Measuring the Race, Ethnic, and Gender Composition of Company Workforces

Extended abstract for NBER/CRIW conference:
Race, Ethnicity, and Economic Statistics for the 21st Century

Alexander Berry
Econ One Research, Inc.

Elizabeth Maloney
The Brattle Group

David Neumark
UCI and NBER

March 16, 2023

Significance

Disparities in economic outcomes with respect to race, ethnicity, and gender are persistent in the United States. There is little doubt that labor market discrimination continues to contribute to these disparities, and that continued if not stronger enforcement of discrimination laws in the United States will help reduce these disparities. However, the data necessary to detect possible discrimination and to act to counter it is not publicly available. Nor – not surprisingly – is it readily provided by companies.

Discrimination laws in the United States are probably the strongest in the world, in part because of the ability to bring class action lawsuits on behalf of large numbers of a company's employees – which serve both to attract resources from attorneys to pursue discrimination claims and to incentivize firms to avoid these claims. Although federal and state agencies (like the EEOC, at the federal level) can file lawsuits against companies alleged to have discriminated, most enforcement stems from private-sector attorneys. Yet the private legal sector on which U.S. anti-discrimination enforcement relies is unable to access data pointing to possible discrimination at the company level. Only the federal government (through EEO-1 reporting) and state governments via similar authority obtain data on, e.g., the racial or ethnic composition of firm's workforces, or of specific occupations within those firms. Unfortunately, EEO-1 data are confidential, and the individual data on demographics is voluntarily reported and can be quite incomplete.

As a consequence, the key agents in the enforcement of discrimination laws are severely hampered in trying to target the companies that are potentially engaging in the most discrimination. Indeed, complaints of discrimination are typically initiated by small number of employees who may have personally experienced discrimination but do not have information on statistical patterns at their

employers. Data to provide more definitive evidence are typically not obtained until discovery. The cost of getting to that stage can thus deter investigations of discriminatory conduct by firms.

Researchers, policymakers, and attorneys are not completely blind to the demographic composition of firms' workforces. Some companies do make public their "diversity reports." For example, *Google* releases an annual diversity report.[1] Its 2022 report provides the percentage of hires by race, ethnicity, and gender, as well as workforce representation and attrition (the only instance we have found of reporting on retention). But it says nothing about the positions different workers occupy in the company. *LinkedIn*'s 2021 report[2] provides the race, ethnic, and gender composition of its overall, tech, and non-tech workforce, and also the composition of a vaguely defined "leadership" category (a category required in confidential EEO-1 reports). It is not difficult to find other similar reports, providing the same limited information.[3] And, conversely, there is ample information on companies' resistance to providing this information.[4] Finally, note that neither these reports, nor the EEO-1 reports, provide information on promotions. The methods we will develop and describe in this paper provide information on the race, ethnicity, and gender dimensions of all of these aspects of firms' workforces.

The project our paper will develop and describe will be useful not only for identifying potentially problematic firms for closer investigation, but also for informing researchers who would like insights into how disenfranchised groups are making inroads in fields in which they have faced substantial barriers in the U.S. economy – in particular, high-paying technical and professional firms. This fills a substantial gap in the labor market data available for the United States. We have rich household data, but these data contain no firm identifiers and typically do not include much information on the positions people hold within companies (and their levels). And we have firm level data (like the LEHD), but any identification of companies in these data is prohibited, and these data also lack information on job positions or levels. The approach we will develop and describe will provide information on the race, ethnic, and gender composition of firms' employment, and of the progress of these employees at the firms we study.[5]

Project Overview

Our project uses information extracted from publicly available *LinkedIn* data to measure the racial, ethnic, and gender composition of workforces at a small number of large companies. We will use predictive tools based on both names and photographs to identify the race, ethnicity, and gender of

