BGT Occupational Mobility Data: Description and FAQ

Gregor Schubert, Anna Stansbury, and Bledi Taska

May 2020

For more details on this data set, its construction, and summary statistics, please see the paper for which it was constructed: Schubert, Stansbury, and Taska (2020) "Monopsony and Outside Options", available at SSRN as paper 3599454. **If you use this data, please cite Schubert, Stansbury, and Taska (2020) "Monopsony and Outside Options".**

1 Summary

This data set contains data on occupational mobility, calculated from 16 million resumes of U.S. workers, obtained and parsed by Burning Glass Technologies (a labor market analytics company). The data is *national* and comes from observations over the period 2002–2016, with the majority of observations in the later years. Individual occupations are defined at a granular level (6-digit Standard Occupational Classificiation or SOC code). There are 840 SOC 6-digit occupations. We have data on a large subset of these occupations. The data set contains 6 variables:

- $soc1$ and $soc1\_name$: The origin occupation for the transitions (SOC 6 digit code and name).
- $soc2$ and $soc2\_name$: The destination occupation for the transitions (SOC 6 digit code and name).

- $total\_obs$: Total number of effective observations of occupation switches from origin occupation *soc1* to any other destination occupation *soc2*, observed in the BGT resume data. Note that this may not be an integer as it is reweighted by age to be reflective of the age distribution in the occupation.

- $transition\_share$: Share of occupation switches from origin occupation *soc1* which move to destination occupation *soc2*.

2 Constructing occupational transitions from BGT Resume Data

Our data on occupational transitions is constructed from a new proprietary data set of 16 million unique resumes with more than 80 million job observations over 2002–2018, provided by labor market analytics company Burning Glass Technologies ("BGT"). Resumes were sourced from a variety of BGT partners, including recruitment and staffing agencies, workforce agencies, and job boards. Since we have all data that people have listed on their resumes, we are able to observe individual workers' job histories and education up until the point where they submit their resume, effectively making it a longitudinal data set.

We apply a number of different filters to the Burning Glass resume data before calculating our occupational mobility matrices: First, we retain only resumes that are from the U.S. Next, we keep only jobs on these resumes that last for longer than 6 months to ensure that we are only capturing actual jobs rather than short-term internships, workshops etc. We also apply a number of filters to minimize the potential for mis-parsed jobs, by eliminating all jobs that lasted longer than 70 years. Moreover, we impute the ages of workers based on their first job start date and education and limit our sample to resumes submitted by workers between the ages of 16 and 100. As we are interested in occupational transitions during the last two decades, we then restrict the data set to jobs held after 2001. The final number of resumes that contain at least two years of job data under these restrictions is 15.8 million. The main job information retained for each resume are the occupation and duration of each job held.

We would ideally use the BGT data to estimate annual transition probabilities between full-time jobs in particular occupations. Unfortunately, resumes often do not list the months in which jobs started and ended, and do not always indicate if jobs were part-time. We therefore approximate the share of workers moving from occupation o to occupation p with the share of all workers observed in occupation o at any point in year t who are observed in occupation p at any point in year $t+1$. We describe this process further below.

For each of these resumes, we start by extracting separate observations for each occupation that the worker was observed in, in each year. These observations are then matched to all other occupation-year observations on the same resume. We retain all matches that are in sequential years - either in the same occupation or in different occupations. For instance, if a worker was a Purchasing Manager in the period 2003-2005, and a Compliance Officer in 2005-2007, we would record 1-year horizon sequential occupation patterns of the form shown in Table **??**.

Table 1: Illustrative example of sequential job holding data.

Year:	2004	2005	2006
Current Occ.	*1-Year Horizon Occ.*		
Purchasing Mgr. (11-3061)	11-3061 13-1040		
Compliance Off. (13-1040)		13-1040	13-1040

In our data we have 80.2 million job observations. This results in 178.5 million observations of year-to-year occupation coincidences (including year-to-year pairs where workers are observed in the same occupation in both years). Below, we describe the characteristics of this data and how it compares to other data sets - with all statistics referring to this final set of filtered sequence observations, or the 15.8 million resumes, unless otherwise noted.

