	(0.037)
Routine × Top21-200	0.009	0.011	0.017	0.017	0.018
	(0.017)	(0.017)	(0.015)	(0.015)	(0.014)
Routine × Top201-1000	0.013	0.015	0.013	0.013	0.014
	(0.014)	(0.014)	(0.012)	(0.012)	(0.012)
Manual × Top20	-0.014	-0.016	-0.005	-0.007	-0.008
	(0.032)	(0.032)	(0.028)	(0.028)	(0.028)
Manual × Top21-200	-0.026	-0.026	-0.042***	-0.042***	-0.040**
	(0.018)	(0.018)	(0.016)	(0.016)	(0.016)
Manual × Top201-1000	-0.044***	-0.043***	-0.048***	-0.048***	-0.046***
	(0.015)	(0.015)	(0.014)	(0.014)	(0.014)
Amenity × Top20		0.092**	0.148***	0.153***	0.153***
		(0.036)	(0.033)	(0.033)	(0.033)
Amenity × Top21-200		-0.003	0.030**	0.030**	0.030**
		(0.013)	(0.013)	(0.013)	(0.013)
Amenity × Top201-1000		-0.028**	0.005	0.005	0.005
		(0.013)	(0.010)	(0.010)	(0.010)
Ldist			-0.077***	-0.069***	-0.067***
			(0.011)	(0.021)	(0.020)
Ldist × Top20			0.071***	0.069**	0.069**
			(0.024)	(0.028)	(0.028)
Ldist × Top21-200			0.027	0.028	0.028
			(0.021)	(0.021)	(0.021)
Ldist × Top201-1000			-0.018	-0.015	-0.015
			(0.013)	(0.013)	(0.013)
CZ			0.440***	0.339***	0.341***
			(0.030)	(0.088)	(0.088)
CZ × Top20			0.186*	0.234**	0.232**
			(0.101)	(0.107)	(0.106)
CZ × Top21-200			0.157***	0.192***	0.191***
			(0.056)	(0.059)	(0.059)
CZ × Top201-1000			0.094**	0.107**	0.106**
			(0.045)	(0.043)	(0.043)

Table 3: Estimates of Equation (16) (Continued)

	(1)	(2)	(3)	(4)	(5)
State			0.073***	-0.027	-0.027
			(0.026)	(0.067)	(0.067)
State × Top20			0.210**	0.249**	0.251**
			(0.097)	(0.099)	(0.099)
State × Top21-200			0.160***	0.169***	0.169***
			(0.043)	(0.044)	(0.044)
State × Top201-1000			0.048	0.055	0.055
			(0.034)	(0.034)	(0.034)
Ldist × In-State				-0.015	-0.016
				(0.024)	(0.024)
CZ × In-State				0.119	0.116
				(0.104)	(0.104)
State × In-State				0.146*	0.147*
				(0.083)	(0.083)
Cognitive × Public					0.016
					(0.018)
Social × Public					-0.010
					(0.017)
Routine × Public					-0.012
					(0.017)
Manual × Public					-0.019
					(0.014)
Observations	84,718	84,175	84,175	84,175	84,175
Adjusted R^2	0.67	0.67	0.74	0.74	0.74

Notes: Columns (1)-(3) report the estimated coefficients for equation (16). Column (1) has the regressors as the interaction of ranking dummies and BGT tasks; Column (2) adds amenities, and Column (3) adds geographic variables. Column (4) adds interactions of in-state student enrollment share. Column (5) adds interactions between public-university status and task measures. The models are estimated using OLS using the full sample. The full sample covers 22 occupations, 25246 firms, and 264 universities. Standard errors are clustered at the firm-occupation level.

Table 4: OLS Estimates of Equation (16) using Sub-Samples

	(1)	(2)
Cognitive × Top20	0.196***	0.186***
	(0.053)	(0.049)
Cognitive × Top21-200	0.084***	0.078***
	(0.019)	(0.020)
Cognitive × Top201-1000	0.040***	0.034**
	(0.014)	(0.016)
Social × Top20	0.025	0.035
	(0.045)	(0.043)
Social × Top21-200	0.004	0.018
	(0.017)	(0.019)
Social × Top201-1000	-0.021	-0.018
	(0.015)	(0.017)
Routine × Top20	-0.029	-0.018
	(0.052)	(0.051)
Routine × Top21-200	0.017	0.018
	(0.014)	(0.015)
Routine × Top201-1000	0.013	0.012
	(0.012)	(0.013)
Manual × Top20	-0.021	-0.016
	(0.041)	(0.039)
Manual × Top21-200	-0.043***	-0.042**
	(0.016)	(0.018)
Manual × Top201-1000	-0.048***	-0.051***
	(0.014)	(0.016)
Amenity × Top20	0.198***	0.198***
	(0.040)	(0.039)
Amenity × Top21-200	0.032**	0.032**
	(0.013)	(0.013)
Amenity × Top201-1000	0.005	0.004
	(0.010)	(0.010)
Ldist	-0.051**	-0.085**
	(0.020)	(0.039)
Ldist × Top20	0.160***	0.168***
	(0.058)	(0.060)
Ldist × Top21-200	0.026	0.027
	(0.020)	(0.024)
Ldist × Top201-1000	-0.016	-0.002
	(0.013)	(0.017)
CZ	0.372***	0.667***
	(0.090)	(0.164)
CZ × Top20	0.364**	0.274*
	(0.148)	(0.151)
CZ × Top21-200	0.185***	0.090
	(0.059)	(0.070)
CZ × Top201-1000	0.103**	0.075
	(0.043)	(0.046)

Table 4: OLS Estimates of Equation (16) using Sub-Samples (Continued)

	(1)	(2)
State	-0.036	-0.120
	(0.067)	(0.111)
State × Top20	0.353**	0.363**
	(0.147)	(0.148)
State × Top21-200	0.166***	0.172***
	(0.044)	(0.051)
State × Top201-1000	0.053	0.046
	(0.034)	(0.039)
Ldist × In-State	-0.040*	-0.014
	(0.024)	(0.044)
CZ × In-State	0.075	-0.232
	(0.106)	(0.182)
State × In-State	0.154*	0.237*
	(0.084)	(0.129)
Observations	80,727	65,612
Adjusted R^2	0.75	0.74

Notes: This table reports the estimated coefficients for equation (16). All columns report OLS estimates. Column (1) uses a subsample that excludes private universities ranked in the global Top 20 or belonging to the Ivy League. Column (2) uses a subsample that includes only public universities. Standard errors are clustered at the university level and reported in parentheses.

Table 5: OLS Estimates of Equation (16) with Additional Characteristics

	(1)	(2)	(3)	(4)
Cognitive × Top20	0.168***	0.187***	0.148***	0.120***
	(0.037)	(0.039)	(0.043)	(0.037)
Cognitive × Top21-200	0.085***	0.093***	0.072***	0.048***
	(0.019)	(0.021)	(0.022)	(0.017)
Cognitive × Top201-1000	0.040***	0.042***	0.033**	0.009
	(0.014)	(0.014)	(0.015)	(0.011)
Social × Top20	0.035	0.016	0.035	0.016
	(0.031)	(0.035)	(0.038)	(0.031)
Social × Top21-200	0.003	-0.013	-0.002	-0.008
	(0.017)	(0.019)	(0.021)	(0.015)
Social × Top201-1000	-0.022	-0.026	-0.020	-0.015
	(0.016)	(0.016)	(0.016)	(0.012)
Routine × Top20	-0.021	-0.027	-0.027	-0.016
	(0.036)	(0.040)	(0.042)	(0.037)
Routine × Top21-200	0.017	0.015	0.015	0.022
	(0.015)	(0.017)	(0.018)	(0.014)
Routine × Top201-1000	0.013	0.013	0.013	0.016
	(0.012)	(0.013)	(0.013)	(0.010)
Manual × Top20	-0.007	-0.024	-0.025	-0.023
	(0.028)	(0.029)	(0.031)	(0.026)
Manual × Top21-200	-0.042***	-0.059***	-0.059***	-0.048***
	(0.016)	(0.017)	(0.018)	(0.013)
Manual × Top201-1000	-0.048***	-0.052***	-0.052***	-0.041***
	(0.014)	(0.014)	(0.015)	(0.010)
Amenity × Top20	0.153***	0.176***	0.138***	0.107***
	(0.033)	(0.035)	(0.035)	(0.033)
Amenity × Top21-200	0.030**	0.039***	0.009	-0.005
	(0.013)	(0.014)	(0.016)	(0.014)
Amenity × Top201-1000	0.005	0.002	-0.018	-0.029**
	(0.010)	(0.012)	(0.014)	(0.012)
Ldist	-0.069***	-0.038	-0.059**	-0.035
	(0.021)	(0.026)	(0.027)	(0.025)
Ldist × Top20	0.069**	0.066**	0.047*	0.032
	(0.028)	(0.029)	(0.028)	(0.026)
Ldist × Top21-200	0.028	0.016	0.002	-0.007
	(0.021)	(0.019)	(0.018)	(0.016)
Ldist × Top201-1000	-0.015	-0.021	-0.028**	-0.023*
	(0.013)	(0.013)	(0.013)	(0.012)
CZ	0.339***	0.223**	0.294***	0.096
	(0.088)	(0.104)	(0.109)	(0.097)
CZ × Top20	0.234**	0.360***	0.374***	0.327***
	(0.107)	(0.121)	(0.125)	(0.115)
CZ × Top21-200	0.192***	0.297***	0.290***	0.208***
	(0.059)	(0.062)	(0.063)	(0.054)
CZ × Top201-1000	0.107**	0.126***	0.121***	0.072**
	(0.043)	(0.044)	(0.045)	(0.037)
State	-0.027	0.089	-0.016	0.078
	(0.067)	(0.075)	(0.081)	(0.073)
State × Top20	0.249**	0.230**	0.156	0.111
	(0.099)	(0.110)	(0.112)	(0.102)
State × Top21-200	0.169***	0.132***	0.091**	0.064
	(0.044)	(0.044)	(0.045)	(0.041)
State × Top201-1000	0.055	0.033	0.013	0.007
	(0.034)	(0.035)	(0.035)	(0.031)

Table 5: OLS Estimates of Equation (16) with Additional Characteristics (Continued)

	(1)	(2)	(3)	(4)
Cognitive × Most Selective		-0.025	-0.027	-0.023
		(0.032)	(0.031)	(0.028)
Cognitive × Highly Selective		0.038	0.029	0.025
		(0.024)	(0.023)	(0.017)
Cognitive × Selective		0.014	0.019	0.004
		(0.015)	(0.015)	(0.011)
Social × Most Selective		-0.016	-0.013	-0.006
		(0.029)	(0.028)	(0.022)
Social × Highly Selective		-0.000	0.006	-0.001
		(0.023)	(0.023)	(0.017)
Social × Selective		-0.032**	-0.031**	-0.016
		(0.015)	(0.015)	(0.011)
Routine × Most Selective		0.009	0.010	0.005
		(0.028)	(0.027)	(0.023)
Routine × Highly Selective		-0.025	-0.025	-0.016
		(0.020)	(0.020)	(0.016)
Routine × Selective		-0.009	-0.009	-0.011
		(0.013)	(0.013)	(0.010)
Manual × Most Selective		-0.010	-0.009	-0.018
		(0.023)	(0.023)	(0.019)
Manual × Highly Selective		0.009	0.010	0.001
		(0.020)	(0.019)	(0.014)
Manual × Selective		-0.032**	-0.031**	-0.013
		(0.014)	(0.014)	(0.010)
Amenity × Most Selective		-0.036*	-0.048**	-0.036**
		(0.020)	(0.020)	(0.017)
Amenity × Highly Selective		-0.014	-0.030	-0.014
		(0.023)	(0.023)	(0.018)
Amenity × Selective		0.012	-0.001	0.008
		(0.010)	(0.011)	(0.010)
Ldist × Most Selective		-0.035	-0.044*	-0.047**
		(0.022)	(0.023)	(0.022)
Ldist × Highly Selective		0.012	-0.000	-0.004
		(0.025)	(0.027)	(0.025)
Ldist × Selective		-0.039***	-0.037**	-0.028**
		(0.015)	(0.015)	(0.014)
CZ × Most Selective		0.062	0.042	0.085
		(0.100)	(0.098)	(0.087)
CZ × Highly Selective		0.071	0.056	0.063
		(0.080)	(0.081)	(0.070)
CZ × Selective		0.243***	0.226***	0.182***
		(0.054)	(0.054)	(0.047)
State × Most Selective		-0.133*	-0.151**	-0.146**
		(0.080)	(0.076)	(0.069)
State × Highly Selective		-0.014	-0.037	-0.026
		(0.067)	(0.068)	(0.061)
State × Selective		-0.129***	-0.115***	-0.103***
		(0.037)	(0.037)	(0.035)

