Preliminary:
Please do not cite, quote, or circulate.

Measuring the Race, Ethnic, and Gender Composition of Company Workforces Using *LinkedIn* Data[*]

Alexander Berry
Econ One Research

Elizabeth Maloney
UCI

David Neumark
UCI, NBER

February 2024

Abstract

Stronger enforcement of discrimination laws can help to reduce disparities in economic outcomes with respect to race, ethnicity, and gender in the United States. However, the data necessary to detect possible discrimination and to act to counter it is not publicly available – in particular, data on racial, ethnic, and gender disparities within specific companies. In this paper, we explore and develop methods to use information extracted from publicly available *LinkedIn* data to measure the racial, ethnic, and gender composition of company workforces. We use predictive tools based on both names and pictures to identify race, ethnicity, and gender. We show that one can use *LinkedIn* data to obtain reasonably reliable measures of workforce demographic composition by race, ethnicity, and gender, based on validation exercises comparing estimates from scraped *LinkedIn* data to two sources – ACS data, and company diversity or EEO-1 reports. And we apply our methods to study the race, ethnic, and gender composition of workers who experienced a mass layoff at a large company.

[*] This paper was prepared for the NBER-CRIW Conference: Race, Ethnicity, and Economic Statistics for the 21st Century. We have received helpful comments from Randall Akee, Lawrence Katz, and Mark Lowenstein. The data used in this paper are proprietary, and hence we cannot produce the full dataset for other researchers to use. However, the paper fully explains how the data can be accessed for purchase. We received modest financial support for this project from Econ One Research ("Econ One"). Alexander Berry is employed by Econ One and David Neumark does consulting work on discrimination cases with Econ One. Econ One had no right of review or comment on this research, and all conclusions are our own.

Introduction

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 to reduce these disparities. However, the data necessary to detect possible discrimination and to act to counter it is not publicly available – in particular, data on racial, ethnic, and gender disparities within specific companies.[1] Nor – not surprisingly – are such data readily provided by companies.

In this paper, we explore and develop methods to use information extracted from publicly available *LinkedIn* data to measure the racial, ethnic, and gender composition of company workforces. We use predictive tools based on both names and pictures to identify the race, ethnicity, and gender of employees. And we explore using this information, along with information from job histories on *LinkedIn*, to develop estimates of racial, ethnic, and gender differences in employment, hiring, and retention.

This paper 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. 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.[2]

[1] As discussed below, there are confidential data sources (like the LEHD) where these disparities can be measured, but not for small numbers of companies and never identifying those companies.

[2] For example, New York recently passed a law that requires firms to post 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's 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). Further, a recent federal

A particularly valuable application of our research is that it can be used to strengthen enforcement of discrimination laws. The strength of discrimination laws in the United States rests on class action or other "pattern and practice" lawsuits on behalf of large numbers of a company's employees. The large potential penalties/awards in these lawsuits 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 – and enforcement against the largest companies – stems from private-sector attorneys. But there are three problems, all of which our research can help address.

First, 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 firms' workforces, or of specific occupations within those firms. These data are confidential.[3]

Second, these data only measure employment (with limited information on occupational distributions). They do not measure hiring or retention.

Third, and most important, private attorneys – the key agents in the enforcement of discrimination laws – are severely hampered in trying to target the companies that potentially engage in the most discrimination. Complaints of discrimination are typically initiated by a small

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. We regard this as a still-open question requiring more research.

[3] 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 block release easily (https://news.bloomberglaw.com/daily-labor-report/labor-department-reluctant-to-reveal-contractor-diversity-data).

number of employees who may have personally experienced discrimination (or believe they have), but do not have information on statistical patterns at their employers. Plaintiffs' attorneys work on contingency fees, and hence have to decide whether to invest large sums in filing charges and commencing discovery before they can see any data on potentially discriminatory behavior. The uncertainty involved can deter them from taking on cases, and reduce the efficiency of how their resources are targeted.[4] The kinds of information we extract from *LinkedIn* data could potentially lead to more efficient targeting of these efforts. By helping attorneys identify where discrimination might be occurring and where it might not be occurring, these data could thus help make discrimination law more efficient, allowing attorneys and enforcement agencies to concentrate their efforts and resources on the companies where there is a higher probability that discrimination is occurring.[5] Moreover, the methods we develop could provide anti-discrimination enforcement agencies with additional tools to monitor companies.

Our core research questions are: How can the *LinkedIn* data best be used to characterize companies' employment by race, ethnicity, and gender? How reliable are these data? And, can the *LinkedIn* data be used to study other company workforce decisions – in particular hiring, retention, and promotions – and their relation to workforce demographics?

Our focus is to a large extent on the Professional, Scientific, and Technical Services Sector, owing in large part to strong representation on *LinkedIn*. That is not to say that our methods cannot be usefully applied to other industries, although some of our conclusions about

[4] This oversimplifies slightly, as state anti-discrimination agencies sometimes partner with private attorneys. But in our experience, this is rare.

[5] Anecdotally, we have spoken with a handful of plaintiffs' attorneys who work on discrimination cases about the potential value of using the *LinkedIn* data in this way. We have confirmed that some indicated they would find it useful. One indicated they had already used the data in this way (although of course absent the kind of validation, etc., we explore in this paper). And below we cite use of *LinkedIn* data in a different manner in a discrimination case.

the representativeness of the *LinkedIn* data might not apply as strongly and future work could assess this. At the same time, jobs in this sector are significant with regard to enforcement of discrimination laws, since this sector is an important source of high-paying jobs and upward mobility.

Social media data vs. scientific samples

The evidence we report in this paper indicates that the *LinkedIn* data correspond reasonably well, with some exceptions, to other probability-sample estimates of workforce demographics. However, the value of what we can do with the *LinkedIn* data does not hinge only on the representativeness of the data being so good as to be able to claim that the estimates (e.g., of the share black) are unbiased, and we would not expect this because the *LinkedIn* data do not provide a probability sample of a company's workforce. At the same time, while the *LinkedIn* database is not the population of workers, and is not a scientific sample, it covers a very large number of workers and can hence be used to generate quite reliable measures.

First, a growing body of research trying to study labor markets using social media data acknowledges the tradeoff between probability sampling and the ability to learn from social media data what we cannot learn from other data. As examples, Schneider and Harknett (2019a, 2019b) use targeted ads on Facebook to study work schedules, based on a 1.2% response rate. Similarly, a number of labor economists use data on job postings or job applications to study monopsony, discrimination, minimum wages, and other topics (e.g., Azar et al., 2022; Borup and Montes Schütte, 2022; Clemens et al., 2021; Marinescu and Wolthoff, 2020; and Neumark et al., 2019) – in our view learning a great deal more than we could otherwise, despite data sources being not fully representative.

Second, it is critical to emphasize that the core value of the *LinkedIn* data for enforcement of discrimination laws is that it provides some reliability in estimates of race, ethnic, or gender differences in the outcomes the data are used to measure, to improve targeting of what would generally be more thorough investigations. As explained earlier, absent this type of information, private attorneys considering class action discrimination claims have only anecdotal evidence to rely on, and hence can be deterred from initiating lawsuits. Information from *LinkedIn* can provide a critical complement to this anecdotal evidence – and hence improve targeting of anti-discrimination efforts – even if estimated differences by race, ethnicity, and gender have some biases due to imperfect representation.[6]

Alternative data sources?

This project considers the development of a data source that can fill substantial gaps in 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 information on the positions people hold within companies. In principle, the LEHD could be used to provide descriptive information similar to some things we can measure with the *LinkedIn* data. However, there are a number of limitations of the LEHD data.

First, a core limitation is that the LEHD could never be used as an enforcement tool the way we are envisioning, both because company names could never be revealed, and because those for whom this tool would be useful would never be able to access these data, to use them to study a single company. Second, the LEHD would pose severe challenges to doing this in a timely manner, both because the LEHD is updated slowly, and because securing access to the

[6] And it is important to keep in mind that *LinkedIn* claims 200 million active U.S. users, so some representativeness is assured.

LEHD and then working with the data is a slow process.[7] Further, the LEHD data does not have any information on job titles that could be used to study employees' positions within companies. In contrast, the *LinkedIn* data do not present these restrictions, and – importantly from both an enforcement and research perspective – are up-to-date and immediately accessible.

The LEHD is an extraordinarily valuable and powerful data set that can be leveraged for the analysis of discrimination,[8] although not as readily for the specific questions we will be studying with the *LinkedIn* data. At the same time, it is possible that our own work with the *LinkedIn* data will prompt us to work with the LEHD to provide a more extensive and higher quality characterization of employment and other dynamics one can measure with the LEHD in relation to race, ethnicity, and gender, or prompt others to do so. We regard these as complementary efforts, with different strengths, weaknesses, and purposes.

The National Establishment Time Series (NETS) is a proprietary data set in which company names are public and can be used in research (e.g., Burnes et al., 2014), but it contains no worker information aside from employment. (As indicated above, however, we do make some use of the NETS data in this project to select companies for our validation work.[9])

Finally, 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.[10] Its 2022 report provides the percentage of hires by race, ethnicity, and gender, as well as workforce representation and

[7] As an example, one of the most recent LEHD publications we could find on *Google Scholar* is a 2022 publication using LEHD data through 2016 (McKinney and Abowd, 2022); and the earliest working paper version of this paper appears to be from 2020.

[8] See, e.g., Barth et al. (2021), Brick et al. (2023), and Hu (2019).

[9] Neumark has access to the NETS data on a contract related to other research.