We use these occupation coincidence pairs to construct our measures of occupational mobility as follows. For each pair of (different) occupations o to p, we count the total number of year-to-year occupation coincidence pairs where the worker is observed in occupation o at any point in year t and is observed in occupation p at

any point in year $t + 1$. We then divide this by the total across all these observations: i.e. the number of occupation-to-occupation coincidence pairs starting in occupation o in year t and moving to *any new occupation* p in the following year $t + 1$.

Note that this measure will also capture mobility between occupations in the form of working in two different occupations *at the same time*, as well as mobility that consists of taking a job in a new occupation while continuing to work in an old job in the origin occupation. Implicitly, we are assuming that taking up a secondary job in an occupation indicates its viability as an outside option to the same degree as moving primary occupations. (Under this assumption, our measure is more appropriate than one that focuses only on transitions that involve abandoning a previous job or occupation entirely.)

Since our data is not fully representative on age within occupations, we compute these occupation transition shares separately for different age categories (24 and under, 25 to 34, 35 to 44, 45 to 54, and 55 and over). We then aggregate them, reweighting by the average proportion of employment in each of these age categories in that occupation in the U.S. labor force over 2012–2017 (from the BLS Occupational Employment Statistics). Our aggregate occupational mobility matrix has therefore been reweighted to correspond to the empirical within-occupation age distribution in the labor force, eliminating any potential bias from the skewed age distribution of our sample.

The BGT resume data set is largely representative of the U.S. labor force in its distribution by gender and location. However, it over-represents younger workers and white-collar occupations. Since we are estimating occupational transition probabilities within each occupation, the over-representation by occupation is not a substantial concern as long as we still have sufficient data for most occupations to have some degree of representativeness *within* each occupation.

Our measure of occupational transitions $transitionshare_{o \to p}$ is the probability of a worker moving from occupation o to occupation p conditional on leaving her

occupation, defined as:

$$\begin{aligned} transitionshare_{o \to p} &= Pr(\text{move from occ } o \text{ to occ } p | \text{leave occ } o) \\ &= \frac{\text{Share from occ } o \text{ moving into occ } p}{\text{Share from occ } o \text{ moving into any new occ}} \\ &= \frac{\text{\# in occ } o \text{ in year } t \text{ observed in occ } p \text{ in year } t+1}{\text{\# in occ } o \text{ in year } t \text{ observed in any new occ in year } t+1} \end{aligned} \tag{1}$$

We estimate these occupation transition probabilities at the national level for a large proportion of the possible pairs of SOC 6-digit occupations. We exclude the occupations for which we have fewer than 500 observations of occupation-to-occupation transition pairs in the BGT data (roughly the bottom 10% of occupations). We also exclude transitions to and from military occupations, as we have reason to believe our data will not be representative of these. This results in 734 origin SOC 6-digit occupations in our data and 278,195 non-empty occupation-to-occupation transition cells out of a total 705,600 possible transition cells (840 x 840). We average the observed annual occupation-to-occupation transitions over all observations over starting years 2002–2015,[1] to capture as much as possible the underlying degree of occupational similarity rather than transitory fluctuations in mobility. We do not use data from 2017 and 2018 in our analysis, because these data capture workers who submitted their resumes to jobs in 2017 or 2018, and we suspect that the sample of workers who switch occupations and then apply for a new job in the same year may not be representative of the majority of workers.