[1] For 2022, see: https://about.google/belonging/diversity-annual-report/2022/.
[2] https://news.*LinkedIn*.com/2021/october/2021-workforce-diversity-report.
[3] See, e.g., https://news.linkedin.com/en-us/2022/october/2022-workforce-diversity-report; https://www.apple.com/diversity/.
[4] See, e.g., https://www.proxypreview.org/2022/contributor-articles-blog/data-transparency-key-to-improving-diversity-equity-and-inclusion-in-the-workplace; https://circaworks.com/articles/eeo-1-report-and-voluntary-disclosure/.
[5] We understand the conference is focused on race and ethnicity. It is very little additional work to include the analysis of gender, but this can be omitted from the conference paper if that is preferred by the organizers.

employees. And we will endeavor to use this information, along with information from job descriptions and titles on *LinkedIn*, to try to map differences in the race, ethnicity, and gender of firm workforces at different job levels, their promotions, and their tenure at the company (which speaks to retention).[6] The research is partly "proof of concept" to assess how well this can be done, and – we hope – partly a road map for how others can do the same.

This project thus builds on an emerging body of research that leverages data from private companies – and especially data on workers, firms, job openings, etc. – to better understand U.S. labor markets.[7] It also dovetails with greater efforts, via legislation, to increase transparency about labor markets, in part to increase the information workers have about jobs to reduce labor market frictions and increase labor market competition, as well as to reduce discrimination.[8]

Methods

Our approach, which we have already tested, is to extract publicly available *LinkedIn* data offered by the company *Proxycurl*. Given a company's *LinkedIn* profile, *Proxycurl* returns *LinkedIn* profile data for that company's employees. This can include both previous and current employees, depending on parameter selection. We will pull *LinkedIn* profile data on all current and former employees for the companies we study (see below). Though there can be variation in how much information an individual chooses to share on his or her *LinkedIn* page, the information returned often includes detailed information about previous work experience (including place of employment, title, description, start date, and end date) and education (including school, field of study, degree obtained, start date, and end date). Because individuals usually include the start and end dates for each role, we can use this to understand both tenure at a company as well as within a specific job title. These data may also include information about a

[6] There is some potential movement towards OFCCP releasing EEO-1 reports under FOIA requests. (See, e.g., https://content.govdelivery.com/accounts/USDOLOFCCP/bulletins/3495276.) It is at this time unclear how easy it will be for companies to stop release of these data by objecting on grounds of trade secrets, financial information, etc. Of course, OFCCP data only cover federal contractors. OFCCP is reluctant to release these data, so objectors may be able to easily block release (https://news.bloomberglaw.com/daily-labor-report/labor-department-reluctant-to-reveal-contractor-diversity-data).

[7] See, e.g., Azar et al. (2022) using *CareerBuilder.com* data to study monopsony; Borup and Montes Schütte (2022) to study forecasting employment growth using *Google trends*; and Clemens et al. (2021) using *Burning Glass* data to study minimum wages; Marinescu and Wolthoff (2020), also using *CareerBuilder.com* data, to study job search.

[8] For example, New York recently passed a law requiring posting of pay ranges in advertisements for all job positions (https://www.littler.com/publication-press/publication/new-york-becomes-latest-state-require-salary-transparency-job-postings) and California recently enacted pay transparency law requires posted pay ranges by demographic group (https://www.shrm.org/resourcesandtools/legal-and-compliance/state-and-local-updates/pages/california-pay-transparency.aspx). And a recent federal Executive Order (13665) prohibits federal contractors from retaliating against workers who disclose or discuss compensation information (Trotter et al., 2017). Despite expectations, some recent work suggests that pay transparency may reduce workers' bargaining power (Cullen and Pakzed-Hurson, 2021), because higher wage offers can lead to more renegotiation when pay is transparent.

worker's skills, activities, volunteer work, languages spoken, certifications, and recommendations, among other topics. It also includes their profile picture.

We use the *DeepFace* package in Python to approximate the race and gender of individuals at each company studied, based on their *LinkedIn* profile photos. For additional validation and in cases in which there is no profile picture available, we use an individual's name to try to predict their gender and race, using the Python packages *ethnicolr*, which uses statistical data to predict race by name and *gender-guesser*, which uses natural language processing patterns to predict gender based on an individual's first name.