Table 5: OLS Estimates of Equation (16) with Additional Characteristics (Continued)

	(1)	(2)	(3)	(4)
Ldist × In-State	-0.015	-0.018	-0.017	-0.017
	(0.024)	(0.028)	(0.028)	(0.025)
CZ × In-State	0.119	0.021	-0.004	0.089
	(0.104)	(0.124)	(0.123)	(0.106)
State × In-State	0.146*	0.133	0.149*	0.029
	(0.083)	(0.087)	(0.086)	(0.075)
Cognitive × Stem			0.223***	0.191***
			(0.068)	(0.052)
Social × Stem			-0.102	-0.070
			(0.064)	(0.046)
Routine × Stem			-0.000	-0.019
			(0.059)	(0.043)
Manual × Stem			0.004	-0.000
			(0.049)	(0.039)
Ldist × Stem			0.119**	0.097*
			(0.056)	(0.053)
CZ × Stem			-0.158	0.019
			(0.191)	(0.173)
State × Stem			0.469***	0.276*
			(0.159)	(0.146)
Amenity × Stem			0.135***	0.114***
			(0.048)	(0.041)
Network				1.669***
				(0.048)
Observations	84,175	84,175	84,089	84,089
Adjusted R^2	0.74	0.74	0.74	0.79

Notes: This table reports the estimated coefficients for equation (16). All columns are OLS estimates using the full sample. The full sample covers 22 occupations, 25246 firms, and 264 universities. Standard errors are clustered by university and reported in parentheses.

References

Abowd, J. M., Kramarz, F. and Margolis, D. N. (1999), 'High wage workers and high wage firms', *Econometrica* **67**(2), 251–333.

Acemoglu, D. and Autor, D. (2011), Skills, tasks and technologies: Implications for employment and earnings, *in* 'Handbook of labor economics', Vol. 4, Elsevier, pp. 1043–1171.

Albert, C. and Monras, J. (2022), 'Immigration and spatial equilibrium: the role of expenditures in the country of origin', *American Economic Review* **112**(11), 3763–3802.

Amanzadeh, N., Kermani, A. and McQuade, T. (2024), Return migration and human capital flows, Technical report, National Bureau of Economic Research.

Atalay, E., Phongthiengtham, P., Sotelo, S. and Tannenbaum, D. (2018), 'New technologies and the labor market', *Journal of Monetary Economics* **97**, 48–67.

Atalay, E., Phongthiengtham, P., Sotelo, S. and Tannenbaum, D. (2020), 'The evolution of work in the united states', *American Economic Journal: Applied Economics* **12**, 1–36.

Atalay, E., Sotelo, S. and Tannenbaum, D. (2024), 'The geography of job tasks', *Journal of Labor Economics* **42**(4), 979–1008.

Autor, D., Dorn, D., Hanson, G. H., Jones, M. R. and Setzler, B. (2025), Places versus people: the ins and outs of labor market adjustment to globalization, Technical report, National Bureau of Economic Research.

Auxier, B., Anderson, M. et al. (2021), 'Social media use in 2021', *Pew Research Center* **1**(1), 1–4.

Barron's Educational Series, I., ed. (2017), *Barron's Profiles of American Colleges 2017*, 33rd edn, Barron's Educational Series, Hauppauge, NY.

Batra, H., Michaud, A. and Mongey, S. (2023), Online job posts contain very little wage information, Technical report, National Bureau of Economic Research.

Becker, G. S. (1973), 'A theory of marriage: Part i', *Journal of Political Economy* **81**(4), 813–846.

Bonhomme, S., Holzheu, K., Lamadon, T., Manresa, E., Mogstad, M. and Setzler, B. (2023), 'How much should we trust estimates of firm effects and worker sorting?', *Journal of Labor Economics* **41**(2), 291–322.

Bound, J., Braga, B., Khanna, G. and Turner, S. (2020), 'A passage to america: University funding and international students', *American Economic Journal: Economic Policy* **12**(1), 97–126.

Braxton, J. C. and Taska, B. (2023), 'Technological change and the consequences of job loss', *American Economic Review* **113**(2), 279–316.

Burstein, A., Morales, E. and Vogel, J. (2019), 'Changes in between-group inequality: computers, occupations, and international trade', *American Economic Journal: Macroeconomics* **11**(2), 348–400.

Cadena, B. C. and Kovak, B. K. (2016), 'Immigrants equilibrate local labor markets: Evidence from the great recession', *American Economic Journal: Applied Economics* **8**(1), 257–90.

Card, D. (2001), 'Immigrant inflows, native outflows, and the local labor market impacts of higher immigration', *Journal of Labor Economics* **19**(1), 22–64.

Card, D., Heining, J. and Kline, P. (2013), 'Workplace heterogeneity and the rise of west german wage inequality', *The Quarterly journal of economics* **128**(3), 967–1015.

Chetty, R., Deming, D. J. and Friedman, J. N. (2023), Diversifying society's leaders? the determinants and causal effects of admission to highly selective private colleges, Technical report, National Bureau of Economic Research.

Chetty, R., Friedman, J. N., Saez, E., Turner, N. and Yagan, D. (2020), 'Income segregation and intergenerational mobility across colleges in the united states', *The Quarterly Journal of Economics* **135**(3), 1567–1633.

Conzelmann, J. G., Hemelt, S. W., Hershbein, B. J., Martin, S., Simon, A. and Stange, K. M. (2025), 'Grads on the go: Measuring college-specific labor markets for graduates', *Journal of Policy Analysis and Management* **44**(3), 741–763.

Costinot, A. and Vogel, J. (2010), 'Matching and inequality in the world economy', *Journal of Political Economy* **118**(4), 747–786.

Costinot, A. and Vogel, J. (2015), 'Beyond ricardo: Assignment models in international trade', *Annual Review of Economics* **7**(1), 31–62.

Dale, S. B. and Krueger, A. B. (2002), 'Estimating the payoff to attending a more selective college: An application of selection on observables and unobservables', *The Quarterly Journal of Economics* **117**(4), 1491–1527.

Deming, D. J. (2017), 'The growing importance of social skills in the labor market', *The Quarterly Journal of Economics* **132**(4), 1593–1640.

Deming, D. J. and Noray, K. (2020), 'Earnings dynamics, changing job skills, and stem careers', *The Quarterly Journal of Economics* **135**(4), 1965–2005.

Deming, D. and Kahn, L. B. (2018), 'Skill requirements across firms and labor markets: Evidence from job postings for professionals', *Journal of Labor Economics* **36**(S1), S337–S369.

Diamond, R. (2016), 'The determinants and welfare implications of us workers' diverging location choices by skill: 1980–2000', *American Economic Review* **106**(3), 479–524.

Diamond, R. and Gaubert, C. (2022), 'Spatial sorting and inequality', *Annual Review of Economics* **14**, 795–819.

Diamond, R. and Moretti, E. (2024), 'Where is standard of living the highest? local prices and the geography of consumption'.

Dorn, D., Schoner, F., Seebacher, M., Simon, L. and Woessmann, L. (2025), 'Multidimensional skills on linkedin profiles: Measuring human capital and the gender skill gap'.

Eeckhout, J. (2018), 'Sorting in the labor market', *Annual Review of Economics* **10**(1), 1–29.

Eeckhout, J. and Kircher, P. (2011), 'Identifying sorting—in theory', *The Review of Economic Studies* **78**(3), 872–906.

Greenland, A., Lopresti, J. and McHenry, P. (2019), 'Import competition and internal migration', *Review of Economics and Statistics* **101**(1), 44–59.

Hall, B. H., Jaffe, A. B. and Trajtenberg, M. (2001), 'The nber patent citation data file: Lessons, insights and methodological tools'.

Hanson, G. H. and Liu, C. (2016), High-skilled immigration and the comparative advantage of foreign-born workers across us occupations, *in* 'Talent Flows in the Global Economy', University of Chicago Press.

Hanson, G. and Liu, C. (2023), 'Immigration and occupational comparative advantage', *Journal of International Economics* **145**, 103809.

Hanushek, E. A. and Kimko, D. D. (2000), 'Schooling, labor-force quality, and the growth of nations', *American economic review* pp. 1184–1208.

Hershbein, B. and Kahn, L. B. (2018), 'Do recessions accelerate routine-biased technological change? evidence from vacancy postings', *American Economic Review* **108**(7), 1737–72.

Kennan, J. and Walker, J. R. (2011), 'The effect of expected income on individual migration decisions', *Econometrica* **79**(1), 211–251.

Khanna, G. and Morales, N. (2025), 'The IT boom and other unintended consequences of chasing the american dream'.

Lagakos, D. and Waugh, M. E. (2013), 'Selection, agriculture, and cross-country productivity differences', *American Economic Review* **103**(2), 948–80.

Lopes de Melo, R. (2018), 'Firm wage differentials and labor market sorting: Reconciling theory and evidence', *Journal of Political Economy* **126**(1), 313–346.

Martellini, P., Schoellman, T. and Sockin, J. (2024), 'The global distribution of college graduate quality', *Journal of Political Economy* **132**(2), 434–483.

Moretti, E. (2013), 'Real wage inequality', *American Economic Journal: Applied Economics* **5**(1), 65–103.

Munshi, K. (2003), 'Networks in the modern economy: Mexican migrants in the us labor market', *The Quarterly Journal of Economics* **118**(2), 549–599.

Pellegrina, H. S. and Sotelo, S. (2021), Migration, specialization, and trade: Evidence from brazil's march to the west, Technical report, National Bureau of Economic Research.

Ruggles, S., Alexander, J. T., Genadek, K., Goeken, R., Schroeder, M. B., Sobek, M. et al. (2010), 'Integrated public use microdata series: Version 5.0 [machine-readable database]', *Minneapolis: University of Minnesota* **42**.

Song, J., Price, D. J., Guvenen, F., Bloom, N. and Von Wachter, T. (2019), 'Firming up inequality', *The Quarterly journal of economics* **134**(1), 1–50.

Spitz-Oener, A. (2006), 'Technical change, job tasks, and rising educational demands: Looking outside the wage structure', *Journal of labor economics* **24**(2), 235–270.

Tombe, T. and Zhu, X. (2019), 'Trade, migration, and productivity: A quantitative analysis of china', *American Economic Review* **109**(5), 1843–72.

Yamaguchi, S. (2012), 'Tasks and heterogeneous human capital', *Journal of Labor Economics* **30**(1), 1–53.

Zimmerman, S. D. (2019), 'Elite colleges and upward mobility to top jobs and top incomes', *American Economic Review* **109**(1), 1–47.