3 BGT Resume Data: description

Below, we describe the characteristics of the Burning Glass Technologies resume data and how it compares to other data sets. All statistics referring to the final set of 15.8 million filtered resumes, or 178.5 million observations of year-to-year occupa-

[1] Where 2015 refers to the *starting* year, i.e. takes individuals in occupation o in 2015 and observes their occupation in 2016. We exclude observations from starting years 2016 or 2017 to avoid bias from the resume collection process: if we observe someone applying for a job in 2017 who has also changed job in 2016, they are not likely to be representative of the average worker.

tion coincidences ('observations') from these resumes, unless otherwise noted.

Job number and duration: The median number of jobs on a resume is 4, and more than 95% of the resumes list 10 or fewer jobs (note that a change of job under our definition could include a change of job title or occupation under the same employer). The median length job was 2 years, with the 25th percentile just under 1 year and the 75th percentile 4 years. The median span of years we observe on a resume (from date started first job to date ended last job) is 12 years. Table **??** shows more information on the distribution of job incidences and job durations on our resumes.

Table 2: Distribution of number of jobs on resume and duration of jobs in BGT data set.

Percentile	10th	25th	50th	75th	90th
# Jobs on resume	2	3	4	6	9
Job duration (months)	4	12	24	48	98

Gender: BGT imputes gender to the resumes using a probabilistic algorithm based on the names of those submitting the resumes. Of our observations, 88% are on resumes where BGT was able to impute a gender probabilistically. According to this imputation, precisely 50% of our observations are imputed to come from males and 50% are more likely to be female. This suggests that relative to the employed labor force, women are very slightly over-represented in our data. According to the BLS, 46.9% of employed people were women in 2018.

Education: 141.3 million of our observations are on resumes containing some information about education. The breakdown of education in our data for these data points is as follows: the highest educational level is postgraduate for 25%, bachelor's degree for 48%, some college for 19%, high school for 8% and below high school for less than 1%. This substantially overrepresents bachelor's degree-holders and post-college qualifications: only 40% of the labor force in 2017 had a bachelor's degree or higher according to the BLS, compared to 73% in this sample (full comparisons to the labor force are shown in Figure **??**). It is to be expected that the sample of the resumes which *provide* educational information are biased towards

those with tertiary qualifications, because it is uncommon to put high school on a resume. Imputing high school only education for all resumes which are missing educational information substantially reduces the overrepresentation of those with a BA and higher: by this metric, only 58% of the BGT sample have a bachelor's degree or higher. This remains an overrepresentation, but this is to be expected: a sample drawn from online resume submissions is likely to draw a more highly-educated population than the national labor force average both because many jobs requiring little formal education also do not require online applications, and because we expect online applications to be used more heavily by younger workers, who on average have more formal education. As long as we have enough data to compute mobility patterns for each occupation, and workers of different education levels *within* occupations do not have substantially different mobility patterns, this should therefore not be a reason for concern.

Age: We impute individuals' birth year from their educational information and from the date they started their first job which was longer than 6 months (to exclude internships and temporary jobs). Specifically, we calculate the imputed birth year as the year when a worker started their first job, minus the number of years the worker's maximum educational qualification requires, minus 6 years. High school is assumed to require 12 years, BA 16 years, etc. For those who do not list any educational qualification on their resume, we impute that they have high school only, i.e. 12 years of education. Since we effectively observe these individuals longitudinally - over the entire period covered in their resume - we impute their age for each year covered in their resume.

As a representativeness check, we compared the imputed age of the people corresponding to our 2002-2018 sample of sequential job observations in the BGT sample to the age distribution of the labor force in 2018, as computed by the BLS. The BGT data of job observations substantially overrepresents workers between 25 and 40 and underrepresents the other groups, particularly workers over 55. 55% of observations in the BGT sample would have been for workers 25-40 in 2017, compared to 33% of the US labor force - see Figure **??** for the full distribution. One would expect a sample drawn from online resume submissions to overweight younger workers for three reasons: (1) because younger workers may be more fa-

miliar with and likely to use online application systems, (2) because older workers are less likely to switch jobs than younger workers, and (3) because the method for job search for more experienced (older) workers is more likely to be through direct recruitment or networks rather than online applications. Moreover, by the nature of a longitudinal work history sample, young observations will be overweighted, as older workers will include work experiences when they are young on their resumes, whereas younger workers, of course, will never be able to include work experiences when they are old on their current resumes. Therefore, even if the distribution of resumes was not skewed in its age distribution, the sample of job observations would still skew younger.