[10] For 2022, see: https://about.google/belonging/diversity-annual-report/2022/.

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[11] provides the race, ethnic, and gender composition of its overall, tech, and non-tech workforces, 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.[12] And, conversely, there is ample information on companies' resistance to providing this information,[13] and we might expect that it is supplied selectively across companies. Finally, note that neither these reports, nor the EEO-1 reports, provide information on retention, and typically not on hiring either. The methods we will develop and describe in this project can provide information on the race, ethnicity, and gender dimensions of all of these aspects of firms' workforces.

Approach and methods

We do our research using extracted, publicly available *LinkedIn* data offered by the company *Proxycurl*.[14] Given a company's *LinkedIn* profile, *Proxycurl* returns *LinkedIn* profile data for that company's employees with public profiles. The data reflect the publicly available *LinkedIn* profiles at the time they are scraped. We can query the current *LinkedIn* profile for each employee who has linked a company of interest as an employer, either past or current, depending on parameter selections. This means we can get a snapshot into past as well as current

[11] https://news.*LinkedIn*.com/2021/october/2021-workforce-diversity-report.

[12] See, e.g., https://news.linkedin.com/en-us/2022/october/2022-workforce-diversity-report; https://www.apple.com/diversity/.

[13] 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/.

[14] See https://nubela.co/proxycurl/linkdb for more details.

employment.[15] However, the data do not simply cover all employees at a single point in time. Rather, the scraping captures all individuals who have ever worked at the company, as long as they have not deleted their data. We also typically know when they were employed at the company, so we can approximate employment by year, hiring, exits, etc. We extract *LinkedIn* profile data on all current and former employees for 7 companies, the selection of which we describe below.

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). All *LinkedIn* data are self-reported, and voluntarily reported. These data may also include information about a worker's skills, activities, volunteer work, languages spoken, certifications, and recommendations, among other topics. Critically, for our purposes, it also often includes their profile picture, along with their name.

We use the *DeepFace* package in Python to classify workers by race, ethnicity, and gender, based on their *LinkedIn* profile pictures and a common training dataset ("picture classification"); see Serengil and Ozpinar (2020, 2021). We also use R packages (*rethnicity*, *gender*), which use statistical data to classify based on names ("name classification"). The *DeepFace* package is trained on the *FairFace* dataset for race/ethnicity identification, using default weights that are the same as the ones we use. The gender prediction model for picture classification is trained on Wikipedia data. These both return probabilities that the worker is in each group. For details on both types of classifications, see Serengil and Ozpinar (2020, 2021). We also supplement with name classification when picture is not available, and we will combine

[15] See https://nubela.co/proxycurl/docs for more details.

information when both are available.

We use binary classifications (black/non-black, Hispanic/non-Hispanic, and female/male), based on highest probabilities returned by these programs. We use names as the primary method for gender and ethnicity, and pictures as the primary method for race, based on evidence described below. Where the classification probability is missing for the primary method, the other method is used. We also change which method we use if the primary method gives a quite uncertain classification and the secondary method gives a far more certain classification, as explained in more detail below.[16]

One of our core goals in this paper is to validate our classifications against external data, to see how reliable the *LinkedIn* data are. For example, one might wonder whether particular demographic groups are under- or over-represented on *LinkedIn*. One approach to validating the *LinkedIn* data is to leverage corresponding information in two other data sources – the National Establishment Time Series (NETS), and the American Community Survey (ACS). For this validation exercise we proceed in two steps. First, we use data from the NETS, along with *LinkedIn* information from the *LinkedIn* website, to identify companies that are in a broad industry category that has good representation on *LinkedIn*. In particular, we focus on the Professional, Scientific, and Technical Services Sector. As documented in Table 1, which reports results for the top 10 *Fortune* companies, this sector (see the highlighted rows) has good representation on *LinkedIn* – in the sense that a large share of the companies' workers appears on

[16] We considered using probabilities to construct estimates of demographic shares of the workforce, etc., based on weighted averages using these probabilities. However, we found that this was not as accurate for the race and ethnicity coding, because there are other minority groups (such as Asians) that can receive some probability weight, which results in lower estimated shares black or Hispanic than we get from using the highest probability.

the website, based on current employment (column (3)).[17]

We then use the NETS data to select companies and areas to make the ACS and *LinkedIn* data comparable – i.e., so when we extract ACS data on workers by POWPUMA and industry, we should be sampling by and large from employees of these companies in the corresponding geographic area. In particular, we identify companies in this sector that meet three criteria: (1) fairly negligible employment at other firms in the same industry and Place-of-Work PUMA (POWPUMA); (2) most of the company's employment in the POWPUMA is in the industry; and (3) the company has strong representation on *LinkedIn*. The idea behind criterion (1) is that these firms constitute most of industry employment in the POWPUMA. Thus, ACS workers in the industry and POWPUMA likely work for these companies. The idea behind criterion (2) is that the company's employment in the POWPUMA is concentrated in one industry. This is critical because we have an industry identifier in the ACS but not in *LinkedIn*. Thus, if the company had POWPUMA employment in other industries, the ACS data for a single industry might not be representative of the company's POWPUMA employment. Because these companies are largely unique in their industry-location cells, if they also have good representation on *LinkedIn* (criterion 3), the measures of race, ethnic, and gender composition from the two data sources should correspond. Thus, we measure the race, ethnic, and gender composition of ACS employment in those industry-location cells, and then compare to our estimates based on the *LinkedIn* data.

These ACS restrictions greatly limit the number of company comparators that we are able

[17] Column (2) captures those ever employed at the company. The low number for CVS Health in the *Proxycurl* database is because of search constraints imposed when using the *Proxycurl* database by buying tokens for a specific number of searches, which constrains the search to workers who include company urls in their *LinkedIn* profiles. Based on our investigations, non-professional workers are much less likely to do this.

to benchmark against. We therefore also use a second validation approach, comparing the *LinkedIn* results to companies' DEI reports, when available, or other sources of information on the demographic composition of their workforces. This approach avoids the constraints on companies dictated by the first approach. In particular, we were able to find information in diversity reports or other information companies provided, and we use these, when available, for both the companies selected to validate against the ACS data, and other companies we selected.[18]

Companies selected

We identified four companies that meet the criteria for the ACS validation discussed above, and that are not too large (allowing the required data extraction within our budget constraints). Table 2 reports the companies meeting the following constraints:

1. Company's NAICS industry employment in POWPUMA vs. all NAICS industry employment in POWPUMA > 70%
2. Company's NAICS industry employment in POWPUMA vs. all company employment in POWPUMA > 80%
3. Firm employment > 800.

The table also shows *LinkedIn* employment – in this case based on ever employed, since our validation with the ACS is not based on only one year of data. We had to constrain the choice among these based on number of employees, given our budget constraint, and we also constrained it based on *LinkedIn* data showing a large share of employment in a single nearby

[18] There is already some limited evidence of the reliability of the *LinkedIn* data. Specifically, in a large gender pay discrimination lawsuit, *LinkedIn* data were extracted on jobs Oracle employees held prior to coming to Oracle. (See *Expert Trial Report of David Neumark in the Matter of Jewett et al. v. Oracle America, Inc.*, December 2021, redacted.) It was possible to match about 55% of Oracle employees in the company's data to *LinkedIn* observations, and to establish that the matched data were representative in one dimension; in particular, in that case the estimated gender pay gap was similar in the full company data and the subsample matched on *LinkedIn*.

geographic area, since otherwise we would not expect the ACS to provide a very relevant comparison. The four companies we then chose for this validation exercise are shown in the shaded rows of Table 2. The non-shaded rows are for companies that met our criteria with regard to POWPUMA, but were either very large – so that we could not afford the data extraction, had low representation on *LinkedIn*, low representation in the geographic area (which likely has to do with international employment not measured in the NETS), or were quite small so that we would not be able to learn that much from the data.

The additional companies selected were GlaxoSmithKline (GSK), SpaceX, and Meta. GSK made our initial list for the ACS validation, but its employment was not sufficiently geographically concentrated to be useful for this validation. Meta and SpaceX are two high-profile technology companies. And for one (Meta), we actually found EEO-1 data posted, whereas for the other (SpaceX), we found nothing. Thus, we could validate the Meta data from *LinkedIn* against the EEO-1 data, and also illustrate the potential value of our approach for a company for which neither EEO-1 nor diversity report information was available. We restricted our choice to other companies in the same sector, for comparability. Finally, we chose among these based on the ability to cover a number of companies while remaining within our data budget constraint. Clearly future work with greater funding could expand the scope of these types of analyses.

Data extraction and classification

For these companies, we extract publicly available data from *LinkedIn* from the *Proxycurl LinkedIn* database. To do this, we provide the company's *LinkedIn* url. We request current and past employees (who can be distinguished in the database). The application then returns all data from public profiles (employment history, education, skills, etc.).

Across the seven companies for which we extracted data, we obtain 112,280 worker profiles, of whom 78,639 are in the United States. The numbers and distributions of these observations are displayed in Table 3. We get a sizable number of observations from all companies except Research Corporation of the University of Hawaii, and very large numbers of observations for GSK, Meta, and SpaceX.

We retain only those working in the United States. We break the data for each person into separate entries for each job at each company at which they worked. Together, this results in 557,329 separate observations, although many of these are not at our companies of interest.

We then use the extracted *LinkedIn* data to do our classifications, based on the *DeepFace* package in Python for picture classification, and R packages (*rethnicity*, *gender*) for name classification.[19] Before reporting our findings on demographic composition and more, there are some results about classification that are of interest, and which dictate how we use this information.