Online Appendix

College Graduates in the Labor Market: Geographic Mobility and Sorting into Firms and Occupations

Chen Liu
NUS

Sibo Liu
HKPU

Yifan Zhang
CUHK

A Data Appendix

A.1 Data Description

The LinkedIn Data. Our primary data are purchased from Revelio Labs, which compiles information from publicly available LinkedIn profiles and other sources. LinkedIn is recognized as a major online platform for professional networking, where individuals voluntarily provide their work experiences and educational backgrounds for job search and career development purposes. As of 2023, LinkedIn has more than 1 billion registered members across over 200 countries and territories.[35] LinkedIn user profiles essentially function as self-reported resumes, containing detailed information on individuals' educational and employment histories. These include the universities they attended, the degrees and fields of study they pursued, their employers, job titles, and the dates during which they held these positions. In the United States, a majority of college graduates use LinkedIn (Auxier, Anderson et al., 2021). Our version of the LinkedIn data from Revelio Labs has only recently begun to be used in economics research. A few recent studies have employed it to examine the returns to international migration (Amanzadeh, Kermani and McQuade, 2024), Indian engineering migrants to the United States (Khanna and Morales, 2025), and the gender skill gap (Dorn, Schoner, Seebacher, Simon and Woessmann, 2025).

Because university names, employer names, and job titles are all self-reported in LinkedIn profiles, we undertake extensive work to harmonize and standardize this information across our three sources (LinkedIn, Burning Glass, and Glassdoor). Employer names are cleaned to ensure consistency across three platforms. The database for LinkedIn (Revelio Labs) provides OCCSOC occupation codes for user self-reported job titles using internal text-based algorithms. For job titles in Glassdoor, we employ a large language model (ChatGPT 4o) to obtain OCCSOC occupation codes. University names are also systematically standardized and linked to external sources of institutional data. To capture university location, we assign each institution a city and state based on its primary campus location, and then map these to geographic codes: commuting zone (CZ) and state codes.

Finally, LinkedIn profiles include detailed geographic information about individuals' current jobs. Crucially, this data specifies the location of each job—not just the employer's headquarters—at multiple levels, including state, metropolitan area, city, and even street address. This granularity enables us to analyze graduates' geographic mobility based on the actual location of their employment. Specifically, we use this job location information to identify where graduates are employed and compare it with the

[35] For more details, please refer to the Wikipedia page of LinkedIn.

location of their alma mater. This comparison allows us to measure geographic mobility relative to where graduates attended college.

To construct a consistent sample across different sources, we create a cross-sectional data snapshot for the year 2018. Specifically, because our focus is on job matching among fresh college graduates, we restrict the LinkedIn sample to individuals who earned a bachelor's degree (as their highest degree) between 2016 and 2018, received their college education at a U.S. institution, and were employed by a U.S. firm in 2018. Our objective is to identify each individual's first "primary" job immediately following graduation. For those reporting multiple concurrent jobs in their profile, we define the primary position as the one in which they had the longest tenure. We exclude individuals who were working as interns or were still enrolled in graduate programs (master's or doctoral) at the time. Finally, we use firm and occupation information from their 2018 job records, which we then match with data from Burning Glass Technologies (BGT), the 2018 American Community Survey (ACS), and the U.S. Bureau of Economic Analysis (BEA).

The Burning Glass Data. The data contain the universal job posting data collected by Burning Glass Technologies (BGT), and were first used in Hershbein and Kahn (2018). We use four BGT task variables commonly used in the literature: cognitive, social, routine, and manual. Following Spitz-Oener (2006), Atalay, Phongthiengtham, Sotelo and Tannenbaum (2020), and Deming and Kahn (2018), we measure these task variables from job advertisements based on keywords.

The BGT data have already processed job titles and mapped them to OCCSOC codes. Location information for vacancies is also standardized and available at multiple geographic levels (MSA, CZ, and state codes), which we can directly use. Employer names, however, are not fully cleaned. We harmonize and standardize these to create a crosswalk with LinkedIn and Glassdoor (see Appendix A.2).

We measure task intensity using Burning Glass Technologies (BGT) job advertisements, focusing on cognitive, social, routine, and manual tasks as captured by the skill requirements listed in job ads. Specifically, BGT reports whether a given skill (from thousands of listed skills) is required for each job title. For each task category, we count the number of required skills and compute task measures as percentile rankings across all postings. We then take the average across postings that share the same firm name and occupation code. These BGT task variables are thus measured at the firm and occupation levels, and each variable is standardized to range between 0 and 1.

For cognitive tasks, we base our selection on two sets of words. The first set includes those related to Spitz-Oener (2006). The second set incorporates words that refer to advanced computer software or skills. The selected words are as follows:

- problem solving, research, analytical, critical thinking, math, statistics, development
- Microsoft C#, Microsoft SQL, Microsoft server, social media platforms, virtual private networking, Microsoft visual C++, C (programming language), statistics, statistical software, software development, simulation software, scripting, sql databases and programming, neuroscience, machine learning, mathematics, aerospace engineering, application programming interface (API), application development, automation engineering, big data, C and C++, cache (computing), chemical engineering, cloud computing, computer hardware, data analysis, data mining, data science

For social tasks, we adopt the keywords based on Deming and Kahn (2018). These words are

- communication, teamwork, collaboration, negotiation, presentation, supervisory, leadership, management, mentoring, staff

For routine tasks, we based our selection on two sets of words. The first set includes words directly related to the administrative nature of jobs. The second set includes software commonly used by administrative staff.

- budget, accounting, cost, account management, admin, billing, administration, education administration
- Microsoft, spreadsheet, Photoshop, Google Docs, Google Maps, Google Drive, Google apps, Macintosh OS, YouTube, Facebook, payroll, accounting and finance software, administrative support, human resource management systems, human resources software, identity management, record keeping

Manual tasks use the following keywords:

- customer, client, service, physical abilities, repair, cleaning, sales

The Glassdoor Data. Our data source for salary is collected from Glassdoor. Glassdoor is an online platform where workers can review employers, report their earnings, and search for jobs. To encourage participation, Glassdoor uses a "give-to-get" model: users who submit an employer review or salary report gain access to others' anonymous data. Users share a wide range of information, including their compensation details such as base pay, bonuses, currency, and job-related information such as years of work experience, employment status, job title, location, and employer. To keep access, users must update their data annually if they have not submitted a recent review or salary report. For our research purposes, the dataset includes rich employer-employee matches with detailed information about individual workers.

We obtain a snapshot of Glassdoor data collected between September and October 2024, which includes detailed wage information by firm, occupation, location, and years of experience. The original sample contains 183,257 companies and 10.5 million wage records. Specifically, Glassdoor provides the average wage for a given job title within a company, aggregated by its internal algorithm. Of these, 5.04 million wage records are at the employer–title level, and 5.45 million are at the employer–title–location level. For our study, we match job titles from Glassdoor to OCCSOC occupation codes, company names to those in the Burning Glass and LinkedIn datasets, and job locations to commuting zone (CZ) codes.

A.2 Data Processing

Step 1: Standardizing Employer Names across Three Sources. Our analysis draws on firm-level information from Burning Glass Technologies (BGT), LinkedIn, and Glassdoor. Since our empirical analysis relies heavily on task variables from BGT, we standardize firm names across all three sources and focus primarily on the set of firms that appear in both LinkedIn and the BGT database. The original BGT records are at the firm-occupation-location-time level and are described in detail in Section A.2. Details on the standardization and matching procedures are provided below.

1. **BGT firm list.** We retrieve the original firm names from BGT, which include 462,295 unique,

non-standardized entries. Since the names in BGT are extracted from job postings and often vary in how company names are displayed, multiple records may correspond to the same firm.

Following the standardization method used in Hall, Jaffe and Trajtenberg (2001), we employ a series of cleaning routines for organization names and create two versions of cleaned firm names. The "standard name" retains basic firm-related words, such as "GROUP" or "INC", while the "stem name" is a shorter version that contains only the company's name that has been simplified to its most basic form.[36] Examples of the standardized results are provided in Table A.1. We use the BGT "stem name" as the tracking benchmark to match firm names across BGT, LinkedIn, and other databases. Firm names are cleaned and standardized using source-specific methods, with different matching techniques applied to align them to the BGT "stem name".

Table A.1: Examples of BGT Name Standardization

BGT name	Standard name	Stem name
Belimo	BELIMO	BELIMO
Belimo Air Controls	BELIMO AIR CONTROLS	BELIMO AIR CONTROLS
Belimo Air Controls Incorporated	BELIMO AIR CONTROLS INC	BELIMO AIR CONTROLS
Belimo Aircontrols Usa Incorporated	BELIMO AIRCONTROLS USA INC	BELIMO AIRCONTROLS USA
Belimo Americas	BELIMO AMERICAS	BELIMO AMERICAS
Tesla	TESLA	TESLA
Tesla Gigafactory	TESLA GIGAFACTORY	TESLA GIGAFACTORY
Tesla Incorporated	TESLA INC	TESLA
Tesla Motors	TESLA MOTORS	TESLA MOTORS

Notes: This table provides examples of how BGT names are standardized and stemmed.

After standardization, some firms may share the same standard name or stem name. For example, there are 455,238 unique standard names and 423,149 unique stem names out of 462,295 original BGT firm names. As shown in the table above, "TESLA", "TESLA MOTORS", and "TESLA GIGAFACTORY" essentially refer to the same firm, which we collapse into a single entity in the fuzzy matching procedure by considering standard names and stem names.

2. **Matching with LinkedIn firm list.** To align with other data sources (BGT, 2018 American Community Survey, etc) and our research design, we filter a list of LinkedIn users who had active job experience in 2018 and the associated jobs.[37] As employer names in LinkedIn are self-reported, multiple records may correspond to the same firm. To reconcile firm names with those in the BGT dataset for research purposes, we perform a similar standardization process to obtain both standardized and stemmed firm names. As before, the "stem name" is used to match firms across the BGT and LinkedIn datasets. Examples are provided in Table A.2.

   Again, after standardization, some firms may share the same standard name or stem name. For example, there are 5,699,863 unique standard names out of 5,899,025 unique original LinkedIn U.S. firm names. As shown in the table above, "JP MORGAN CHASE", "JP MORGAN CHASE BANK", and "JP MORGAN CHASE BANK NA" essentially refer to the same firm, which we collapse into a single entity in the fuzzy matching procedure. The number of records in LinkedIn firms is larger than that in the BGT firm list. We use the shorter BGT firm list as the master file

[36] The algorithm used can be downloaded here.

[37] Specifically, in our construction procedure, we extract all job records for those users from the LinkedIn database across all time periods. This approach yields a sample of 196,167,307 job records from 5,899,025 U.S. firm names, each assigned a unique ID by the data vendor. We clean and match with those firm names with BGT and Glassdoor.

Table A.2: Examples of LinkedIn Name Standardization

LinkedIn name	Standard name	Stem name
BJ TERRONI CO	BJ TERRONI CO	BJ TERRONI
B.J. Terroni Co., Inc.	BJ TERRONI CO INC	BJ TERRONI
B.J. Terroni Company, Inc	BJ TERRONI CO INC	BJ TERRONI
BJ Terroni Company, Inc	BJ TERRONI CO INC	BJ TERRONI
JPMorgan Chase	JP MORGAN CHASE	JP MORGAN CHASE
JP Morgan Chase	JP MORGAN CHASE	JP MORGAN CHASE
J.P. Morgan Chase	JP MORGAN CHASE	JP MORGAN CHASE
JPMorgan Chase Bank	JP MORGAN CHASE BANK	JP MORGAN CHASE BANK
JPMorgan Chase Bank, N.A.	JP MORGAN CHASE BANK NA	JP MORGAN CHASE BANK NA

Notes: This table provides examples of how company names found on LinkedIn are standardized and stemmed for analysis.

and perform a left join with the matched LinkedIn firms. To this end, we carry out a three-step matching procedure as follows.