As noted above, we directly address this issue by computing occupational mobility only after reweighting observations to adjust the relative prevalence of different ages in our sample relative to the labor force. For instance, this means that we overweight our observations for 45-49 year olds, as this age category is underrepresented in our sample relative to the labor force.

Occupation: The BGT automatic resume parser imputes the 6-digit SOC occupation for each job in the dataset, based on the job title. Of 178.5 million useable observations in the data set, 169.6 million could be coded into non-military 6-digit SOC occupations by the BGT parser. 833 of the 840 6-digit SOC occupations are present, some with few observations and some with very many. Ranking occupations by the number of observations,[2] the 10th percentile is 1,226 observations, 25th percentile is 4,173, the median is 20,526, 75th percentile is 117,538, and the 90th percentile is 495,699. We observe 216 occupations with more than 100,000 observations, 83 occupations with more than 500,000 observations, and 19 occupations with more than 2 million observations.[3]

[2]As defined above, for our purposes, an observation is a person-occupation-year observation for which we also observe another occupation in the following year: i.e. the start of a year-to-year occupation coincidence sequence.

[3]The occupations with more than 2 million observations are: General and Operations Managers; Sales Managers; Managers, All Other; Human Resources Specialists; Management Analysts; Software Developers, Applications; Computer User Support Specialists; Computer Occupations, All Other; First-Line Supervisors of Retail Sales Workers; Retail Salespersons; Sales Representatives, Wholesale and Manufacturing, Except Technical and Scientific Products; First-Line Supervisors of Office and Administrative Support Workers; Customer Service Representatives; Secretaries and Administrative Assistants, Except Legal, Medical, and Executive; Office Clerks, General; Heavy and

Figure **??** compares the prevalence of occupations at the 2-digit SOC level in our BGT data to the share of employment in that occupation group in the labor force according to the BLS in 2017. As the figure shows, at a 2-digit SOC level, management occupations, business and finance, and computer-related occupations are substantially overweight in the BGT data relative to the labor force overall, while manual occupations, healthcare and education are substantially underrepresented. However, this does not bias our results, as we compute mobility at the occupation-level.

Location: Since not all workers list the location where they work at their current job, we assign workers a location based on the address they list at the top of their resume. 115.4 million of our observations come from resumes that list an address in the 50 U.S. states or District of Columbia. Comparing the proportion of our data from different U.S. states to the proportion of workers in different U.S. states in the BLS OES data, we find that our data is broadly representative by geography. As shown in figure **??**, New Jersey, Maryland and Delaware, for instance, are 1.5-2x as prevalent in our data as they are in the overall U.S. labor force (probably partly because our identification of location is based on residence and the BLS OES data is based on workplace), while Nebraska, Montana, South Dakota, Alaska, Idaho and Wyoming are less than half as prevalent in our data as they are in the overall U.S. labor force. However, the figure also suggests that the broad patterns of the demographic distribution of populations across the U.S. is reflected in our sample. Aggregating the state data to the Census region level, the Northeast, Midwest, South, and West regions represent 24%, 22%, 38%, and 16% of our BGT sample, while they constitute 18%, 22%, 37%, and 24% of the BLS labor force. This shows that our sample is very close to representative for the Midwest and South regions, and somewhat overweights the Northeast, while underweighting workers from the West region.

Tractor-Trailer Truck Drivers; Financial Managers; Food Service Managers; Medical and Health Services Managers.

4 BGT Resume Data: caveats/concerns

The BGT dataset does, however, have other features which should be noted as caveats to the analysis.