There are certainly caveats to using these methods. First, some pictures on *LinkedIn* make identification difficult. For example, some show multiple people, obscured images, or have bad lighting, as shown in the examples in Figure 1. In addition, sometimes there is no picture available. Overall, we have pictures, for which we run the *DeepFace* classification code, for 96,651 profiles (which includes pictures from non-U.S. profiles). Of these, 22.95% were missing a picture, and 1.09% had an image in which a face could not be detected. The distributions of these cases were roughly stable across the companies, as shown in Table 4.

Second, we cannot always classify people by name. We run the name classification code

[19] We utilized *RetinaFace* for the face detection backend, and otherwise used a pre-trained neural network that comes with the package.

for 75,393 names.[20] We are unable to classify gender for 9.89% of names, and unable to classify race/ethnicity for 4.55% of names. This can occur if the name recorded on an individual's LinkedIn profile does not include their first name (e.g., "Lt. Higgins") or if their name only includes foreign characters, for example. For predicting gender, if a name is sufficiently uncommon that it does not appear in the Social Security database used to predict gender, then no gender probability will be assigned to it. For race prediction, we do not use an individual's last name to predict race when they only provide a single initial in place of their last name (the program assumes all one-letter names are Asian).

Third, some names were problematic. For example, some first names – like "Alex" – are not strongly gendered. Some last names – like "Monte" – could be of Hispanic origin or another ethnicity. Some names are classified as more likely to be black or white, without providing a strong confirmation – e.g., "Steve Fulton," with a 72% probability of being black. And, of course, last name (or even first name) changes can obscure race or ethnicity. In these cases, pictures may provide more definitive information.

We use additional information on how the two programs classify people by race, ethnicity, and gender to settle on our classification "algorithm." First, as shown in Figure 1A, the distributions of probabilities that observations are female, whether using names or pictures, are bimodal, with probabilities clustered near zero or one. This reflects the fact that names are highly gendered, as is physical appearance. There is a little more mass at the endpoints using names (about 85%) than using pictures (about 80%), which is why we use names as the first source of classification by gender.

Second, the story for ethnicity and race classification is more complicated. The charts in

[20] Note that this differs from the number of profiles above because we do not run the algorithm separately for repeated names as the predicted gender/race will not differ for individuals with the same name.

Figure 2A for both ethnicity and race show large spikes at zero probability, but do not show much evidence of bimodality. Similarly, there is far less mass at the lower and upper ends of the range for Hispanic or black classifications than for gender classifications, and correspondingly more mass between these points for Hispanic or black classifications – and more so for Hispanic classifications. These findings reflect a combination of lower shares Hispanic or black than female, of course, but also reflects less definitive assignments of probabilities, possibly a reflection of less distinct physical differences than those by gender, in part because shading of pictures can obscure race or ethnicity. And it is likely because these differences are less pronounced for Hispanics that there is, in the bottom panel of Figure 2A, a good deal more mass at lower values but above zero probability for Hispanic than black, and conversely much more mass at zero (more accurately, in the band 0-2) probability for blacks. We learn more about what is happening at the higher probabilities from Figure 2B, which shows more details at the higher probabilities by focusing in on the upper half of the distributions.[21] We now see much more clearly that for black classifications the distribution of probabilities is more bimodal, with a spike at 100. We thus rely on pictures as the first source of classification for blacks. For Hispanic classifications, there is a much more pronounced mass of probabilities at the top of the distribution using names than using pictures, so we use names as the first source of classification for Hispanics.

As a result of these considerations, as well as the inability sometimes to classify people by gender, race, or ethnicity based on either a picture or a name, we use the following algorithm to classify people.

[21] Note that in Figure 2B the vertical scales are not the same in each graph, so that we could better highlight the details.

Race

1. If the picture probability is non-missing, we classify people as black based on the picture if the probability black based on picture is the highest among all "race" categories.[22]
2. If the picture probability is missing but the name probability is non-missing, we classify people as black based on the name if the probability black based on name is the highest among all "race" categories.
3. If both are non-missing, we rely on pictures, except when the picture classification is highly uncertain but the name classification is not. Specifically, when no race probability based on pictures (among the 6 groups)[23] > .5, but the probability black based on name > .9, we classify people as black.[24]

Ethnicity

1. If the name probability is non-missing, we classify people as Hispanic based on the name if the probability Hispanic based on name is the highest among all "race" categories.
2. If the name probability is missing but the picture probability is non-missing, we classify people as Hispanic based on the picture if the probability Hispanic based on picture is the highest among all "race" categories.
3. If both are non-missing, we rely on names, except when the name classification is highly uncertain but the picture classification is not. Specifically, when no race

[22] The classification programs do not separately code race and ethnicity as commonly defined by, e.g., the U.S. Census, but rather include these in the same overall "race" classification.

[23] For pictures, these are Asian, black, Indian, Latino/Hispanic, Middle Eastern, and white. For name, these are Asian, black, Latino/Hispanic, and white.

[24] This only results in a re-classification if the probability black based on the picture was not the highest.

probability based on name (among the 4 groups) > .5, but the probability Hispanic based on picture > .9, we classify people as Hispanic.[25]

Gender

1. If the name probability is non-missing, we classify people as female based on the name if the probability female is higher.
2. If the name probability is missing but the picture probability is non-missing, we classify people as female based on the picture if the probability female based on picture is higher.

We ran the classification code for all worker-company observations. After creating the race, ethnicity, and gender identifiers, we had 29,843 observations that were missing either race or gender, approximately 5.4% of the data. In terms of race, we have 97.44% coverage, and our gender variable covers 96.82% of the data.[26]

Table 5 reports on the probabilities of classification by each category based on pictures and names, including the initial classifications and the final classifications. We can see in this table some of the same results from Figures 2A and 2B, and also the consequences of our final rules for classification. For example, we noted that pictures are far more reliable for classifying black vs. non-black than names. This is reflected in the first row, columns (1)-(8), in the higher probabilities black for pictures at each of the percentiles reported. Note, though, that we obtain far more classifications based on name, so as a result the probabilities in columns (10)-(11), which are often based on name, are lower than in columns (2)-(5). In contrast, the probabilities

[25] This only results in a re-classification if the probability Hispanic based on the name was not the highest.

[26] Evaluations of these methods point to fairly reliable classification. For *DeepFace*, see Serengil and Ozpinar (2021). As discussed in Blevins and Mullen (2015), the *gender* package in R assigns a probability that an individual is female based on the historical frequency at which women are observed with that name using Social Security Administration data since 1930s.

Hispanic are far higher for names. For gender, the probabilities are very high in both cases, with the difference (lower probabilities based on pictures) only apparent at lower percentiles (column (2) vs. columns (6)).

It is worth mentioning here that we deliberately did not tune or adjust our methodology regarding classifications to try to better match the proportions we see in the ACS or the diversity/EEO-1 reports. We are trying to demonstrate the value of these data in doing these classification exercises for other companies. Since a potential user presumably would not have any data source other than the *LinkedIn* data, there would be no basis for adjustments to the classifications. As a result, potential users should be most interested in the accuracy of race, ethnic, and gender classifications that use out-of-the-box algorithms (as we do), without any further fine-tuning.[27]

Classification of companies' workforces and validation with ACS data

Based on race, ethnic, and gender classifications, we first report our results for demographic classifications of companies' workforces for the four companies for which we can perform our validation exercise with ACS data. For these comparisons, we restrict the relevant areas in the *LinkedIn* data as follows: Virginia for BWXT; Hawaii for Research Corporation of the University of Hawaii; Virginia for Chesterfield County, and Illinois for Fermi Research Corporation. For the ACS comparison we limit observations to those observations with at least a portion of relevant experience at the company between 2012 and 2021 (a 10-year window). Our *LinkedIn* dataset treats each job at which a person works as an observation. It is the case that *LinkedIn* employment at each company generally trends upwards over time, which we imagine is

[27] At the same time, we recognize that additional exploration could help identify other code that works better, refine the existing code, or find alternative ways to use the resulting probability estimates.

due in part to increasing numbers of employees on *LinkedIn*. Regardless of the reason, we weight the ACS data by year to be proportional to the representation in the *LinkedIn* data.

These results are reported in Table 6. The ACS numbers are weighted annual averages, based on 2012-2021 data for the same POWPUMA and 4-digit NAICS code. We also show similar results for professional, technical, and managerial occupations, which are likely over-represented on *LinkedIn*.[28] The data are also displayed in more digestible form in Figure 3. The results for the percent female do not indicate tight concurrence of the estimates, but there is some correspondence. For example, looking at the overall ACS numbers, the rank order across the four companies is the same for the ACS and *LinkedIn* data, and the *LinkedIn* percentages are notably higher where the ACS estimates are (for Research Corporation of Hawaii and Chesterfield County), and vice versa. The data for the Research Corporation of the University of Hawaii should probably be ignored, given the low representation in the *LinkedIn* data and the ACS data. The results for the percent Hispanic do not correspond very well. The estimates for the percent black, excluding Research Corporation of the University of Hawaii, exhibit reasonable correspondence between the two measures, with the rank order the same in both data sets, and the values matching reasonably well. The results are very similar, generally, with the occupational restrictions, except for the sharp decline in the percent black at BWXT.