A natural question is how reliable this approach is likely to be. We have some evidence already that it is reliable. Specifically, in a large gender pay discrimination lawsuit, we used extracted data from *LinkedIn* to recover information on jobs Oracle employees held prior to coming to Oracle.[9] We were able to match about 55% of Oracle employees in the company's data to *LinkedIn* observations. Moreover, we were able to establish that the matched data were representative; in that case, the estimated gender pay gap was similar in the full company data and the subsample matched on *LinkedIn*.

In addition, we have ways to validate the data using public data sources. In particular, for the stage of the project that is planned for this paper, we have selected to extract and study companies for which other sources of data are available to validate our methods. First, we have chosen a set of large companies that we believe to have employment concentrated in a single geographic area, preferably an area with little or no presence of other firms in the same industry. For firms meeting this criterion, we can use ACS data based on industry and PUMA of employment to compare the demographic composition of workers (and counts) in these cells in the ACS with what we obtain from *LinkedIn*.[10] Second, we will provide some evidence for publicly-traded companies for which there is other available information on firm employment. Third, we will use firms in technology/software companies, to avoid firms with sizable numbers of manufacturing workers at different sites. (We could use firms in other industries that meet this criterion as well, but we think technology/software companies are much more likely to have geographically-concentrated employment.)

Of course another criterion – which also reflects a limitation of our approach – is that employees are likely to be present on *LinkedIn*. Our sense is this is more likely the case for higher-level professional and technical workers. In fact, we are in the process of confirming this for a dataset we have on Texas A&M employees (based on a state-level FOIA-type request), where we can see that higher-level employees are much more likely to be present on *LinkedIn*.

[9] See *Expert Trial Report of David Neumark in the Matter of Jewett et al. v. Oracle America, Inc.*, December 2021, redacted.

[10] We will also do comparisons with the companies' DEI reports, when available.

We are currently reviewing publicly available information for a number of companies to determine which firms may be appropriate candidates for our analysis based on the criteria above. Potential candidates include:[11]

- *Epic Systems*, a healthcare software company established in 1979: HQ: Verona, Wisconsin; Employee Count: 10,000 (2019); NAICS Code: 541618 - Other Management Consulting Services.[12]
- *Crowdstrike*, a cybersecurity company established in 2011: HQ: Austin, Texas; Employee Count: 6,000 (2022); NAICS Code: 541511 - Custom Computer Programming Services.[13]
- *Ancestry*, a genealogy company established in 1985: HQ: Provo, Utah; Employee Count: 1,000-5,000; NAICS Code: 518210 - Data Processing, Hosting, and Related Services.[14]
- *iRobot*, a technology company established in 1990: HQ: Bedford, Massachusetts; Employee Count: 1,254 (2022); NAICS Code: 541611 - Administrative Management and General Management Consulting Services.[15]
- *Match Group*, an IT company established in 1986: HQ: Dallas, Texas; Employee Count: 1,880 (2020); - NAICS Code: 812990 - All Other Personal Services.[16]

Note that we will also be using the NETS data to further verify that there is fairly negligible employment at other firms in the same industry and PUMA. Also, although our resources likely will not permit the analysis of more companies, we have ways to identify other candidate companies. For example, we can search for STEM companies on *builtin.com*[17] that meet employment level thresholds and have only one location listed, and compare those companies to location results returned on *Indeed.com*'s company search.[18]

Because for these companies we can validate our methods, our plan is to implement our procedure for a few (or perhaps all) of these companies. We will then fine-tune our methods improve our ability to match other sources of data for these companies. And then we will use our methods to do the same analysis for one very large company (for which we would not have these validation methods available). Finally, we will try to find a company that was involved in a large race or ethnic