1/ Sample selection: There are three areas of concern over sample selection: first, our data is likely to over-sample people who are more mobile between jobs, as the data is collected only when people apply for jobs; second, our data is likely to over-sample the types of people who are likely to apply for jobs online rather than through other means; and third, our data is likely to over-sample the types of people who apply for the types of jobs which are listed through online applications.

2/ Individuals choose what to put on their resume: We only observe whatever individuals have chosen to put on their resume. To the extent that people try to present the best possible picture of their education and employment history, and even sometimes lie, we may not observe certain jobs or education histories, and we may be more likely to observe "good" jobs and education histories than "bad" ones. The implication of this concern for our measure of job opportunities depends on the exact nature of this distortion. If workers generally inflate the level of occupation that they worked at, this would not necessarily distort our estimates of job transitions systematically, unless transition probabilities across occupations vary systematically with the social status / level of otherwise similar jobs. At the same time, if workers choose to highlight the consistency of their experiences by describing their jobs as more similar than they truly were, we may underestimate the ability of workers to transition across occupations. Conversely, if workers exaggerate the breadth of their experience, the occupational range of transitions would be overestimated. In any case, this issue is only likely to be significant, if these types of distortions exist for many observed workers, do not cancel out, and differ systematically between workers in different occupations.

We are only aware of a very limited number of studies directly trying to estimate the incidence of misrepresentations on resumes. For instance, Sloane (1991) surveys HR executives in banking and finds that 51 responding executives are jointly aware of a total of 17 incidences of meaningfully falsified job titles, which, given the presumably large number of resumes and contained job listings that would have

been processed under these executives seems small. All but one of the respondents estimated the incidence of falsification of *any* part of the resume to be below 20%, with most opting for lower estimates. Note that this study was done before online search made verification of basic resume information much faster and more affordable. footnoteSloane, Arthur A, "Countering resume fraud within and beyond banking: no excuse for not doing more," Labor Law Journal, 1991, 42 (5), 303. More recently, Nosnik et al (2010) found that 7% of the publications listed by a sample of urology residency applicants on their resumes could not be verified.[4]

While such low rates of misrepresentation seem unlikely to introduce systematic bias into our data, it is also important to keep in mind that we are trying to estimate the *plausibility* in a bargaining setting of other jobs constituting relevant outside options. If the skills of a job that they haven't actually held are plausibly consistent with *other* jobs on their resume in the eyes of jobseekers - and ultimately of employers - then this still constitutes evidence that these jobs are perceived as pertaining to the same labor market.

3/ Parsing error: Given the size of the dataset, BGT relies on an algorithmic parser to extract data on job titles, firms, occupations, education and time periods in different jobs and in education. Since there are not always standard procedures for listing job titles, education, dates etc. on resumes, some parsing error is likely to exist in the data. For example, the database states that 25,000 resumes list the end date of the most recent job as 1900.

4/ Possible duplicates: The resume data is collected from online job applications. If a worker over the course of her career has submitted multiple online job applications, it is possible that her resume appears twice in the raw database. BGT deduplicates the resume data based on matching name and address on the resume, but it is possible that there are people who have changed address between job applications. In these cases, we may observe the career history of the same person more than once in the data. Preliminary checks suggest that this is unlikely to be a major issue.

[4]Nosnik, Israel P, Patricia Friedmann, Harris M Nagler, and Caner Z Dinlenc, "Resume fraud: unverifiable publications of urology training program applicants," The Journal of urology, 2010, 183 (4), 1520–1523.