Of course, one issue is that the ACS samples are not very large.[29] In addition, despite our best efforts, we pick up workers at other companies, and the geographic match is not exact. We

[28] That said, the sample sizes when we restrict to these occupations are only a bit smaller, consistent with (a) most workers at these companies being in these occupations, and (b) little apparent bias from the hypothesized over-representation of these occupations on *LinkedIn.* A potential caveat, however, is that differences between the occupations represented on *LinkedIn* and overall workforces may be more marked for other industries.

[29] The ACS is a 1% random sample for the years we use (https://usa.ipums.org/usa/acs.shtml).

thus, in the next section, turn to comparisons between the *LinkedIn* data and other sources of direct measures of workforce composition at the companies in question.

Classification of other companies' workforces and other validation efforts

We looked for other sources of information on company demographics, including company diversity reports and other information posted on their websites. We also found that some companies post actual EEO-1 reports.[31] Table 7 indicates, for each company, what kind of information we could identify. As described in the notes to the table, which also provide the source, we sometimes had to do some computations with the available numbers and make some assumptions.[32]

As the table indicates, we obtain numbers to compare, either for a single year or, in the case of Meta, many years, for three companies: Chesterfield County and Fermi Research Alliance, LLC, for which we also did the ACS validation exercise, and Meta, which fortuitously provides data from two sources, in one case (its diversity report) for many years.[33]

For the first two, the observation counts in *LinkedIn* are quite low, both because of smaller companies and the restriction to a single year. And as the table notes, for Fermi, in particular, there are so few observations that the comparison should probably be ignored. For Chesterfield County, however, there is some correspondence. In both data sources, the percent

[31] We explored with both the U.S. EEOC and the California equivalent – the Civil Rights Division – about obtaining such data. But at the company level they are confidential and could not be shared.

[32] There are some SpaceX numbers available from "Zippia." It is unclear where data come from. According to its website, Zippia gets company information from employee self-reporting, public and open data sources on the internet, and proprietary data licensed from other companies. Data sources include, but are not limited to, the BLS, company filings, H1B filings, public websites on the internet, and other public and private datasets. https://www.zippia.com/employer/zippia-faq/.

[33] For the overlapping data – the percentages black and Hispanic for 2021 – the data are very close but do not match exactly.

female is the highest, followed by the percent black, and then the percent Hispanic, and the numbers roughly correspond (e.g., the percent female is much higher in both data sources).

The data for Meta are perhaps the most interesting, because (i) we have far more observations, and (ii) we can check some data by year. The numbers for the percent black are close, and the pattern of increase in this percentage is similar in the two data sources, as is the amount (2.9 percentage points in the diversity report, and 3.1 percentage points in *LinkedIn*). The numbers for percent Hispanic are not as close, perhaps because of how Hispanics are defined (although we find no mention of this in the Meta documents), but again the pattern of increase is similar, as is the amount (2.7 percentage points in the diversity report, and 1.6 percentage points in *LinkedIn*). The comparison with the 2021 EEO-1 Report looks similar for the percentages black and Hispanic, which is not surprising. And the percentage female is reasonably close.

What do we conclude from the validation exercise? First, there is clearly some correspondence between measures of workforce demographic composition for the *LinkedIn* data and other sources. For our ACS comparisons, given that we do not have an exact match, and that there is sampling error in the ACS data, we would not expect exact matches, so this is encouraging. Put differently, there is no reason to assume the ACS data are more reliable. But the rough correspondence is encouraging for using the *LinkedIn* data, and of course the *LinkedIn* data can in principle be used to study any company, whereas the ability to use the ACS to learn something about a company's workforce is highly limited to companies with a sizable share of industry employment in a POWPUMA. Second, and reassuringly for the *LinkedIn* data, the correspondence appears to be much tighter for large companies – although admittedly this is based on data for one company (Meta).

Recall, though, our perspective on the utility of the *LinkedIn* data. The data do not provide scientifically valid estimates of workforce composition with known sampling properties. Rather, they are interesting as a guide to further exploration by government agencies or attorneys seeking to enforce discrimination laws. Our take is that these data may be useful for both larger and smaller companies, although of course more reliable – and hence more useful – for larger companies. That actually meshes well, however, with the way data are likely to be used in enforcing discrimination laws, as the class action suits that rely on statistical evidence typically are against large companies.[34]

Overall demographic composition

Having established what we view as reasonable reliability of the *LinkedIn* data, in Table 8 we report the overall demographic composition for each company – the shares black, Hispanic, and female. For the four companies with which we did the ACS validation, we have much larger numbers of observations. This is mainly because we do not restrict to the geographic areas identified in *LinkedIn* to correspond to the ACS POWPUMA.[35] In addition, we do not restrict the time period to the 10 years covered by the ACS analysis. The percent black varies substantially, from a low of 6.3% at Research Corporation of the University of Hawaii to 26.9% at Chesterfield County. The variation in the percent Hispanic is less pronounced, but ranges from 5.7% at BWXT to 18.4% at SpaceX. In contrast, the percent female is lowest at SpaceX (17.5%) and

[34] As examples, there have been fairly recent class action discrimination lawsuits against IKEA (https://www.consolelaw.com/court-unseals-order-conditionally-certifying-age-discrimination-collective-action-suit-against-ikea-filed-by-console-mattiacci-law/), Google (https://www.nytimes.com/2022/06/12/business/google-discrimination-settlement-women.html), Walmart (https://www.cohenmilstein.com/case-study/wal-mart/, including a very large case two decades ago), and Twitter (https://www.reuters.com/legal/twitter-beats-disabled-workers-lawsuit-over-layoffs-now-2023-05-08/).

[35] We cannot map directly to POWPUMA in the *LinkedIn* data for two reasons. First, the geographic information is less specific in *LinkedIn* (we use state). Second, the geographic information in *LinkedIn* is current and may not correspond to when the person worked at the company.

highest at Chesterfield County (56.2%). It is also substantially higher than SpaceX at some of the other private employers, like GSK (44.9%) and Meta (38.1%). We caution, again, that any inference of discrimination would have to consider the composition of potential workers, which can vary with other factors – probably most notably geography and educational levels and fields.

An application to mass layoffs

Based on what we have learned about the reliability of the *LinkedIn* data, we explore using these data to study potential discrimination in mass layoffs. In particular, we examine layoffs at Meta, which is reported to have laid off 11,000 employees (13% of its workforce) in 2022, and 10,000 in 2023.[36] We use the *LinkedIn* data to measure all separations, identifying the spell of time an employee spends at the same company and inferring that an employee has separated with a company if they either stop working for at least 4 months or if their next employment is with a different company.[37] We see inordinately high numbers of separations in 2022 and 2023, as shown in Table 9.[38] We suspect the 2023 numbers may be lower than reported layoffs because either the layoffs were not yet implemented when we extracted the data in fall of 2023, or some people do not update their profiles until they get a new jobs. Similarly, some of the 2022 layoffs may be reflected in the 2023 data. Given that many separations in normal times

[36] See: https://www.forbes.com/sites/qai/2022/12/07/meta-layoffsfacebook-continues-to-cut-costs-by-cutting-headcount/?sh=5e36a1898456; https://www.washingtonpost.com/technology/2023/05/23/meta-layoffs-misinformation-facebook-instagram/.

[37] If a spell of time at one company is entirely overlapped by a spell at another company, we drop it from our data set. This can happen, for example, if a Ph.D. student is employed as a teaching assistant with a university for 5 years, but also lists a summer internship on their resume for a summer during their Ph.D. We would then drop the spell of time at the internship because it is completely overlapped by the teaching assistant position, rather than considering the teaching assistant as having separated from the university during the tenure of their internship. We do not drop partially overlapping positions, however.

[38] Table 9 also exhibits the rising employment over time at companies, as noted earlier.

are voluntary quits,[39] from the point of view of learning about discrimination in layoffs is it more informative to look at a period of mass layoffs.

We compare the demographic composition of those laid off vs. the workforce as a whole. Given the large numbers of separations in 2021, 2022, and 2023, we present comparisons for these three years, separately and combined. As reported in Table 10, for all three groups – blacks, Hispanics, and women – the layoff rate is higher than their representation in the workforce. Looking at 2021-2023 combined, for blacks the difference is relatively small in absolute terms (0.61 percentage point), although the difference is large in relative terms, with blacks over-represented by 9.7% among those laid off. The absolute differences are larger for Hispanics and women, but the relative differences in the same ballpark.[40] Finally, in a simple statistical test used as a heuristic, treating the layoff and workforce samples as independent and testing the equality of proportions, the difference is statistically significant for each group (the numbers of observations and layoffs are both very high).

To be clear, this is not a "test" of discrimination because other factors could account for disproportionate layoffs among some groups. That said, this kind of evidence could be far more reliable than anecdotal evidence one or a few plaintiffs present to a government agency or private attorney.

We did similar calculations for hires. In general, the representation of blacks, Hispanics, and women among hires was very similar to that among the workforce, as shown in Table 11 (with some variation from year to year). Of course, with regard to hiring the more relevant

[39] See, e.g., https://www.bls.gov/news.release/pdf/jolts.pdf, indicating about a 2-to-1 ratio of quits vs. layoffs/discharges.