- Exact matching using standard names. Initially, records are compared using the standard names from both lists. If the standard names are identical, the records are considered a match.
- Exact matching using stem names. For records not yet matched, the next step uses "stem names." Similar to the first step, but using a potentially simplified or base form of the names (e.g., removing suffixes or prefixes). Matches are made when stem names match exactly between the two lists.
- Fuzzy matching. For records that remain unmatched after exact matching attempts, a fuzzy matching technique is applied. This process is restricted to name pairs where the first six characters are identical. Fuzzy matching algorithms generate similarity scores based on standard names and based on stem names. Records are retained if both generated similarity scores are above 0.85.[38]

Through the procedure described above, we identify 51,363 disambiguated firm names that link LinkedIn records with BGT data. These disambiguated employers are associated with 75,037,667 jobs (firm–occupation–location), accounting for 38.3% of the original sample. Non-matched records can be attributed to three main factors: (1) we currently consider only LinkedIn positions from 2018, and therefore the LinkedIn sample does not fully overlap with the coverage of BGT; (2) the matching scheme may not capture all potential firm matches, as we adopt a conservative approach that prioritizes the accuracy of matched records; and (3) all firm names with stemmed name lengths shorter than three characters are excluded from the analysis. Among all matched records, 92% are based on exact matches using either standardized or stemmed names, while the remaining 8% rely on fuzzy matching. Examples of matched records are provided in Table A.3.

3. **Matching with Glassdoor firm list.** The original Glassdoor sample consists of 5,448,727 wage records at the employer-title-location level and an additional 5,048,711 records at the employer-title level. These wage records are associated with 183,267 unique employer names. Using the same matching procedures described above, we disambiguate these names and match them with firms in the BGT dataset. Specifically, we apply the standardization algorithm, match based on

[38]We use Stata function matchit to create the similarity scores. We experiment with alternative methods using different lengths of the initial characters as a starting point and different score cutoffs. The current approach yields similar results to other optimized combinations.

Table A.3: Examples of BGT and LinkedIn Name Matching Methods

Matching method	BGT name	LinkedIn name
Exact standard/stem name	Blue Ridge Sales	blue ridge sales inc
Exact standard/stem name	Blue Rose Consulting Llc	blue rose consulting group, inc.
Exact standard/stem name	Blue Sky Property	blue sky property group
Fuzzy matching	Blue Streak Reprographic	blue streak reprographics
Fuzzy matching	Blue Water Automotive Systems	blue water automotive, inc
Fuzzy matching	Blue Willow Counseling	blue willow counseling ctr

Notes: This table illustrates examples of exact and fuzzy matching between BGT and LinkedIn company names.

standardized and stemmed names, and employ the same fuzzy matching approach. As a result, we identify 12,175 disambiguated firm names that can be matched with our BGT sample. These matched employers account for 66.8% of the wage records in the Glassdoor dataset. Examples of matched records are provided in Table A.4.

Table A.4: Examples of BGT and Glassdoor Name Matching Methods

Matching method	BGT name	Glassdoor name
Exact standard/stem name	Bahwan Cybertek, Inc	Bahwan CyberTek
Exact standard/stem name	Bain Company Incorporated	Bain & Company
Exact standard/stem name	Worldwide Flight Services Incorporated	Worldwide Flight Services
Fuzzy matching	Zeeland Lumber Supply Company	Zeeland Lumber & Supply
Fuzzy matching	Marmic Fire Safety Company Incorporated	Marmic Fire & Safety
Fuzzy matching	Hard Rock Hotels And Casinos	Hard Rock Hotel & Casino

Notes: This table provides examples showing how company names from BGT are matched with Glassdoor names using exact and fuzzy matching methods.

Step 2: Standardizing OCCSOC Codes of Job Titles for Three Sources. Our study requires occupation information that can be linked to standard OCCSOC codes. The occupation titles associated with these codes are obtained from IPUMS USA. In total, 487 distinct occupations are identified. We summarize how this occupation information is utilized across the three main datasets used in our analysis.

1. **LinkedIn job titles.** The LinkedIn data provider, Revelio Labs, assigns an OCCSOC code to each job title using an internal text-based algorithm. In the 2018 LinkedIn dataset, 385 distinct OCCSOC codes are identified. Occupations not captured by this process are primarily labor-intensive roles, which are typically underrepresented in LinkedIn data.

2. **Burning job titles.** The job posting data from Burning Glass Technologies (BGT) includes OCCSOC codes directly, which have been validated and used in the literature (e.g., Braxton and Taska (2023)).

3. **Glassdoor job titles.** Job titles in the Glassdoor wage data are highly non-standard and noisy. From the full sample, we observe 572,691 distinct job title descriptions. To standardize these, we utilize ChatGPT-4o to match each title to the closest OCCSOC code, resulting in 563,294 successful matches (match rate=98.4%). A summary of the results is provided in Table A.5.

Step 3: Standardizing LinkedIn University Name and Matching to World Ranking and IPEDS. Our analysis requires university ranking and location. We perform the following procedures to obtain this information.

Table A.5: Examples of Job Title Matching Between Glassdoor and OCCSOC

Glassdoor Job Title	Matched OCCSOC Code	Matched OCCSOC Title
Assistant Vice President Consultant Risk Tech	15-1199.09	Risk Management Specialists
Senior Systems Data Analyst	15-2041.01	Data Scientists
Associate Audiovisual Technician	27-4011.00	Audio and Video Technicians
Tax Servicing Specialist	13-2082.00	Tax Preparers
Debt Consolidation	13-2072.00	Loan Officers
Financial Controller	11-3031.00	Financial Managers

Notes: This table shows examples of how job titles from Glassdoor are matched to OCCSOC codes and standardized occupational titles.

1. **LinkedIn university names.** LinkedIn provides user-generated education information, resulting in high variability. This includes differences in the type of degree, field of study, program details (if provided), and the university name. As shown in Table A.6, the self-reported university names are particularly noisy, as they often include various name variants, abbreviations, department names, campus or school designations, and entries in multiple languages. For example, column (1) shows how Harvard University is self-reported by LinkedIn users, and column (2) shows the example for Purdue University.

Table A.6: Examples of University Name Variations

Example 1 (Harvard)	**Example 2 (Purdue)**
Harvard	Purdue University
Harvard University	Purdue Global University
Harvard Law School	Purdue North Centeral
Harvard University Extension School	Purdue School of Engineering and Technology IUPUI
Harvard College	Purdue U Indiana U
Harvard Business School	Purdue College of Technology Columbus
HBS	Purdue University Calumet
Harvard University Kennedy School of Government	Purdue University Daniels School of Business
John F. Kennedy School of Government	Purdue University Global
Harvard T.H. Chan School of Public Health	Purdue University Krannert School of Management

Notes: This table presents examples of how the same institution can appear under different names, illustrating the need for name normalization in university data.

We filter degrees and associated university names for analysis using the following criteria:

- In our main sample, we first consider a list of users with a U.S.-based job and extract all education information for those users, obtaining 79,407,923 degree records.
- We retain degrees at the bachelor's level and above—specifically, bachelor's, master's, and doctoral degrees.
- We keep only records associated with U.S. universities. Since the country information is missing for many institutions, we infer the university's country based on the most frequent job location of its graduates.
- We drop records with a missing degree start year.
- We conduct preliminary cleaning by removing special symbols from university names.

The above procedure results in 31,696,635 degree records associated with 426,388 unique university name entries, which remain highly noisy.

2. **University rankings and a list of universities to consider.** The noisy nature of LinkedIn university names hinders our ability to disambiguate institutions and match them with external

datasets. To address these challenges, we construct a list of major U.S. universities to restrict our sample to a meaningful subset of records for further data cleaning. This list is based on prominent university rankings, which naturally provide the ranking needed for additional analysis. Specifically, we combine the following three sources:

- World University Rankings (WUR): WUR provides annual rankings for 2,000 global universities, among which we consider 348 universities located in the United States. For our study period, we use the 2019 edition of the WUR rankings.
- U.S. News[39]: The U.S. News ranking includes 2,145 universities, but only the Top 1,000 institutions are assigned a specific rank; universities beyond the Top 1,000 are unranked. Among these, 284 are U.S.-based universities, of which 197 are ranked.
- Universities considered in Chetty et al. (2020), including over 2000 U.S. universities.

There is substantial overlap across the three university lists, although institution names often appear in varied forms. We merge the lists and standardize university names before matching them to the LinkedIn database, ultimately identifying 2,300 distinct U.S. universities. The composition of the final university list used in our analysis is summarized in Table A.7. Specifically, we report the percentage of cleaned school names that appear in only a single data source versus those that overlap across multiple sources. Notably, 235 universities (10.2%) appear in all three lists, while the list from Chetty et al. (2020) alone covers 1,925 universities (83.7%).

Table A.7: Source of the University List

Sources	Freq	Percent
Chetty et al. (2020) (only)	1,925	83.70
U.S. News (only)	16	0.70
WUR (only)	30	1.30
U.S. News, WUR (two sources)	94	4.09
U.S. News, WUR, Chetty et al. (2020) (three sources)	235	10.22
Total	2,300	100.00

Notes: This table summarizes the sources used to determine the universities analyzed in our study. WUR refers to the World University Rankings.

3. **Matching LinkedIn university names.** We implement a multi-step matching procedure between LinkedIn university entries and the standardized school list. This process begins with 426,388 unique university name entries from LinkedIn, as filtered in the preceding stage.

    - **Exact and fuzzy matching.** To obtain matched records, we experiment with both exact and fuzzy matching approaches in an iterative fashion. We begin by matching records with exact character strings. We then compute similarity scores using a fuzzy matching algorithm.[40] Several heuristics prove useful in identifying matches. For example, when the first 12 characters of university names are identical and the similarity score exceeds 0.8, the records are generally correctly matched.
    - **First-30-character matching.** University names with the first 30 characters identical are also found to match with high probability.

[39]We retrieve the ranking information from U.S. News as of 3/13/2023.
[40]We again use the Stata function `matchit` to generate similarity scores.

- **Manual verification.** We manually verify unmatched records where university names appear more than 2,000 times in the entire database.
- **ChatGPT-assisted matching.** As explained above, raw university names from LinkedIn often include abbreviations, department names, school names, program names, or even multiple languages. We utilize ChatGPT to assist in identifying the correct matched institutions in such cases.

Through the matching procedure described above, we successfully linked 20,543 raw LinkedIn university name entries to our standardized list of U.S. universities, corresponding to 2,170 major institutions. These matched universities are associated with 26,691,378 degrees held by LinkedIn users working in 2018, representing 84.2% of the original filtered dataset (31,696,635 records).[41] A summary of the matched records is presented in Table A.8.

Table A.8: LinkedIn University Matching Methods

Match Method	Raw LinkedIn Universities Matched	Degrees Matched
Exact and fuzzy	2,684	8,639,110
First-30-character	11,320	50,746
Manual	119	742,885
ChatGPT-assisted	6,420	17,258,637
Total	20,543	26,691,378

Notes: This table provides the number of universities and degrees matched using various methods, including traditional exact/fuzzy matching, manual review, and ChatGPT-assisted matching.

Step 4: Standardizing the Geographic Information of Jobs and Universities. For our research purposes, we require city-level geographic information for the job locations listed in LinkedIn and Glassdoor, as well as for users' graduating universities.

1. **Job location in LinkedIn.** As explained above, we begin by extracting LinkedIn users with an active job position in 2018, and then retrieve all available job history associated with those users. The resulting dataset includes 3.89 million unique job location records. However, the data are noisy, as it includes geographic information at varying levels of granularity—such as state, county, city, and street levels. For our research purposes, we focus on extracting U.S. city-level information from the raw address field.

   - **Extract Cities.** We start with a list of 31,254 city names[42]. We search the job location addresses using city names as keywords and require the state information to be consistent. We also manually correct various issues, such as abbreviations (e.g., "NYC" for New York City) and ambiguous city names that appear in multiple states. After cleaning and validation, we successfully identify 2,998,495 records with reliable city-level information. The remaining records are manually reviewed, found to be inaccurate or incomplete, and subsequently excluded. Examples are provided below in Table A.9.