[11] Per the footnote above about FOIA releases of EEO-1 data, none of these companies appear on the list of *non-objectors* to the OFCCP's release of data. See https://content.govdelivery.com/accounts/USDOLOFCCP/bulletins/3468912.
[12] https://en.wikipedia.org/wiki/Epic_Systems#History; https://siccode.com/business/epic-systems-corp.
[13] https://en.wikipedia.org/wiki/CrowdStrike; https://www.infotelligent.com/company/crowdstrike-holdings/184301.
[14] https://www.indeed.com/cmp/Ancestry; https://siccode.com/business/ancestry.
[15] https://en.wikipedia.org/wiki/IRobot; https://siccode.com/business/irobot-corp.
[16] https://en.wikipedia.org/wiki/Match_Group; https://siccode.com/business/match-group-inc.
[17] https://builtin.com/companies/size/over-1000-employees?page=1.
[18] https://www.indeed.com/companies/search.

discrimination lawsuit regarding either employment or promotions, for which there are non-redacted estimates available in *LinkedIn* expert reports, and see how well we can match these findings using data.

Limitations

One limitation of this approach is that it is more applicable (and reliable) for companies with large shares of employees on *LinkedIn*. We believe lower-skilled and lower-paid jobs are less well represented on *LinkedIn* (something we will try to document in the course of the research). So this approach cannot provide information on the race, ethnic, and gender composition of the workforces of a representative set of companies. On the other hand, the composition of employment at higher-pay, higher-skilled firms, and advancement through the ranks, is critically important because these companies, and the higher-level jobs within them, are among the best jobs in the U.S. economy, and are the jobs at which minority groups (and women, in tech jobs) are under-represented.[19]

A second limitation is that we are by no means endeavoring to assemble data on, for example, all companies that have a sizable representation of their employees on *LinkedIn*. Rather, what we are doing can be viewed as a prototype, or proof of concept (a limitation also driven by our limited resources available for *LinkedIn* searches). We will develop (and make available) tools that others can use to study specific companies. And it is conceivable that a well-funded project could use our approach to construct much more comprehensive data.

Relation to issues in the NBER/CRIW call for papers

This research is related to the following issues in the call for papers:

- Challenges with the collection of data on race and/or ethnicity.
- Gaps in the availability of data by race and/or ethnicity and options for filling those gaps.
- The future of data on race/ethnicity and opportunities for innovative methods.
- Methods for combining multiple data sources, whether they be carefully designed surveys or experiments or administrative data, to produce innovative statistics on race/ethnicity.
- How the collection of more detailed race/ethnic data could affect our understanding of economic and social trends.

[19] See National Center for Science and Engineering Statistics (2023).

References

Azar, José, Ioana Marinescu, and Marshall Steinbaum. 2022. "Labor Market Concentration." *Journal of Human Resources* 58(1): 167-199.

Borup, Daniel, and Erik Christian Montes Schütte. 2022. "In Search of a Job: Forecasting Employment Growth Using Google Trends." *Journal of Business and Economic Statistics* 40(1): 186-200.

Clemens, Jeffrey, Lisa B. Kahn, and Jonathan Meer. 2021. "Dropouts Need Not Apply? The Minimum Wage and Skill Upgrading." *Journal of Labor Economics* 39(S1): S107-49.

Cullen, Zoe B., and Bobak Pakzad-Hurson. 2021. "Equilibrium Effects of Pay Transparency." NBER Working Paper No. 28903.

Marinescu, Ioana, and Ronald Wolthoff. 2020. "Opening the Black Box of the Matching Function: The Power of Words." *Journal of Labor Economics* 38(2): 535-68.

National Center for Science and Engineering Statistics (2023). *Diversity and STEM: Women, Minorities, and Persons with Disabilities*. National Science Foundation Directorate for Social, Behavioral and Economic Sciences, NSF 23-315.

Trotter, Richard G., Susan Rawson Zacur, and Lisa T. Stickey. 2017. "The New Age of Pay Transparency. *Business Horizons* 60(4): 529-39.