[40] Recall the result documented in Table 7 that the percent Hispanic at Meta in the *LinkedIn* data is much higher than in the data reported by the company.

question is the comparison of the race, ethnic, or gender composition of hires relative to the hiring pool. Nonetheless, this table suggests that one could use estimates like these, relative to "benchmark" estimates of the composition of the hiring pool sometimes used in discrimination cases, usually from the ACS, to obtain provisional evidence on discrimination in hiring.[41,42]

Potential limitations

One inherent limitation of our approach is that it is more applicable (and reliable) for companies with large shares of employees on *LinkedIn*. We suspect that lower-skilled and lower-paid jobs are less well represented on *LinkedIn*. 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 at these companies, is critically important because these companies, and the higher-level jobs within them, are among the best jobs in the U.S. economy. These are also the jobs at which minority groups (and women, in tech jobs) are under-represented.[43]

A second limitation is the possibility of fake profiles on *LinkedIn.* We have no information on how pervasive this is. We also do not know any algorithm to identify these

[41] For an example in a discrimination case, see Expert *Report of David Neumark in the matter of Heldt et al. v. Tata Consultancy Services, Ltd.*, February 2017. For examples and discussion of EEOC guidance using benchmarks from the Decennial Census or the ACS, see https://www.eeoc.gov/federal-sector/management-directive/instructions-federaal-agencies-eeo-md-715-1 and Amano-Patiño et al. (2022). For an example at the state level see https://www.twc.texas.gov/sites/default/files/enterprise/docs/equal-employment-opportunity-minority-hiring-practices-report-2016-twc.pdf.

[42] We considered trying to study promotions as well. However, tabulating job titles over time a worker spends at a company did not clearly signal whether a promotion occurred, in part because most job titles do not indicate a clear promotion (such as moving to a job title of the same name with a level indicator), no doubt exacerbated by the fact that job titles to not appear to be reported in any uniform way on *LinkedIn*. For example, in our sample of 57,326 work experiences at Meta, we observe 24,732 unique job titles, 21,422 of which only appear once.

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

profiles.[44] Still, this is another reason one should not interpret the type of analysis we do with the *LinkedIn* data – or that others could do – as definitive with respect to measuring discrimination, but rather as indicative of possible discrimination. Finally, there is some evidence that *LinkedIn* seems to be fairly successful at stopping fake accounts.[45]

It is also important to clarify that the kind of evidence that can be produced with the *LinkedIn* data is not intended to be rigorous evidence of discrimination. Our goal is to try to use the *LinkedIn* data to produce estimates of race, ethnic, and gender differences in employment, hiring, and separations (especially layoffs). Our ability to look within companies can provide new descriptive evidence that is currently not available to researchers. And this evidence and methods, as we have argued, will be useful in enforcing discrimination laws. We want to be clear, however, that we are not proposing our measures based on the *LinkedIn* data as evidence of discrimination per se. Other factors can explain sorting of workers across firms, as well as differences in retention and promotion.

The intention is not that the *LinkedIn* data would necessarily be the data actually used to establish the definitive evidence of discrimination for either legal proceedings or research. The strongest evidence would typically require richer company data, both for reliability and comprehensiveness, and to rule out other non-discriminatory explanations.[46] These data typically become available at later stages of litigation. But our findings suggest that the *LinkedIn* data can

[44] Rather, “advice” on spotting them is based on reading and assessing individual profiles, and does not appear very systematic. (E.g., https://www.forbes.com/sites/forbesbusinesscouncil/2022/11/17/how-to-identify-a-fake-linkedin-profile-in-five-minutes-or-less/?sh=1421f73f1d7c; https://www.linkedin.com/pulse/dangers-fake-profiles-how-spot-one-david-smith-cv-writer).

[45] See https://www.cnbc.com/2022/12/10/not-just-twitter-linkedin-has-fake-account-problem-its-trying-to-fix.html.

[46] On the latter point, for example, one might get information on promotions from job titles at a company and be able to test whether education and prior jobs explain any difference. But more rigorous evidence would likely require performance ratings, as well as perhaps a more definitive identifier of promotions.

be used to increase the precision with which potentially discriminatory companies could be targeted for further legal exploration, including filing of discrimination claims and opening of legal discovery to access the richer data that companies have on both workforce outcomes and potential factors accounting for those differences. And, conversely, the use of these data might prevent spurious lawsuits against companies less likely to be discriminating.

Summary and conclusions

We have shown that one can use *LinkedIn* data to obtain reasonably reliable measures of workforce demographic composition by race, ethnicity, and gender, based on validation exercises comparing estimates from scraped *LinkedIn* data to two sources – ACS data, and company diversity or EEO-1 reports. To be clear, though, we do this validation for a small number of companies, limited by a restriction to one industry (which we suspect is better represented on *LinkedIn*) and to a small number of companies dictated by our research budget. This validation is further restricted to companies that that can be compared to ACS data because they represent a high share of industry employment in a POWPUMA, and companies that make public diversity reports or EEO-1 reports. Our evidence cannot speak to the universe of industries or companies.

That said, we emphasize that the research we present in this paper is to some extent a "proof of concept" (or more accurately an assessment of a proof of concept), exploring how well our ideas for measuring the demographic composition of companies' workforces, and the relation of other decisions of these companies to demographics, can be measured. The methods we develop and explore – which we anticipate will be improved upon by others – can be used in three principal ways.

First, and most directly related to this paper, the *LinkedIn* data can be used by plaintiffs' attorneys or agencies charged with enforcing discrimination laws. We illustrate how this might be done by studying mass layoffs at one of the companies for which we extracted *LinkedIn* data. We first show that the data are sufficiently comprehensive to detect mass layoffs. And we then illustrate using the data to compare the race, ethnic, and gender composition of laid off workers to the workforce as a whole. Again, we caution that this is not definitive evidence of discrimination, both because the *LinkedIn* data do not provide a scientific sample (indeed, in an actual discrimination case, data on all employees covered by the case would typically be available), and because of a lack of control variables (although some could in principle be constructed from the *LinkedIn* data). Still, this kind of evidence would be much more convincing to attorneys or government agencies than anecdotal evidence from a handful of laid-off workers.

Second, we imagine that researchers will develop other ways to use these data to measure and study the demographics of the workplace, such as the evolution of employment by race, ethnicity, and gender, overall (or at least in industries well-represented on *LinkedIn*), and extending to other questions like the progress women and minorities are making in reaching higher-level positions within companies. Indeed, while in this paper we are only able to study data on limited number of companies (for cost reasons), the company from which we draw *LinkedIn* data does make available the entire public *LinkedIn* database.[47]

Finally, we think our demonstration of how we use the *LinkedIn* data may prompt consideration of other potential research uses one could make of these data. One can, for example, construct work histories and educational histories, as well as job titles (although the job titles are quite idiosyncratic). Thus, for example, the *LinkedIn* data could in principle be used to

[47] Currently, the cost to purchase this database is $40,000. The cost is about four cents per record (with bulk purchases of thousands of searches). This cost quickly becomes prohibitive for large companies, and purchasing the whole dataset obviously would be very cost-effective.

study the impact of education on careers, and to study career trajectories including both changes within companies and mobility across companies.

References

Amano-Patiño, Noriko, Julian Aramburu, and Zara Contractor. 2022. "Is Affirmative Action in Employment Still Effective in the 21st Century?" U.S. Census Bureau, Center for Economics Studies Working Paper 22-54.

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

Barth, Erling, Sari Pekkala Kerr, and Claudia Olivetti. 2021. "The Dynamics of Gender Earnings Differentials: Evidence from Establishment Data." *European Economic Review* 134: 103713.

Bayard, Kimberly, Judith Hellerstein, David Neumark, and Kenneth Troske, 2003, "New Evidence on Sex Segregation and Sex Differences in Wages from Matched Employer-Employee Data," Journal of Labor Economics, pp. 887-922.

Blevins, Cameron, and Lincoln Mullen. 2015. "Jane, John... Leslie? A Historical Method for Algorithmic Gender Prediction." *DHQ: Digital Humanities Quarterly* 9(3).

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.

Brick, Carmen, Daniel Schneider, and Kristen Harknett. 2023. "The Gender Wage Gap, Between-Firm Inequality, and Devaluation: Testing a New Hypothesis in the Service Sector." Forthcoming in *Work and Occupations*.

Burn, Ian, Patrick Button, David Neumark, & Luis Felipe Munguia Corella. 2022. "Does Ageist Language in Job Ads Predict Age Discrimination in Hiring?" *Journal of Labor Economics* 40: 613-667.

Burnes, Daria, David Neumark, and Michelle J. White. 2014. "Fiscal Zoning and Sales Taxes: Do Higher Sales Taxes Lead to More Retailing and Less Manufacturing?" *National Tax Journal* 67(1): 7-50.

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.

Deng, Jiankang, et al. 2019. "RetinaFace: Single-Stage Dense Face Localization in the Wild." arXiv: 1905:00641v2.

Google. 2022. "2022 Diversity Annual Report." https://about.google/belonging/diversity-annual-report/2022/.

Hellerstein, Judith K., Mark Kutzbach, & David Neumark. 2019. "Labor Market Networks and

Recovery from Mass Layoffs: Evidence from the Great Recession Period." *Journal of Urban Economics* 113: 103192.

Hellerstein, Judith, David Neumark, and Melissa McInerney, 2008, "Spatial Mismatch or Racial Mismatch?" Journal of Urban Economics, pp. 464-79.

Hellerstein, Judith, Melissa McInerney, and David Neumark, 2011, "Neighbors and Co-Workers: The Importance of Residential Labor Market Networks," Journal of Labor Economics, pp. 659-95.

Hellerstein, Judith, and David Neumark, 2004, "Ethnicity, Language, and Workplace Segregation: Evidence from a New Matched Employer-Employee Data Set," Annales d'Economie et de Statistique, pp. 19-78.