[41] The 26,691,378 LinkedIn users include individuals who graduated in any year, regardless of the employer or occupation reported.
[42] Source: SimpleMaps.

- **Match with CZ Codes.** The geographic coordinates of the extracted cities are obtained using the Google Geocoding API and are mapped to Commuting Zone (CZ) boundaries (based on the 1990 definition)[43], as well as to Metropolitan and Micropolitan Statistical Areas. Among the 75,037,667 job positions matched with firm names from the Burning Glass database, 68,663,767 (91.5%) are successfully associated with a city name and thus assigned with a CZ code.

Table A.9: Examples of Location Extraction and CZ Code Inference

Raw Location in LinkedIn	State	Extracted City	Inferred CZ Code
45 W. 111th Street, Chicago, IL 60627	Illinois	Chicago	16001
1356 Bellevue St. Green Bay, WI	Wisconsin	Green Bay	55004
5500 Cloverleaf Pkwy Cleveland	Ohio	Cleveland	39002
Shiawassee Michigan area.	Michigan	–	–
United States, MI, Orion Township	Michigan	–	–
Unit Number 4, DDA Local Shopping Centre, Hemkunt Co	Colorado	–	–
Central, Nebraska	Nebraska	–	–

Notes: This table shows examples of how raw LinkedIn location strings are parsed to extract U.S. states, cities, and commuting zone (CZ) codes.

2. **Job location in Glassdoor.** We also require city-level information for the Glassdoor wage records. Among 3,952,145 wage records with location information linked to employers matched with Burning Glass firms, there are 24,246 unique location descriptions. We apply the same matching approach described above. After this process, 8,913 records remain unmatched. Given that Glassdoor location data are relatively clean, we use the Google Geocoding API to extract precise geographic coordinates for these remaining records. In the final sample, we successfully obtained city and CZ information for 3,950,469 out of 3,952,145 wage records. Some examples are provided in Table A.10.

Table A.10: Examples of Location Extraction from Glassdoor and CZ Code Inference

Raw Location in Glassdoor	State	Extracted City	Inferred CZ Code
Austin, TX	Texas	Austin	31201
Bala Cynwyd, PA	Pennsylvania	Bala Cynwyd	19700
La Jolla, CA	California	La Jolla	38000
Whippany, NJ	New Jersey	Whippany	19600

Notes: This table shows how structured location data from Glassdoor job postings is parsed and linked to commuting zone (CZ) codes.

3. **University location in LinkedIn.** For each university, we extract geographic coordinates using the Google Geocoding API and map them to cities, Commuting Zone (CZ) boundaries, Metropolitan and Micropolitan Statistical Areas.

The data we processed includes college graduates from all U.S. institutions across all graduating years, covering 13.7 million U.S. employees. For our analysis, we restrict the sample to LinkedIn users who: (1) received a bachelor's degree (as their highest degree) between 2016 and 2018 from a U.S. institution ranked in the Top 2000 of the WUR; (2) were employed in the United States in 2018; (3) have employer names and occupational titles that can be clearly identified and matched to the BGT data; and (4) have identifiable employer geographic locations. To improve estimation precision, we further restrict the sample to U.S. universities with at least 100 LinkedIn users.

[43] Source: The Health Inequality Project.

We also restrict the analysis to 264 U.S. universities ranked in the WUR.[44] Our final sample covers 264 universities, 25246 distinct firms, and a total of 244,632 LinkedIn users.[45]

[44]WUR ranks the top 2,000 universities globally, of which 348 are U.S. universities.

[45]Throughout the paper, we define a firm as the combination of a company name and the location of its establishment. A firm enters our sample if at least one LinkedIn user is observed as employed there.

B Validating the Sample

Since BGT data have been widely used and validated in previous studies (Hershbein and Kahn, 2018, Atalay et al., 2024), we focus on validating our LinkedIn sample in various ways.

B.1 Validating the LinkedIn Sample

We first assess the spatial representativeness of the LinkedIn data by comparing it with the ACS. For both datasets, we restrict attention to workers whose highest degree is a college degree and who were employed in 2018. The left panel of Figure B.1 plots each commuting zone's share of national college graduate employment using LinkedIn data (y-axis) and ACS data (x-axis). The dashed line represents the 45-degree line. Most commuting zones lie close to this line, indicating a high degree of similarity in employment shares between the two datasets. Nonetheless, LinkedIn users appear to be overrepresented in a few major cities, such as New York, San Francisco, and Seattle, and underrepresented in others, such as Los Angeles, Newark. Across U.S. commuting zones, the correlation between city sizes implied by the two datasets is high: the correlation equals 0.95, and the OLS regression slope equals 1.09 (s.e. = 0.01).

We next examine the occupational representativeness of the LinkedIn data. The right panel of Figure B.1 plots the shares of employment across two-digit SOC occupations using LinkedIn data (y-axis) and ACS data (x-axis). Again, most points lie close to the 45-degree line. Perhaps as expected, LinkedIn users are disproportionately represented in high-skilled, business- and technology-oriented occupations such as business and finance, and computer and mathematics, while they are underrepresented in service and administrative occupations such as education, sales, and office administration. Once more, we find a strong correlation between the two datasets: the correlation equals 0.93, and the OLS regression slope equals 1.02 (s.e. = 0.09).

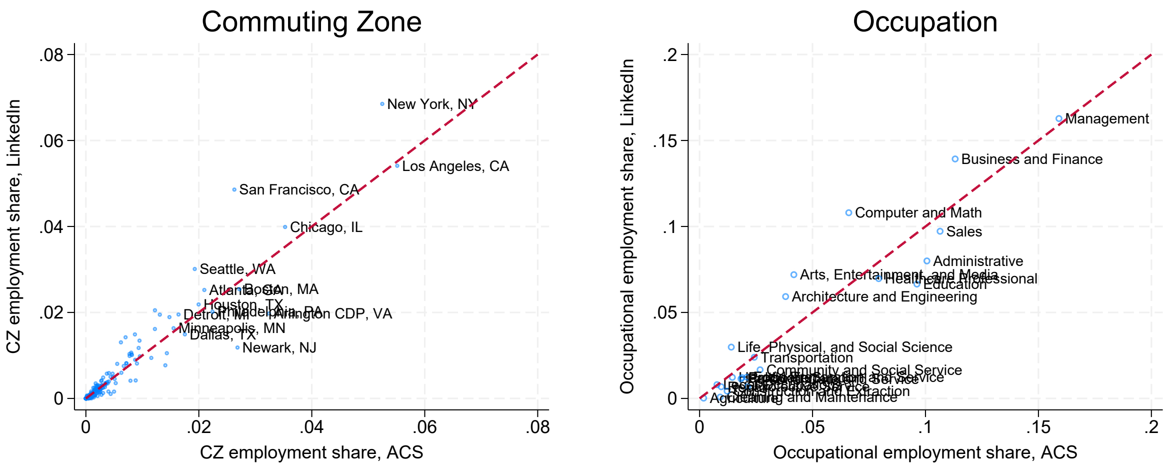

Figure B.1: Comparison of LinkedIn and IPUMS-ACS Data: Commuting Zones (left) and Occupations (right)

Our third validation exercise examines the extent to which LinkedIn data represent graduating class sizes across U.S. universities. Specifically, we compare the share of graduates from each university in

LinkedIn with that reported in the Integrated Postsecondary Education Data System (IPEDS), published by the National Center for Education Statistics. For both data, we use LinkedIn users who graduated between 2016 and 2018. Figure B.2 plots the national share of graduates by university, with LinkedIn data on the y-axis and IPEDS data on the x-axis. For comparability, shares are expressed as percentages. We find a strong positive relationship between the two measures. An Ordinary Least Squares (OLS) regression yields a coefficient of 1.23 (standard error = 0.063), with an R^2 of 0.50.

These results indicate that LinkedIn data are, to a considerable extent, representative of the U.S. college-educated labor force and graduating class sizes.

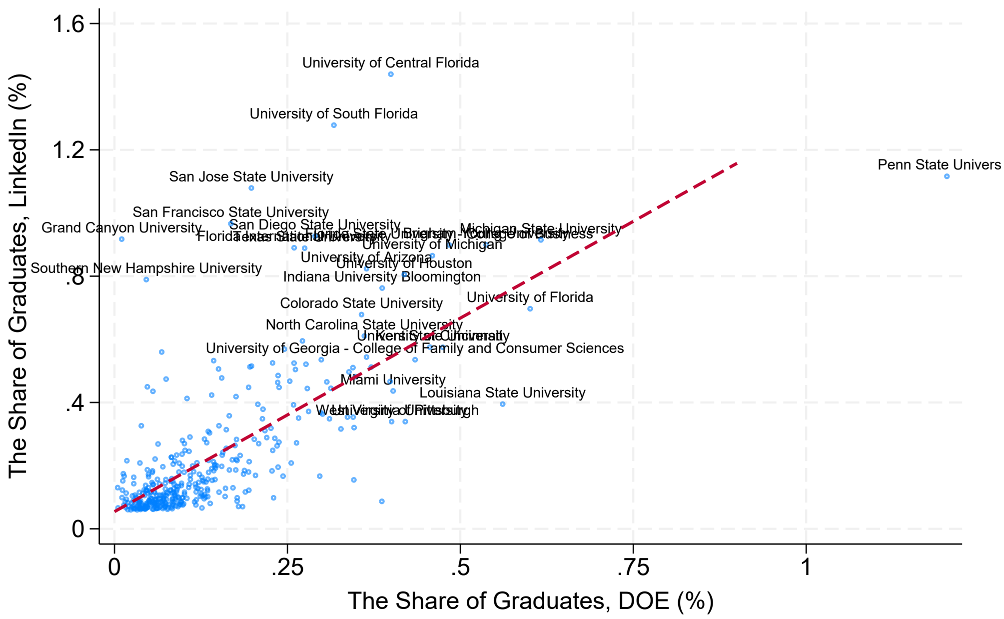

Figure B.2: Comparison of College Class Sizes: LinkedIn vs. IPEDS

B.2 Validating the Glassdoor Wages

Glassdoor wage data are available at the firm, city, and occupation levels. We validate these data in three ways. First, we compute the average Glassdoor wage by commuting zone and compare it with corresponding estimates from the ACS. In the ACS, wages are calculated for college graduates who are full-time workers (defined as working more than 35 hours per week and more than 40 weeks per year).

The left panel of Figure B.3 compares the average annual salaries from the two sources across 722 U.S. CZs. The dashed line represents the 45-degree line. The two measures are strongly correlated: a simple regression of Glassdoor wages on ACS wages yields a coefficient of 0.71 and an R^2 of 0.23. The right panel of Figure B.3 compares average annual salaries across occupations. Using 22 two-digit OCCSOC occupations, a simple regression produces a coefficient of 1.03 and an R^2 of 0.84.

Finally, Figure B.4 compares average annual salaries across CZ–occupation pairs, where a simple regression yields a coefficient of 0.87 and an R^2 of 0.55. Taken together, the evidence indicates that Glassdoor wage data capture much of the variation in wages across cities and occupations.