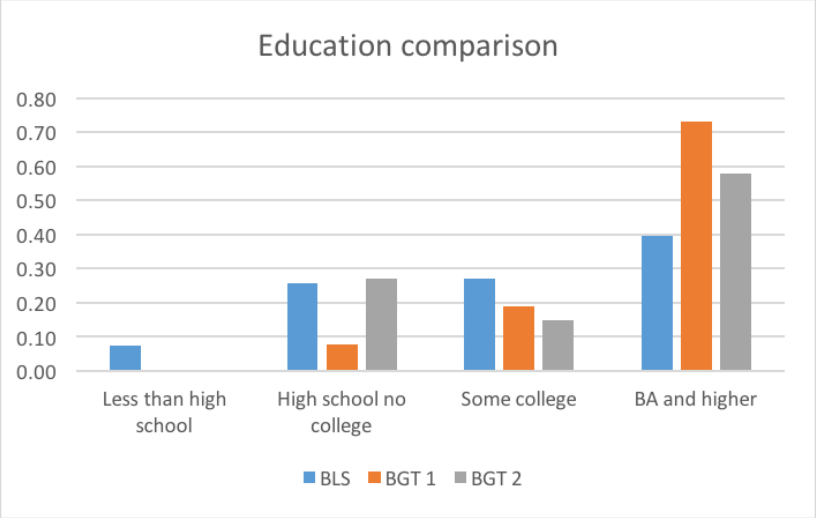
5 SOC occupational codes

Our occupational mobility data uses 6-digit occupational codes following the Standard Occupational Classification 2010 (SOC 2010) system. There are 840 narrowly-defined 6-digit SOC occupations ("detailed occupations"). These can be aggregated using the SOC hierarchy into bigger groups: 461 "broad groups", 97 "minor groups", and 23 "major groups". The Bureau of Labor Statistics has more detail on the SOC occupational codes on their website at www.bls.gov/soc.

6 Appendix: Figures

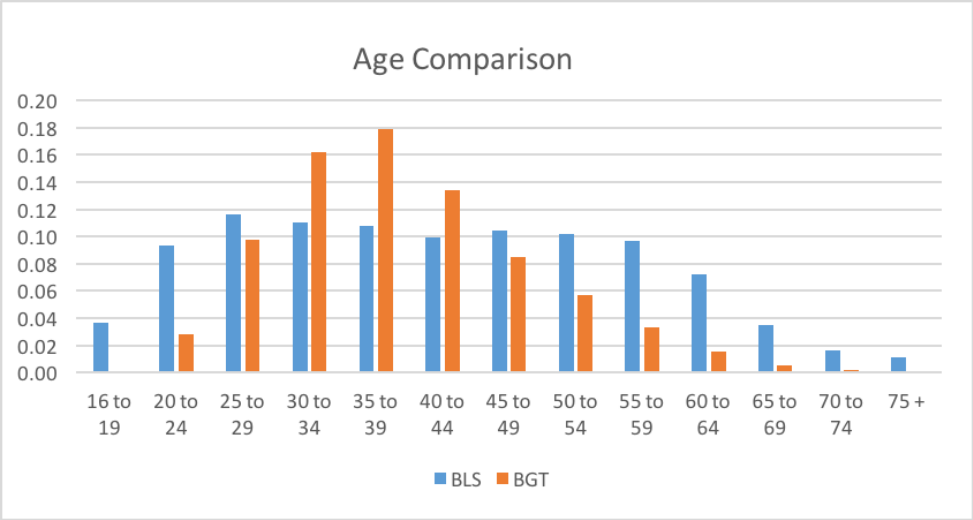
BGT Data: education relative to 2018 labor force

Figure A1: Comparison of distribution of highest educational attainment in the labor force, according to BLS data, to distribution in BGT data. Two versions are shown: BGT 1 excludes all resumes missing educational information, while BGT 2 assumes all resumes missing educational information have high school education but no college