Hellerstein, Judith K., David Neumark, & Kenneth Troske. 1999. "Wages, Productivity, and Worker Characteristics: Evidence from Plant-Level Production Functions and Wage Equations." *Journal of Labor Economics* 17: 409-446.

Hu, Lingqian. 2019. "Racial/Ethnic Differences in Job Accessibility Effects: Explaining Employment and Commutes in the Los Angeles Region." *Transportation Research Part D: Transport and Environment* 76: 56-71.

LinkedIn. 2021. "Our 2021 Workforce Diversity Report." https://news.*LinkedIn*.com/2021/october/2021-workforce-diversity-report.

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.

McKinney, Kevin L., and John M. Abowd. 2022. "Males Earnings Volatility in LEHD Before, During, and After the Great Recession." *Journal of Business and Economic Statistics* 41(1): 33-9.

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.

Neumark, David. 2018. "Experimental Research on Labor Market Discrimination." *Journal of Economic Literature* 56: 799-866.

Neumark, David. 2012. "Detecting Evidence of Discrimination in Audit and Correspondence Studies." *Journal of Human Resources* 47: 1128-1157.

Neumark, David, Ed., 2007, Improving School-to-Work Transitions. Russell Sage Foundation.

Neumark, David, Ian Burn, & Patrick Button. 2019. "Is It Harder for Older Workers to Find Jobs? New and Improved Evidence from a Field Experiment." *Journal of Political Economy* 127: 922-970.

Neumark, David. 1999. “Labor Market Information and Wage Differentials by Race and Sex.” *Industrial Relations* 38: 414-445.

Schneider, Daniel, and Kristen Harknett. 2019a. “Consequences of Routine Work-Schedule Instability for Worker Health and Well-Being.” *American Sociological Review* 84(1): 82-114.

Schneider, Daniel, and Kristen Harknett. 2019b. “What’s to Like? Facebook as a Tool for Survey Data Collection.” *Sociological Methods & Research* 51(1): 108-40.

Serengil, Sefik Ilkin, and Alper Ozpinar. 2020. “LightFace: A Hybrid Deep Face Recognition Framework.” *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, Istanbul, Turkey, 2020, pp. 1-5.

Serengil, Sefik Ilkin, and Alper Ozpinar. 2021. "HyperExtended LightFace: A Facial Attribute Analysis Framework," *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, Istanbul, Turkey, 2021, pp. 1-4.

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

Xie, Fangzhou. 2022. “Rethnicity: An R Package for Predicting Ethnicity from Names.” *Software X* 17: 100965.

Figure 1: Examples of *LinkedIn* Pictures

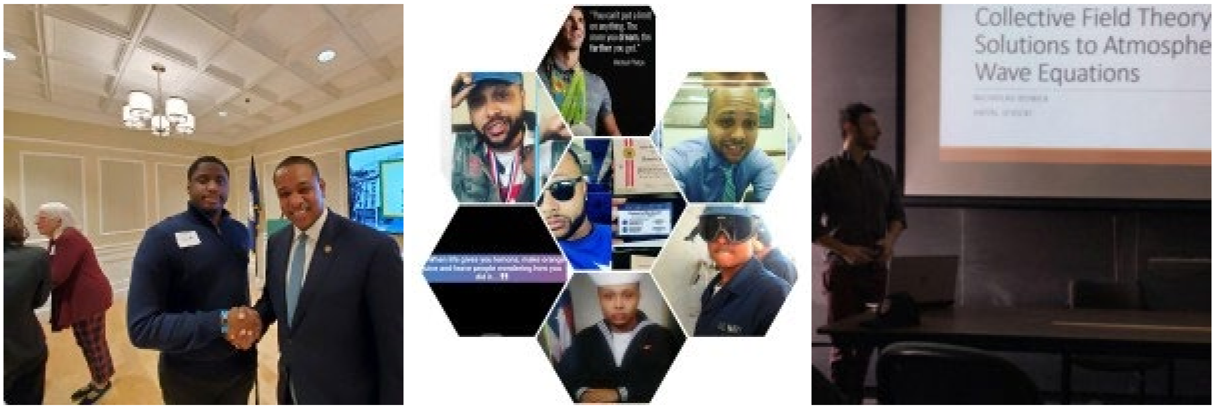


Figure 2A: Distributions of Probabilities of Gender, Ethnicity, and Race

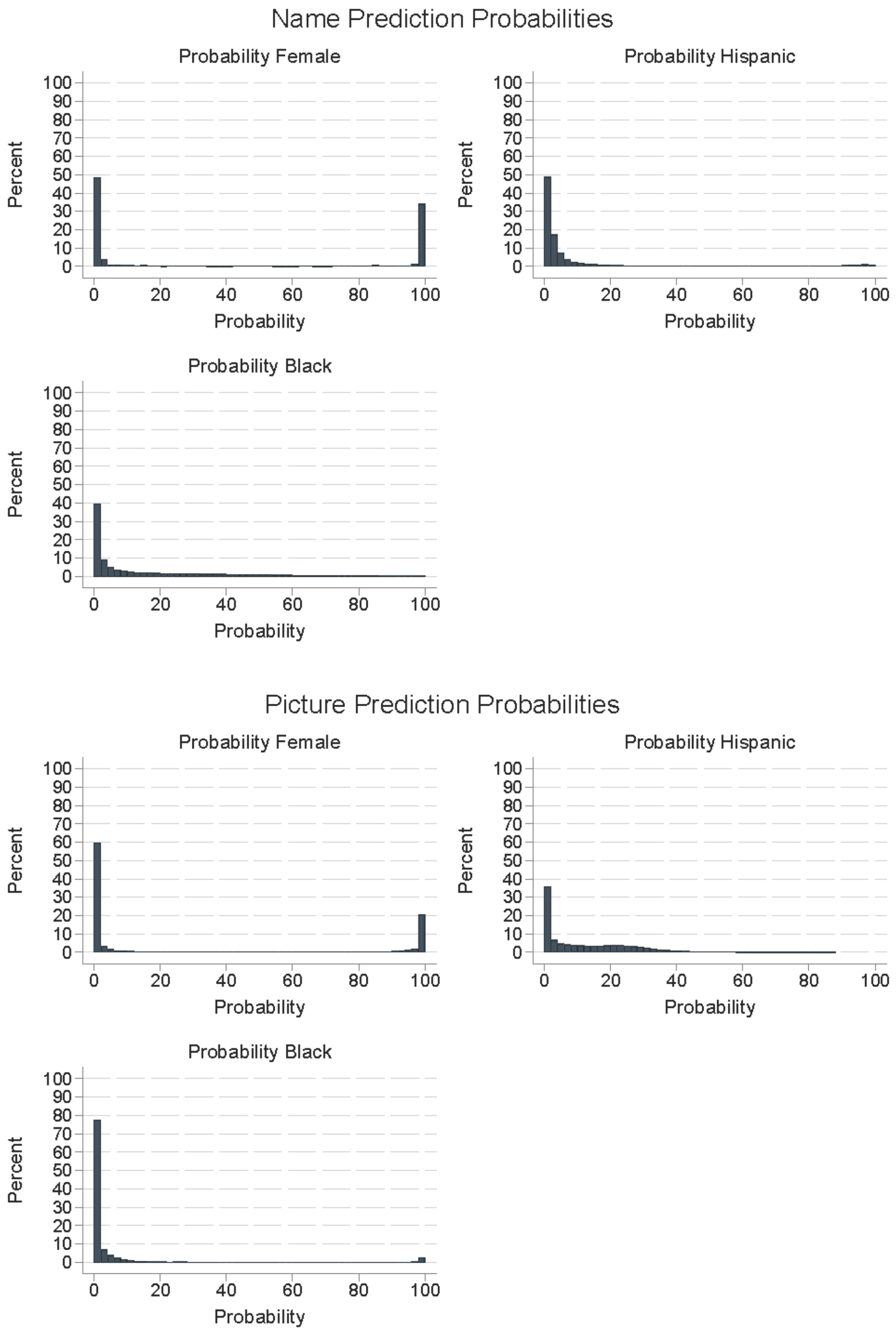


Figure 2B: Distributions of Probabilities of Gender, Ethnicity, and Race – Upper Tails