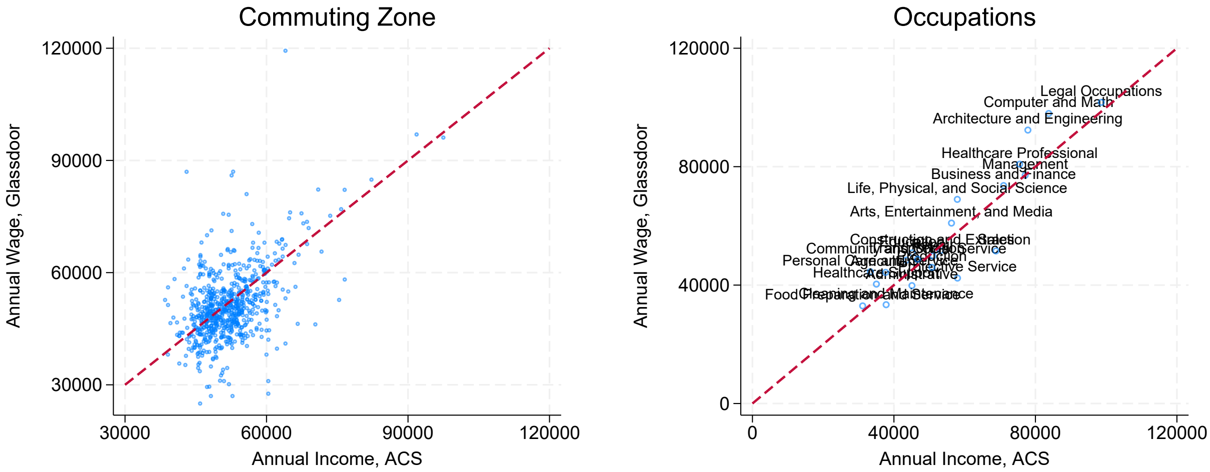

Figure B.3: Annual Wage by Commuting Zones (left) and Occupations (right): Glassdoor vs. ACS

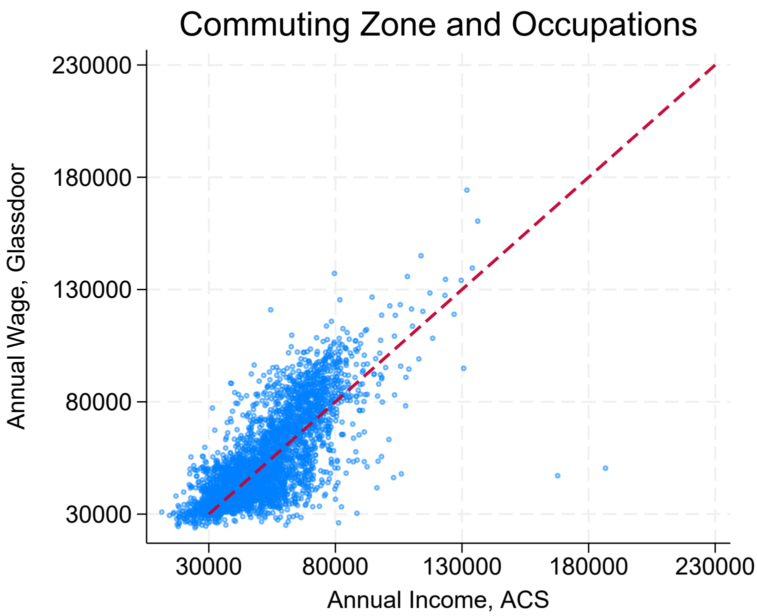

Figure B.4: Annual Wage by Commuting Zone-Occupation Pairs: Glassdoor vs. ACS

C Stylized Facts

Fact C1. Fresh graduates are highly mobile across cities, with the majority moving within the same state or to neighboring states.

Column (1) in Table C.1 reports the fraction of graduates who remain in the same CZ as their college. Overall, more than half of graduates leave their college city for employment, with 47.7% staying and working in the same CZ.

Despite the low retention rate within the city, column (2) shows that 70% of graduates are employed within the same state. On average, graduates travel 320 miles to secure a job, a distance similar to that between Boston and Philadelphia, or from Phoenix to San Diego. Compared to the vast distances spanned across the U.S. continent (nearly 3,000 miles from northeast to northwest, and over 3,300 miles from southeast to northwest), these figures indicate that, on average, graduates tend to work relatively close to their place of study.

Table C.1: Retention Rates by College Rankings

College	(1) Retention Rate Commuting Zones	(2) Retention Rate State	(3) Distance Traveled (Miles)	(4) Weekly Wages of Hosting City
All	0.477	0.703	320	1547
Top20	0.376	0.551	605	1733
Top21-200	0.431	0.654	380	1520
Top201-1000	0.475	0.704	279	1489
Top1001-2000	0.561	0.800	243	1623

Fact C2. Graduates from higher-ranked universities are more mobile, with the most notable group being graduates from the Top 20 universities.

Column (1) shows that within-CZ retention is 56.1% for universities outside the Top 1000 but drops sharply to 37.6% for Top 20. Column (2) shows that within-state retention is as high as 80% for universities outside the Top 1000, falling to 70% for Top 201–1000 universities, 65.4% for Top 21–200, and only 55.1% for Top 20. Column (3) shows that graduates from Top 20 universities travel, on average, 605 miles for jobs, compared to 380 miles for Top 21–200 universities, 279 miles for Top 201–1000, and 243 miles for universities outside the Top 1000.[46]

This pattern persists when we examine the distance traveled by movers.[47] Figure C.1 shows that, for graduates of lower-ranked universities, more than half relocate within 300 miles, and the vast majority move within 600 miles. In contrast, graduates of the top 20 universities are far more geographically mobile: 25.7% relocate more than 1,800 miles for a job.

The last column of Panel A reports the average weekly wage in the university-hosting cities.[48] Top 20 universities, on average, are situated in the highest-paying cities than other groups.

[46] These average distances traveled are estimated using both movers and stayers.

[47] Movers are defined as graduates who relocate to a different CZ for employment after graduation.

[48] We compute the average weekly wage using the ACS, restricting the sample to individuals who work full-time and hold a college degree.

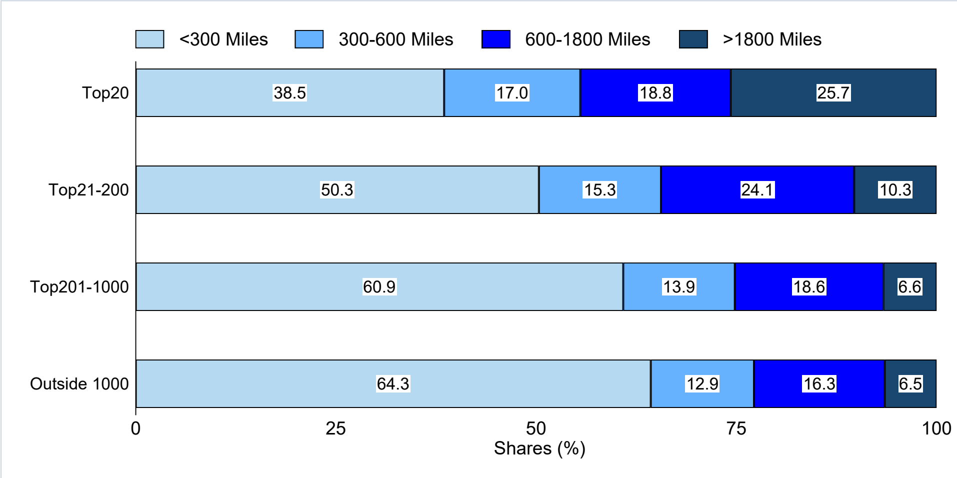

Figure C.1: Distribution of Distance Traveled by University Group, Among Movers

Notes: The plot displays the share of graduates who match to jobs in each distance category, conditional on movers (graduates who relocate to a different CZ for employment after graduation). County-to-county distances are obtained from the NBER County Distance Database, and CZ-to-CZ distances are computed as a population-weighted average.

D The Probability of Matching to High-Paid Jobs

Figure D.1 displays the probability of matching to top 5% jobs using blue circles and the effect of geography on this probability, $\text{ProbTop}_g^{\text{Geo}}$, using triangles. We display 30 universities: the fifteen with the highest geographic effect and the fifteen with the lowest.[49] There are systematic differences across universities in the probability of interest, reflecting a combination of factors, including university reputation, educational quality, student ability, and geography. Among the listed universities, the highest probabilities are observed at Harvard (34.6%), Carnegie Mellon (34.5%), Georgetown (32.8%), UC Berkeley (29.8%), and Columbia (28.9%). In contrast, the lowest values are found in institutions such as San Diego State (8.8%) and Fordham (9.0%).

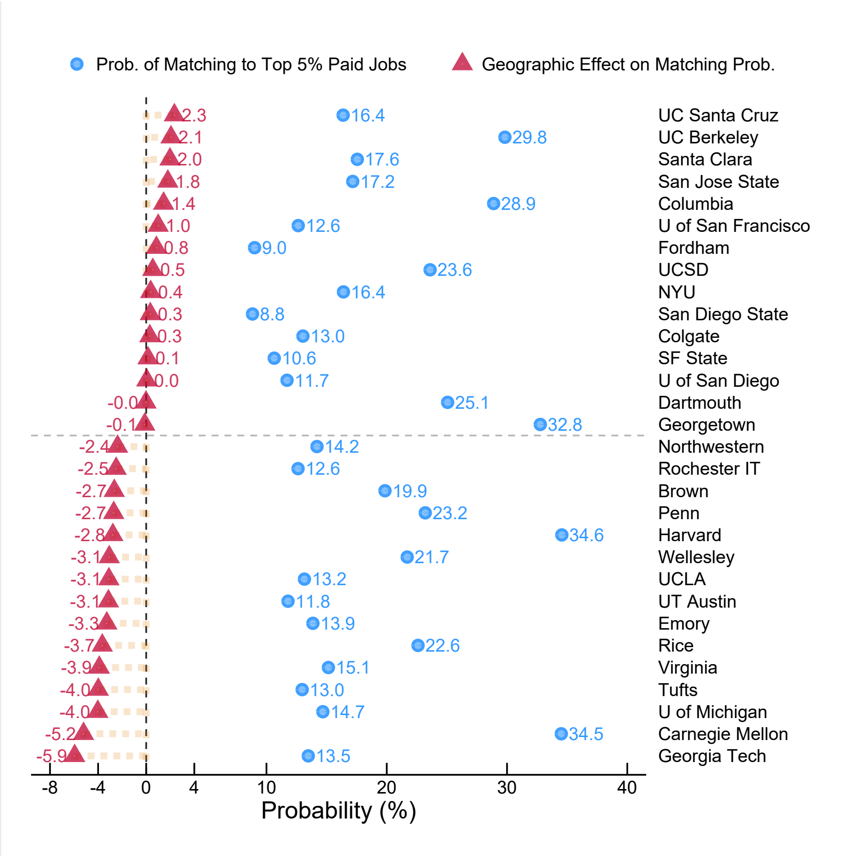

Figure D.1: Geographic Premium on the Probability of Matching to Top 5% of Paid Jobs

Notes: The figure shows the probability of matching to Top 5% of paid jobs. The x-axis represents the probability in percentage points. The blue circles plot the overall probability of such matches. The red triangle plots the value of $\text{ProbTop}_g^{\text{Geo}}$, capturing the effect of geography on this probability. We display 30 universities: the fifteen with the highest geographic premium and the fifteen with the lowest (most negative).

The red triangles show that geography substantially increases the probability of matching to high-paid jobs for universities located in the Bay Area and New York City. This result is unsurprising, given that the Bay Area and NYC together host nearly half of the top 5% highest-paid entry-level jobs according

[49] We select these 30 universities among those that have at least an 8% probability of matching to top 5% jobs.

to Glassdoor. Notably, geographic proximity increases the probability of matching by 2.3 percentage points (ppts) for UC Santa Cruz, 2.1 ppts for UC Berkeley, 2.0 ppts for Santa Clara University, 1.8 ppts for San Jose State, and 1.4 ppts for Columbia.

In contrast, geographic disadvantage reduces the probability most strongly for Georgia Tech (5.9 ppts), followed by Carnegie Mellon (5.2 ppts), and by 4.0 ppts for the University of Michigan and Tufts. Graduates from other elite universities—including Harvard, Northwestern, Brown, and Penn—also appear to face geographic disadvantages in matching to high-paid jobs. Between the two extremes, geographic differences account for up to an 8.2 ppt gap in the probability of obtaining a top-paid job. This value is based on the difference between UC Santa Cruz (+2.3 ppts) and Georgia Tech (–5.9 ppts). In a recent study, Chetty, Deming and Friedman (2023) show that attending an Ivy-Plus college triples the likelihood of working in a prestigious firm, relative to attending an average flagship public university. Complementing this literature, we find a sizable geographic premium associated with university location.