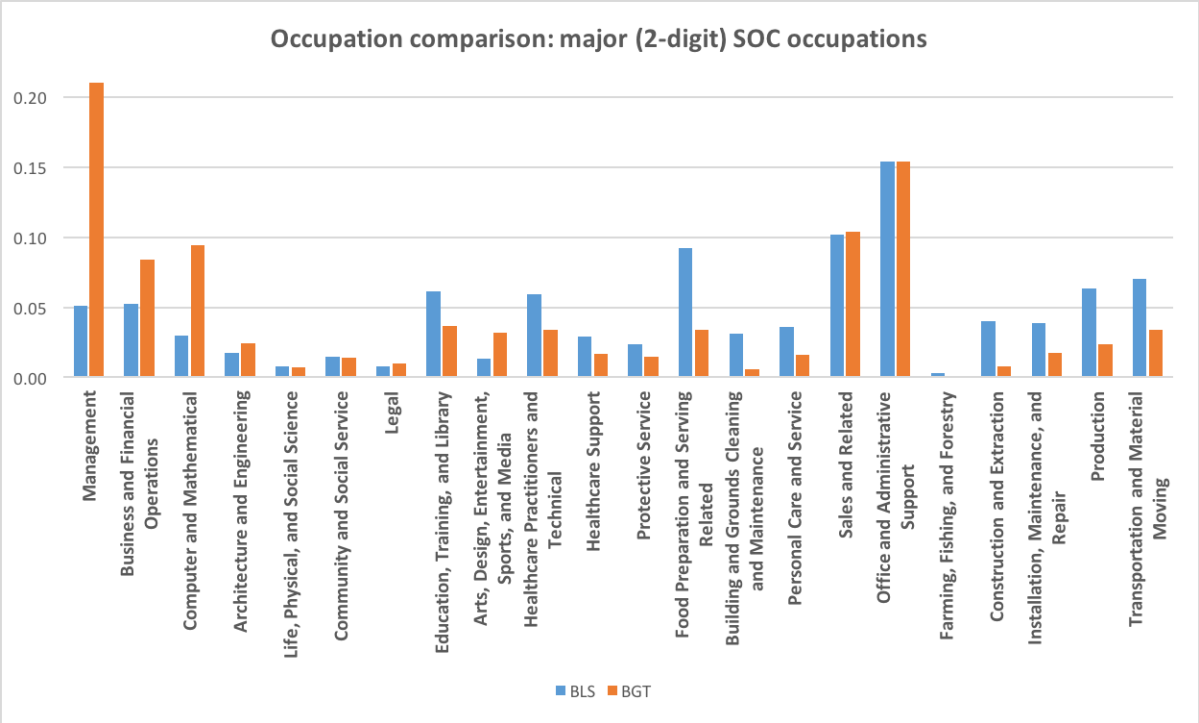
BGT Data: age distribution relative to 2018 labor force

Figure A2: Comparison of distribution of age in the labor force, according to 2018 BLS data, to distribution of imputed worker ages in BGT job sequence data.