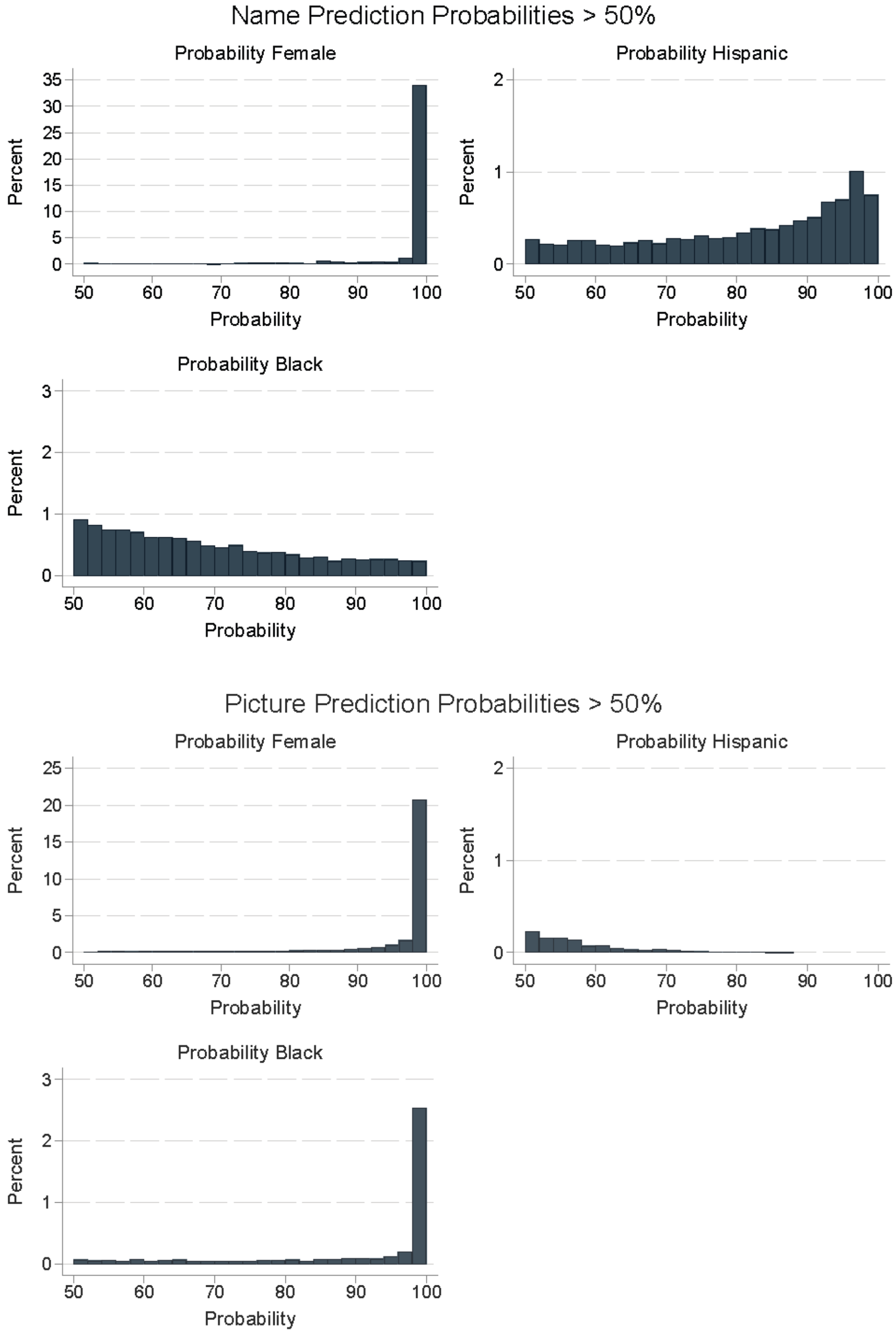
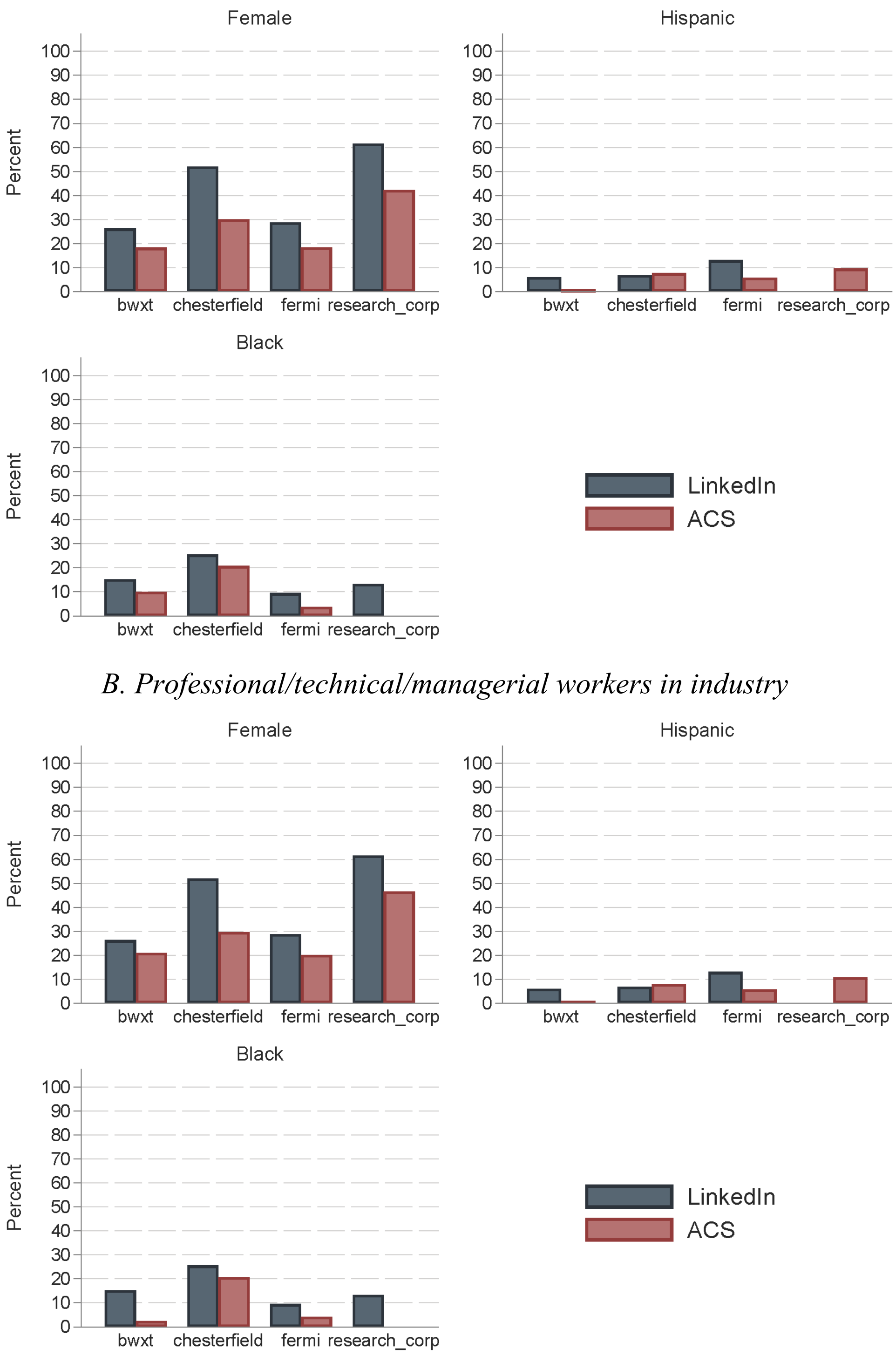


Figure 3: Validation Estimates for Percentages Black, Hispanic, and Female in *LinkedIn* and American Community Survey Data



A. All workers in industry

B. Professional/technical/managerial workers in industry

Notes: See notes to Table 6.

Table 1: Fortune 1-10 Companies (by Revenue) Comparing 10k and *LinkedIn* Employment

Company	**10k employment**	***LinkedIn* webpage employment**	***LinkedIn* database *(Proxycurl)* employment (current)**
	(1)	(2)	(3)
Walmart	2.2 million	389,386	94,192
Amazon	1.54 million	841,260	182,960
Apple	132,000	289,924	97,927
CVS Health	300,000	115,472	175
UnitedHealth Group	400,000	167,345	39,871
Exxon Mobil	63,000	57,735	20,939
Berkshire Hathaway	372,000 (many subsidiaries)	8,198	990
Alphabet	190,000	280,107 (Google)	124,743
McKesson	68,000	21,260	10,370
AmerisourceBergen	44,000	19,566	4,556

Source: https://www.zyxware.com/articles/4344/list-of-fortune-500-companies-and-their-websites and 10k reports.

Note: Proxycurl database employment is for those who link to company URL in their LinkedIn profile.

Table 2: Companies Identified using National Establishment Time Series Data as having Large Share of Total Industry Employment in Place-of-Work PUMA, Large Share of Total Company's Employment in Place-of-Work PUMA, Good Representation on *LinkedIn*, and of Moderate Size for Initial Data Extraction

NAICS	Company	*LinkedIn* webpage empl.	*LinkedIn* webpage employment in nearby geographic "area"	% of POWPUMA empl.	% of firm empl.
(1)	(2)	(3)	(4)	(5)	(6)
5417	Research Corporation of the University of Hawaii	286	229	82	100
5413	BWXT Technical Svcs Group Inc	1928	722	81	100
5415	Chesterfield County	2296	2022	73	99
5417	Fermi Research Alliance LLC	2069	1771	96	100
5415	Cognizant Tech Sltions US Corp	317047	34429 (US)	91	100
5417	Corning Research & Dev Corp	19106	5008	92	100
5417	Charles River Labs Intl Inc	14288	1619	86	100
5413	Tungland Corporation	215	172	76	100
5417	GlaxoSmithKline LLC	97000	5000	70	100

Table 3: *LinkedIn* Worker Profiles Extracted

Company	Total	U.S.
BWXT Technical Svcs Group Inc	2,187	1,705
Chesterfield County	2,104	2,093
Fermi Research Alliance LLC	4,215	3,738
GlaxoSmithKline LLC	26,636	10,761
Meta	61,265	45,274
Research Corporation of the University of Hawaii	66	65
SpaceX	15,807	15,003
Total	112,280	78,639

Table 4: Missing Pictures or No Face Detected

Company	Total	Classified	Missing Picture	No Face Detected
BWXT Technical Svcs Group Inc	1753	1319	388	46
Chesterfield County	1249	952	265	32
Fermi Research Alliance LLC	3198	2486	652	60
GlaxoSmithKline LLC	22985	17106	5684	195
Meta	54256	41855	12030	371
Research Corporation of the University of Hawaii	60	48	11	1
SpaceX	13150	9652	3150	348
Total	96651	73418	22180	1053