E Tables and Figures

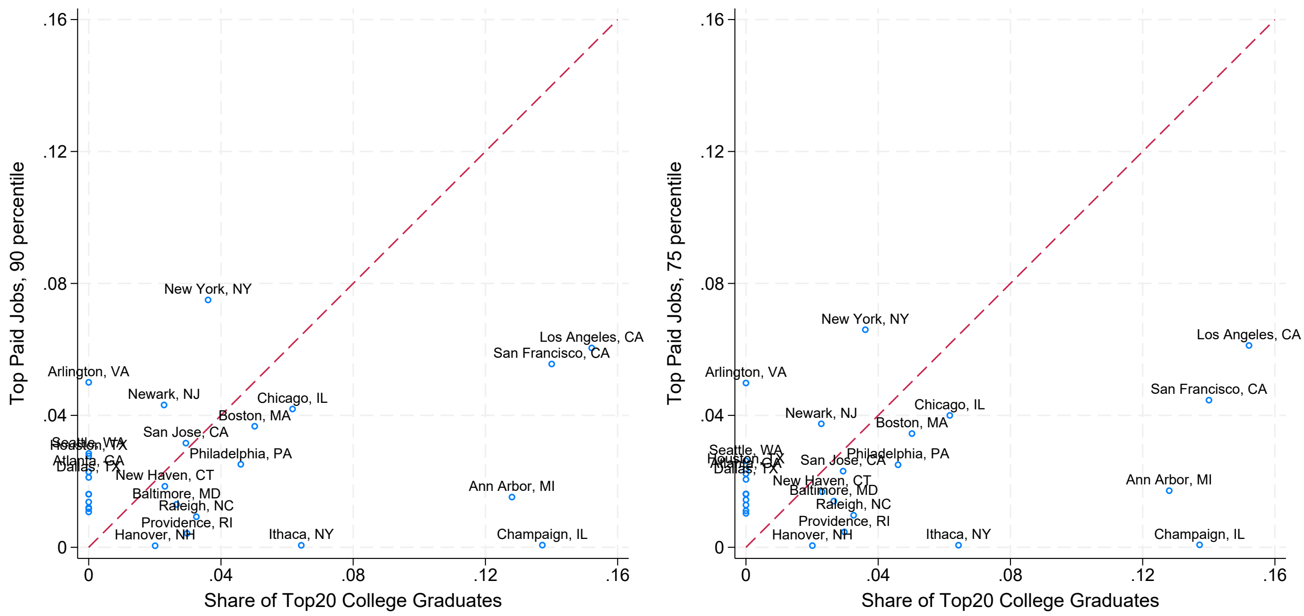

Figure E.1: Spatial Distribution of Top 20 university Graduates and Jobs: 90th Percentile (Left) and 75th Percentile (Right) as High-Paying Jobs

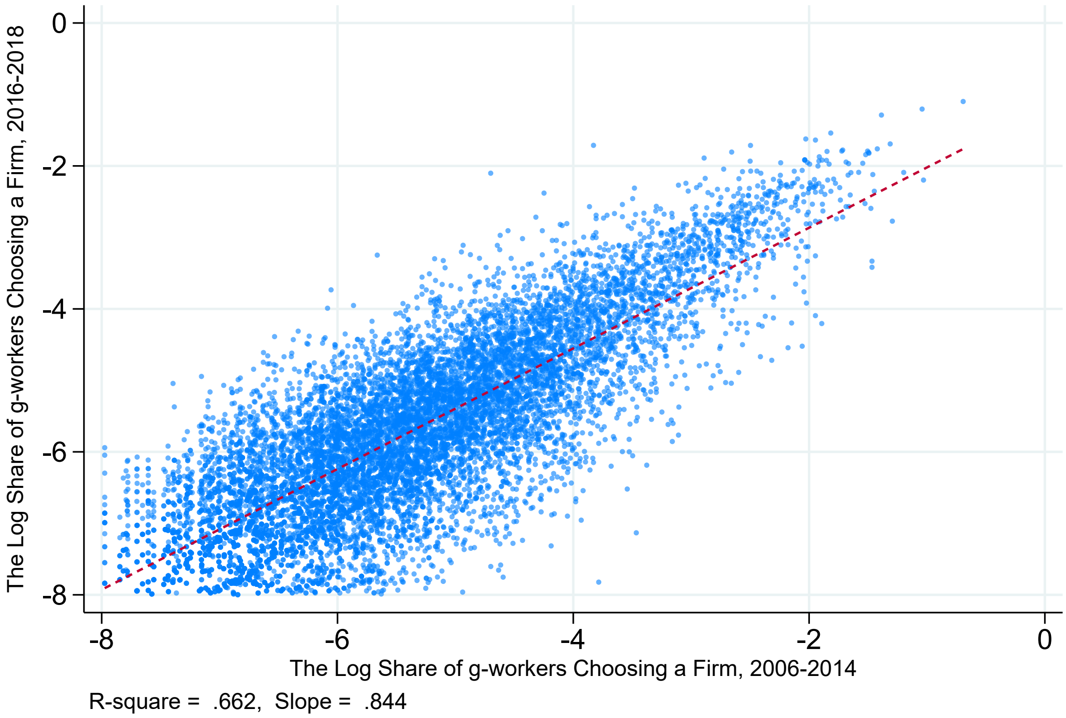

Figure E.2: The Log Share of g-worker Choosing a Firm, Recent Graduates (2016-2018) versus Former Graduates (2006-2014)

This figure plots the log share of graduates from a given college choosing a firm, for fresh graduates (2016-2018) against early cohorts who graduated before 2014. The dashed line plots the linear fit, which shows a slope coefficient of 0.844 (s.e. = 0.006) and an R^2 of 0.662 .

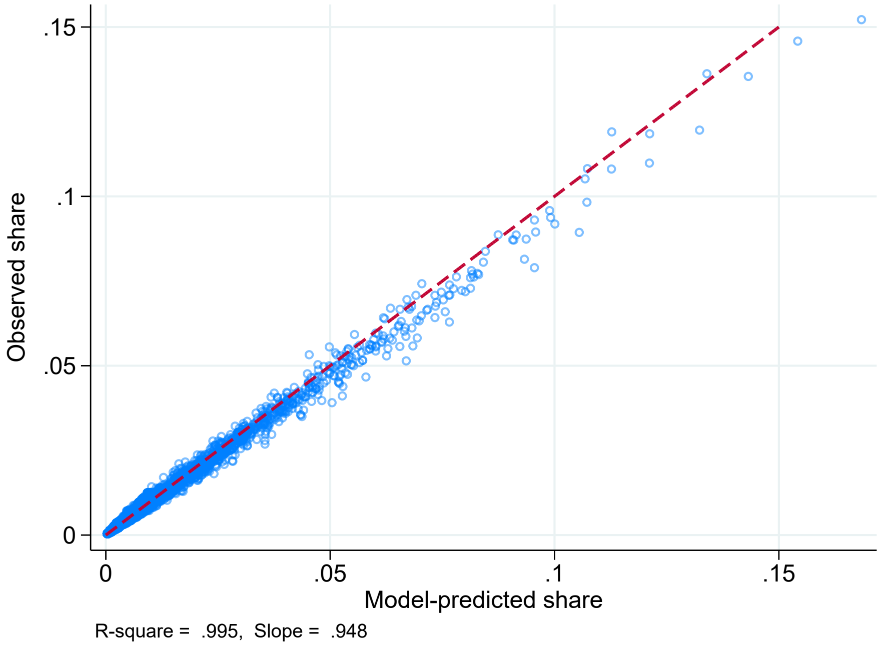

Figure E.3: The University-to-Job Allocation Shares, Data vs. Model-predicted Shares

This figure plots the share of graduates from each college choosing a given firm–occupation pair in the data (y-axis) against the corresponding model-predicted shares (x-axis). The dashed line shows the linear fit, with a slope coefficient of 0.948 and an R^2 of 0.995.

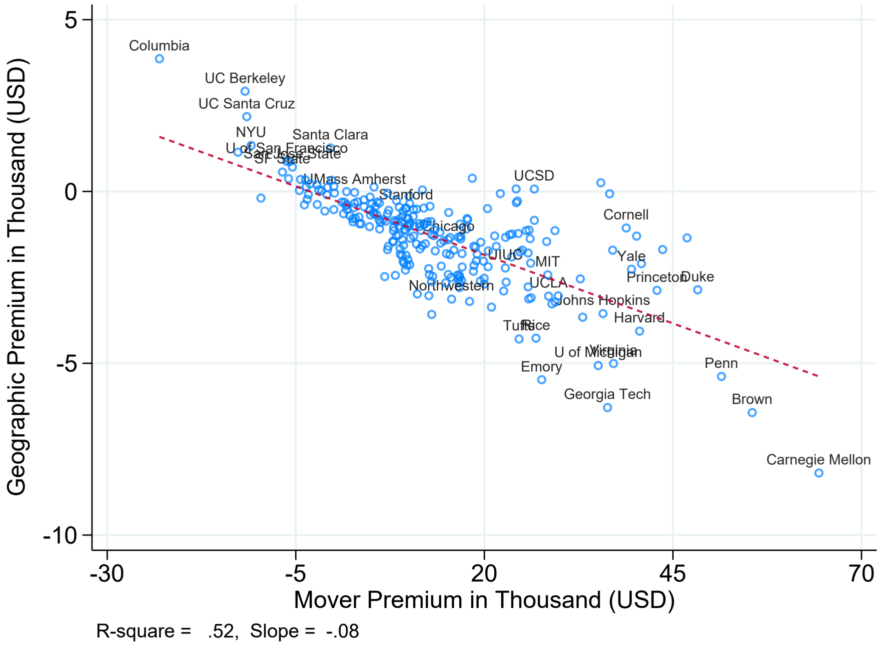

Figure E.4: Geographic Premium (y-axis) Against Mover Premium (x-axis)

This Figure plots the geographic premium (y-axis) against the mover premium (x-axis) across universities. We measure the mover premium as the difference in average wages between movers and stayers among fresh graduates (2016–2018), directly estimated from our sample. A negative mover premium implies that, on average, local stayers earn higher wages than those who migrate.