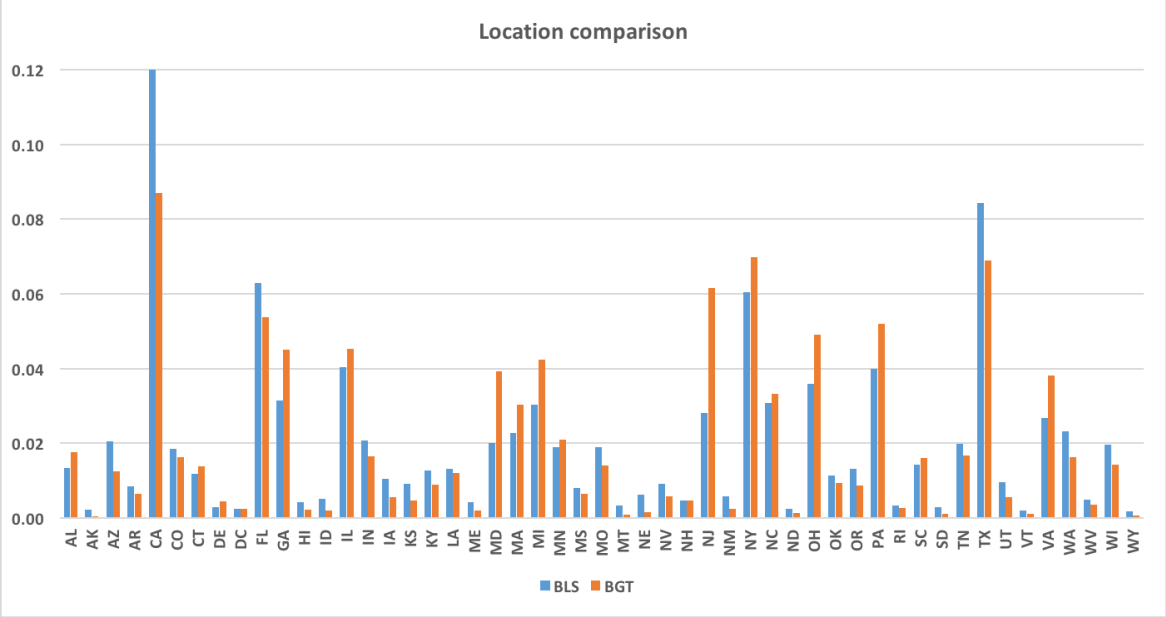
BGT Data: occupations relative to 2017 labor force

Figure A3: Comparison of distribution of 2-digit SOC occupations in the labor force, according to 2017 BLS data, to distribution of occupations in BGT job sequence data.



BGT Data: locations relative to 2017 labor force

Figure A4: Comparison of distribution of employment by U.S. state, according to 2017 BLS data, to distribution of resume addresses in BGT job sequence data. Graph shows share of total in each state.



Examples of probabilistic labor markets

The graphs below illustrates the most common occupation transition paths for counter attendants and registered nurses, respectively. For both of these occupations, the majority of people who leave their SOC 6-digit occupation also leave their SOC 2-digit occupation group, but the pattern is very different. Counter attendants' outside-occupation job options are very diverse, and are mostly lateral moves into jobs in sales, office & administrative work, and food preparation and service. In contrast, almost all registered nurses who leave their occupation do so through a promotion, becoming medical and health service managers.

Figure A5: Occupational transitions for counter attendants in the food industry. Each bubble is a SOC 6-digit occupation, and the colors represent SOC 2-digit occupational groups. The size of each bubble is proportional to the share of counter attendants in the BGT data who are observed in each destination occupation in the following year.

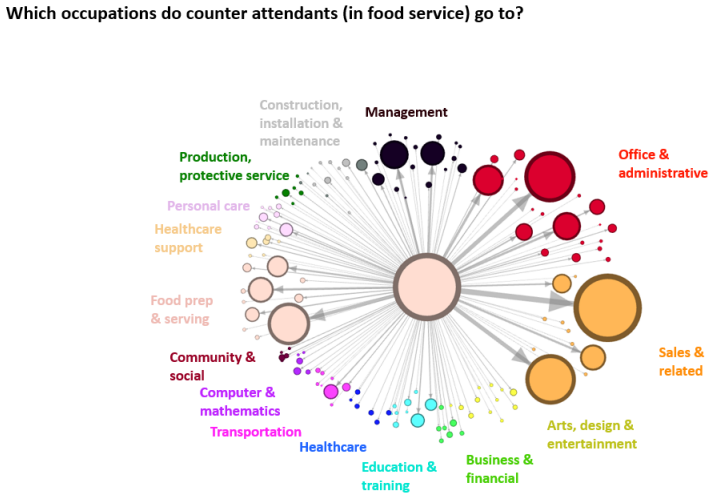


Figure A6: Occupational transitions for registered nurses. Each bubble is a SOC 6-digit occupation, and the colors represent SOC 2-digit occupational groups. The size of each bubble is proportional to the share of registered nurses in the BGT data who are observed in each destination occupation in the following year.