Table 5: Classification Methods Used and Result Probabilities

	Picture: probability black/Hispanic/female				Name: probability black/Hispanic/female				Final: probability black/Hispanic/female			
	N	25th percentile	Median	75th percentile	N	25th percentile	Median	75th percentile	N	25th percentile	Median	75th percentile
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Initial classifications:												
Black vs. non-black	2,648	62.71%	97.90%	100.00%	11,117	50.96%	60.78%	74.65%	6,289	53.79%	71.33%	95.84%
Hispanic vs. non-Hispanic	4,966	30.96%	36.21%	42.96%	7,964	61.53%	81.68%	93.43%	9,979	49.31%	72.87%	91.20%
Female vs. male	15,017	96.92%	99.83%	99.99%	24,851	99.49%	99.72%	99.89%	26,098	99.40%	99.72%	99.90%

Table 6: Validation Estimates for Percentages Black, Hispanic, and Female in *LinkedIn* and American Community Survey Data

Company	Data source	*LinkedIn* employment in area (N)	*LinkedIn* database public profiles (N)	ACS (N)	% black	% Hispanic	% female
Research Corporation of the University of Hawaii	*LinkedIn*	229	15		13.0	0.0	61.5
	ACS			67	0	9.6	42.3
	ACS, Prof/Tech Occ's			58	0	11.8	46.6
BWXT Technical Svcs Group Inc	*LinkedIn*	722	403		15.2	6.2	26.3
	ACS			116	10.0	0.7	18.3
	ACS, Prof/Tech Occ's			95	2.4	0.9	21.0
Chesterfield County	*LinkedIn*	2,022	859		25.5	7.1	52.0
	ACS			227	20.8	7.7	30.2
	ACS, Prof/Tech/Mgr Occ's			217	20.7	8.0	29.7
Fermi Research Alliance LLC	*LinkedIn*	1,771	721		9.5	13.4	28.7
	ACS			174	3.8	5.8	18.4
	ACS, Prof/Tech/Mgr Occ's			151	4.2	5.8	20.2

Sources/notes: *LinkedIn* and ACS data, 2012-2021, same POWPUMA and 4-digit NAICS. The *LinkedIn* data in this table are restricted to the relevant areas to correspond to the ACS POWPUMA, as described in the text. We use ACS weights, and for each company also reweight by year corresponding to the representation by year in the *LinkedIn* data. Professional, technical, and managerial occupations exclude, based on 2018 occupation codes: Community and Social Services Occupations 2001-2060; Service Occupations 3601-4655; all occupation 6005 and higher (Farming, Fishing, and Forestry; Construction Trades; Installation, Maintenance, and Repair; Production; Transportation and Material Moving; Military). See: https://www2.census.gov/programs-surveys/cps/methodology/Occupation%20Codes.pdf.

Table 7: Validation Against Other Sources of Data on Workforce Composition

Company	Source	Statistics reported	Comparable *LinkedIn* data, for corresponding years
Research Corporation of the University of Hawaii	Nothing from company		
BWX Technologies, Inc.	Nothing from company		
Chesterfield County of Virginia	EEO Utilization Report, 2020	% black: 18.9 % Hispanic: 3.7 % female: 47.7	% black: 17.2 % Hispanic: 11.2 % female: 60.3
Fermi Research Alliance, LLC	Fermilab webpage, 2023	% black: 5.6 % Hispanic: 9.3 % female: 28.1	% black: 0 % Hispanic: 0 % female: 33 Note: based on only 4 *LinkedIn* observations in 2023.
GlaxoSmithKline (GSK)	GSK website	40% of senior roles were held by women, up from 38% in 2020; 50% of manager roles held by women 27.1% of senior leaders in the U.S. were "ethnically diverse" in 2021, up from 23.2% in 2020	Cannot compare because no way to reliably match roles, and "ethnically diverse" is vague.
SpaceX	Nothing from company		
Meta	2022 Annual Diversity Report	% black 2014: 2 2015: 2 2016: 2 2017: 3 2018: 3.5 2019: 3.8 2020: 3.9 2021: 4.4 2022: 4.9 Hispanic 2014: 4 2015: 4 2016: 4 2017: 5 2018: 4.9 2019: 5.2 2020: 6.3 2021: 6.5 2022: 6.7	% black 2014: 4.0 2015: 4.8 2016: 5.2 2017: 5.7 2018: 6.2 2019: 6.3 2020: 6.5 2021: 7.1 2022: 7.1 Hispanic 2014: 12.6 2015: 13.3 2016: 13.4 2017: 13.2 2018: 13.3 2019: 13.4 2020: 13.8 2021: 14.4 2022: 14.2
	2021 EEO-1 Report	Total/Professionals % black: 4.6/4.4 % Hispanic: 6.7/6.4 % female: 36.2/34.6	Total % black: 7.1 % Hispanic: 14.4 % female: 41.9

Notes and sources: **Chesterfield**. https://www.chesterfield.gov/DocumentCenter/View/446/EEOP-DOJ-Utilization-Report-PDF. Numbers reported for 6 categories of jobs: Officials/Administrators; Professionals; Technicians; Protective Services (Sworn); Protective Services (Unsworn); and Administrative Support. We have assumed these include the entire workforce. **Fermi**. https://www.fnal.gov/pub/about/demographics/. Numbers reported for 6 categories of jobs: Technical; Scientists; Postdocs; Mission Support; Engineers, Computing. We have assumed these include the entire workforce. Some numbers are reported as < 5 but < 0. We assume values of 2. **GSK**: https://us.gsk.com/en-us/responsibility/diversity-equity-and-inclusion/#inside-gsk. **Meta**: https://about.fb.com/wp-content/uploads/2022/07/Meta_Diversity-Data-Summary-Report_2022.pdf. % female is also reported by year, but only globally. % by ethnic group is also reported for Tech, Non-Tech, and Leadership. 2021 EEO-1 Report also reports numbers for: Executive/Senior Officials & Managers; First/Mid Officials & Managers; Professionals; Technicians; Sales Workers; Administrative Support; Craft Workers; Operatives; Laborers & Helpers; and Service Workers. Professionals are the vast majority.

Table 8: *LinkedIn* Demographic Composition Estimates for Each Company

Company	N	% Black	% Hispanic	% Female
Research Corporation of the University of Hawaii	65	6.3	7.8	45.9
BWXT Technical Svcs Group Inc.	1,704	11.4	5.7	22.7
Chesterfield County	2,090	26.9	6.4	56.2
Fermi Research Alliance LLC	3,735	9.3	13.8	28.5
GlaxoSmithKline (GSK)	10,616	10.0	10.3	44.9
SpaceX	14,929	8.4	18.4	17.5
Meta	45,035	6.6	12.5	38.1

Note: This table reports the demographic composition for each company. Observations are at the employee-firm level, and proportions are calculated based on the number of classified individuals in that company. In contrast to Table 6, this table does not restrict to the relevant *LinkedIn* area to correspond to the ACS POWPUMA.

Table 9: Separations at Meta in *LinkedIn* Data

	Workforce	**Separations**
2012	393	11
2013	667	20
2014	1,129	35
2015	1,778	56
2016	2,903	106
2017	5,102	189
2018	8,576	359
2019	12,857	611
2020	18,834	736
2021	29,777	2,512
2022	39,751	9,336
2023	30,964	5,104
Total	152,731	19,075

Table 10: Meta Workforce and Layoffs Demographic Composition

	% black in workforce	% black in layoffs	% Hispanic in workforce	% Hispanic in layoffs	% female in workforce	% female in layoffs
2021	6.47%	7.32%	12.47%	13.65%	37.22%	41.04%
2022	6.28%	6.45%	11.95%	13.34%	36.16%	38.60%
2023	6.24%	7.64%	11.54%	13.75%	35.36%	40.50%
2021-2023	6.33%	6.94%	11.98%	13.51%	36.23%	39.54%
Absolute difference, layoff % vs. workforce %		-0.61%		-1.53%		-3.31%
Relative difference, layoff % vs. workforce %		-9.66%		-12.78%		-9.13%
P-value for difference, 2021-2023	0.00		0.00		0.00	

Notes: P-values based on tests of equal proportions assuming independence samples. Across the three years 2021-2023, there are 100,492 employees and 16,952 layoffs.

Table 11: Meta Workforce and Hires Demographic Composition

	% black in workforce	% black in hires	% Hispanic in workforce	% Hispanic in hires	% female in workforce	% female in hires
2012	4.83%	4.60%	7.89%	8.62%	35.62%	33.91%
2013	4.65%	4.21%	8.70%	10.18%	35.08%	34.04%
2014	4.61%	4.37%	9.74%	11.43%	33.13%	30.56%
2015	4.44%	4.25%	11.70%	14.64%	33.86%	34.70%
2016	4.68%	5.08%	12.44%	13.47%	34.86%	35.25%
2017	5.25%	5.81%	12.31%	12.45%	36.52%	38.25%
2018	5.63%	6.25%	12.20%	12.14%	36.91%	37.29%
2019	5.72%	5.82%	12.24%	12.38%	36.81%	36.30%
2020	5.96%	6.57%	12.22%	12.15%	36.24%	35.62%
2021	6.47%	7.34%	12.47%	12.93%	37.22%	38.76%
2022	6.28%	6.03%	11.95%	11.04%	36.16%	34.64%
2023	6.24%	6.79%	11.54%	12.29%	35.36%	32.11%