F Additional Table and Estimation Results

Table F.1: Estimates of Equation (16)

	(1)	(2)	(3)	(4)	(5)
Cognitive × Top20	0.112***	0.109***	0.086***	0.087***	0.087***
	(0.032)	(0.032)	(0.027)	(0.027)	(0.027)
Cognitive × Top21-200	0.029*	0.030*	0.038***	0.038***	0.039***
	(0.016)	(0.016)	(0.015)	(0.015)	(0.015)
Cognitive × Top201-1000	0.015	0.016	0.020*	0.021*	0.021*
	(0.013)	(0.013)	(0.012)	(0.012)	(0.012)
Social × Top20	0.082**	0.077**	0.056**	0.056**	0.056**
	(0.033)	(0.032)	(0.027)	(0.027)	(0.027)
Social × Top21-200	0.018	0.018	0.013	0.013	0.014
	(0.015)	(0.015)	(0.014)	(0.014)	(0.014)
Social × Top201-1000	-0.015	-0.016	-0.010	-0.009	-0.009
	(0.014)	(0.014)	(0.012)	(0.012)	(0.012)
Routine × Top20	-0.058*	-0.053	-0.030	-0.029	-0.029
	(0.033)	(0.033)	(0.027)	(0.027)	(0.027)
Routine × Top21-200	0.000	0.001	0.000	-0.000	-0.000
	(0.014)	(0.014)	(0.012)	(0.012)	(0.012)
Routine × Top201-1000	0.016	0.017	0.014	0.013	0.013
	(0.011)	(0.011)	(0.010)	(0.010)	(0.010)
Manual × Top20	0.024	0.023	0.034	0.034	0.035
	(0.028)	(0.028)	(0.025)	(0.025)	(0.025)
Manual × Top21-200	0.011	0.010	-0.003	-0.003	-0.004
	(0.014)	(0.014)	(0.014)	(0.013)	(0.013)
Manual × Top201-1000	-0.014	-0.015	-0.013	-0.013	-0.014
	(0.012)	(0.012)	(0.011)	(0.011)	(0.011)
Amenity × Top20		0.087**	0.146***	0.150***	0.150***
		(0.034)	(0.030)	(0.031)	(0.031)
Amenity × Top21-200		-0.005	0.031**	0.031**	0.031**
		(0.013)	(0.013)	(0.012)	(0.013)
Amenity × Top201-1000		-0.024*	0.009	0.009	0.009
		(0.013)	(0.010)	(0.010)	(0.010)
Ldist			-0.076***	-0.050***	-0.050***
			(0.011)	(0.018)	(0.018)
Ldist × Top20			0.087***	0.073***	0.073***
			(0.024)	(0.027)	(0.027)
Ldist × Top21-200			0.027	0.024	0.024
			(0.019)	(0.020)	(0.020)
Ldist × Top201-1000			-0.012	-0.011	-0.011
			(0.013)	(0.013)	(0.013)
CZ			0.439***	0.372***	0.373***
			(0.029)	(0.088)	(0.088)
CZ × Top20			0.192*	0.219**	0.219**
			(0.099)	(0.105)	(0.105)
CZ × Top21-200			0.141***	0.168***	0.168***
			(0.053)	(0.056)	(0.056)
CZ × Top201-1000			0.093**	0.102**	0.102**
			(0.042)	(0.041)	(0.041)

Table F.1: Estimates of Equation (16) (Continued)

	(1)	(2)	(3)	(4)	(5)
State			0.055**	-0.020	-0.020
			(0.025)	(0.067)	(0.067)
State × Top20			0.289***	0.311***	0.311***
			(0.102)	(0.107)	(0.106)
State × Top21-200			0.178***	0.183***	0.183***
			(0.043)	(0.044)	(0.044)
State × Top201-1000			0.079**	0.083**	0.084**
			(0.034)	(0.034)	(0.034)
Ldist × In-State				-0.038*	-0.038*
				(0.021)	(0.021)
CZ × In-State				0.075	0.074
				(0.103)	(0.103)
State × In-State				0.108	0.108
				(0.082)	(0.082)
Cognitive × Public					-0.006
					(0.014)
Social × Public					-0.003
					(0.015)
Routine × Public					0.000
					(0.013)
Manual × Public					0.014
					(0.013)
Observations	84,100	83,549	83,549	83,549	83,549
Adjusted R^2	0.76	0.75	0.81	0.81	0.81

Notes: Columns (1)-(3) report the estimated coefficients for equation (16). Column (1) includes the interaction of ranking dummies and BGT tasks; Column (2) adds amenities, and Column (3) adds geographic variables. Column (4) adds interactions of the in-state student enrollment share. Column (5) adds interactions between public-university status and task measures. All models are OLS estimates using the full sample. The full sample covers 22 occupations, 25246 firms, and 264 universities. Standard errors are clustered by university and reported in parentheses.

Table F.2: Estimates of Equation (16) with Disaggregated Groups

	(1)	(2)	(3)	(4)
Cognitive × Top20	0.200***	0.190***	0.168***	0.170***
	(0.043)	(0.042)	(0.037)	(0.037)
Cognitive × Top21-100	0.078***	0.078***	0.085***	0.085***
	(0.023)	(0.024)	(0.022)	(0.022)
Cognitive × Top101-200	0.076**	0.083**	0.085***	0.084***
	(0.036)	(0.037)	(0.032)	(0.032)
Cognitive × Top201-500	0.047**	0.049**	0.059***	0.059***
	(0.020)	(0.020)	(0.018)	(0.018)
Cognitive × Top501-1000	0.018	0.021	0.017	0.017
	(0.017)	(0.018)	(0.015)	(0.015)
Social × Top20	0.052	0.043	0.035	0.034
	(0.037)	(0.037)	(0.031)	(0.031)
Social × Top21-100	-0.025	-0.027	-0.017	-0.017
	(0.023)	(0.022)	(0.020)	(0.020)
Social × Top101-200	0.071**	0.071**	0.049**	0.050**
	(0.031)	(0.031)	(0.025)	(0.025)
Social × Top201-500	-0.037	-0.037	-0.025	-0.024
	(0.024)	(0.024)	(0.020)	(0.020)
Social × Top501-1000	-0.040**	-0.042**	-0.017	-0.017
	(0.019)	(0.019)	(0.016)	(0.016)
Routine × Top20	-0.058	-0.058	-0.021	-0.020
	(0.046)	(0.046)	(0.036)	(0.036)
Routine × Top21-100	0.035*	0.035*	0.029*	0.029*
	(0.018)	(0.018)	(0.016)	(0.016)
Routine × Top101-200	-0.047	-0.041	-0.012	-0.012
	(0.034)	(0.034)	(0.026)	(0.026)
Routine × Top201-500	0.013	0.013	0.014	0.014
	(0.018)	(0.018)	(0.016)	(0.016)
Routine × Top501-1000	0.011	0.016	0.011	0.011
	(0.015)	(0.015)	(0.013)	(0.013)
Manual × Top20	-0.010	-0.011	-0.006	-0.007
	(0.032)	(0.032)	(0.028)	(0.028)
Manual × Top21-100	-0.022	-0.022	-0.036*	-0.036*
	(0.021)	(0.021)	(0.019)	(0.019)
Manual × Top101-200	-0.040*	-0.037	-0.058***	-0.057***
	(0.023)	(0.023)	(0.021)	(0.021)
Manual × Top201-500	-0.060***	-0.059***	-0.060***	-0.060***
	(0.019)	(0.019)	(0.017)	(0.017)
Manual × Top501-1000	-0.029*	-0.028*	-0.031**	-0.031**
	(0.017)	(0.017)	(0.015)	(0.015)
Amenity × Top20		0.349***	0.520***	0.539***
		(0.127)	(0.114)	(0.115)
Amenity × Top21-100		0.105*	0.136**	0.137**
		(0.060)	(0.056)	(0.056)
Amenity × Top101-200		-0.350***	0.059	0.056
		(0.080)	(0.056)	(0.056)
Amenity × Top201-500		-0.015	0.048	0.057
		(0.066)	(0.051)	(0.051)
Amenity × Top501-1000		-0.171***	-0.022	-0.034
		(0.066)	(0.048)	(0.047)

Table F.2: Estimates of Equation (16) with Disaggregated Groups (Continued)

	(1)	(2)	(3)	(4)
Ldist			-0.077***	-0.070***
			(0.011)	(0.021)
Ldist × Top20			0.072***	0.070**
			(0.024)	(0.028)
Ldist × Top21-100			0.055**	0.056**
			(0.025)	(0.025)
Ldist × Top101-200			-0.037*	-0.035
			(0.023)	(0.023)
Ldist × Top201-500			-0.023	-0.019
			(0.015)	(0.016)
Ldist × Top501-1000			-0.007	-0.005
			(0.015)	(0.015)
CZ			0.439***	0.338***
			(0.030)	(0.089)
CZ × Top20			0.188*	0.237**
			(0.101)	(0.107)
CZ × Top21-100			0.195***	0.231***
			(0.065)	(0.068)
CZ × Top101-200			0.065	0.102
			(0.090)	(0.092)
CZ × Top201-500			0.132**	0.148***
			(0.057)	(0.057)
CZ × Top501-1000			0.050	0.058
			(0.054)	(0.052)
State			0.074***	-0.036
			(0.026)	(0.068)
State × Top20			0.210**	0.253**
			(0.097)	(0.099)
State × Top21-100			0.174***	0.183***
			(0.050)	(0.051)
State × Top101-200			0.128**	0.141**
			(0.065)	(0.066)
State × Top201-500			0.029	0.036
			(0.045)	(0.044)
State × Top501-1000			0.074*	0.082**
			(0.038)	(0.038)
Ldist × In-State				-0.013
				(0.024)
CZ × In-State				0.119
				(0.106)
State × In-State				0.159*
				(0.084)
Observations	85,824	85,277	84,175	84,175
Adjusted R^2	0.67	0.67	0.74	0.74

Notes: Columns (1)-(3) report the estimated coefficients for equation (16). Column (1) includes the interaction of ranking dummies and BGT tasks; Column (2) adds amenities, and Column (3) adds geographic variables. Column (4) adds interactions of the in-state student enrollment share. All models are OLS estimates. Standard errors are clustered by university and reported in parentheses.

Table F.3: Estimates of Equation (16) using U.S. News Rankings

	(1)	(2)	(3)	(4)
Cognitive × Top20	0.181***	0.165***	0.121***	0.122***
	(0.052)	(0.050)	(0.042)	(0.042)
Cognitive × Top200	0.065***	0.066***	0.063***	0.064***
	(0.020)	(0.020)	(0.017)	(0.017)
Social × Top20	0.057	0.053	0.043	0.042
	(0.046)	(0.046)	(0.036)	(0.036)
Social × Top200	-0.000	-0.000	-0.004	-0.005
	(0.020)	(0.020)	(0.016)	(0.016)
Routine × Top20	-0.059	-0.066	-0.030	-0.030
	(0.057)	(0.056)	(0.043)	(0.042)
Routine × Top200	-0.007	-0.006	-0.004	-0.004
	(0.019)	(0.019)	(0.015)	(0.015)
Manual × Top20	-0.009	-0.009	0.015	0.015
	(0.038)	(0.038)	(0.032)	(0.032)
Manual × Top200	0.018	0.019	0.013	0.013
	(0.017)	(0.017)	(0.015)	(0.015)
Amenity × Top20		0.088**	0.144***	0.151***
		(0.037)	(0.035)	(0.035)
Amenity × Top200		-0.007	0.035***	0.036***
		(0.014)	(0.013)	(0.013)
Ldist			-0.106***	-0.074***
			(0.010)	(0.020)
Ldist× Top20			0.074***	0.059**
			(0.021)	(0.023)
Ldist× Top200			0.070***	0.066***
			(0.021)	(0.021)
CZ			0.516***	0.423***
			(0.036)	(0.122)
CZ× Top20			0.143	0.159
			(0.132)	(0.140)
CZ× Top200			0.097	0.119*
			(0.062)	(0.067)
State			0.113***	0.098
			(0.025)	(0.083)
State× Top20			0.117	0.126
			(0.150)	(0.150)
State× Top200			0.167***	0.168***
			(0.044)	(0.045)
Ldist × In-State				-0.050*
				(0.028)
CZ × In-State				0.120
				(0.139)
State × In-State				0.027
				(0.108)
Observations	57,788	57,479	57,479	57,105
Adjusted R^2	0.67	0.67	0.75	0.75

Notes: Columns (1)-(3) report the estimated coefficients for equation (16). Column (1) includes the interaction of ranking dummies and BGT tasks; Column (2) adds amenities, and Column (3) adds geographic variables. Column (4) adds interactions of the in-state student enrollment share. All models are OLS estimates using the full sample. Standard errors are clustered by university and reported in parentheses.

Table F.4: Admission Criteria by College Tiers

College Tier	Minimum SAT Scores	Minimum ACT Scores	Rankings in High school
Most selective	655	29	Top 10-20%
Highly selective	620	27	Top 20-35%
Very selective	573	24	Top 35-50%
Selective	500	21	Top 50-65%
Less selective	below 500	below 21	Top 65%
Nonselective	None	None	None

Notes: This table shows the SAT/ACT scores and the ranking in high school transcripts or class rank that are typically required by college admission. The information is based on Barron's Profiles of American universities (Barron's Educational Series, 2017).