NBER WORKING PAPER SERIES

FIRMING UP INEQUALITY

Jae Song
David J. Price
Fatih Guvenen
Nicholas Bloom

Working Paper 21199
http://www.nber.org/papers/w21199

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2015

For comments, we thank seminar and conference participants at Stanford University, the 2014 Winter Meeting of the American Economic Association, the 2014 Summer Meeting of the Econometric Society and the 2014 meeting of the Society for Economic Dynamics. To mitigate alphabetical inequality we have reversed the order of coauthors. We are grateful to Gerald Ray at the Social Security Administration for his help and support. We thank the National Science Foundation for financial support. The views expressed herein are those of the authors and do not necessarily reflect the views of the Social Security Administration or the National Bureau of Economic Research.

Firming Up Inequality
Jae Song, David J. Price, Fatih Guvenen, and Nicholas Bloom
NBER Working Paper No. 21199
May 2015
JEL No. E24,E25,J31,L23

ABSTRACT

Earnings inequality in the United States has increased rapidly over the last three decades, but little is known about the role of firms in this trend. For example, how much of the rise in earnings inequality can be attributed to rising dispersion between firms in the average wages they pay, and how much is due to rising wage dispersion among workers within firms? Similarly, how did rising inequality affect the wage earnings of different types of workers working for the same employer—men vs. women, young vs. old, new hires vs. senior employees, and so on? To address questions like these, we begin by constructing a matched employer-employee data set for the United States using administrative records. Covering all U.S. firms between 1978 to 2012, we show that virtually all of the rise in earnings dispersion between workers is accounted for by increasing dispersion in average wages paid by the employers of these individuals. In contrast, pay differences within employers have remained virtually unchanged, a finding that is robust across industries, geographical regions, and firm size groups. Furthermore, the wage gap between the most highly paid employees within these firms (CEOs and high level executives) and the average employee has increased only by a small amount, refuting oft-made claims that such widening gaps account for a large fraction of rising inequality in the population.

Jae Song
Social Security Administration
Office of Disability Adjudication
and Review
5107 Leesburg Pike, Suite 1400
Falls Church, VA 22041
jae.song@ssa.gov

David J. Price
Stanford University
Department of Economics
579 Serra Mall
Stanford, CA 94305
djprice@stanford.edu

Fatih Guvenen
Department of Economics
University of Minnesota
4-101 Hanson Hall
1925 Fourth Street South
Minneapolis, MN, 55455
and NBER
guvenen@umn.edu

Nicholas Bloom
Stanford University
Department of Economics
579 Serra Mall
Stanford, CA 94305-6072
and NBER
nbloom@stanford.edu

1 Introduction

The dramatic rise in U.S. wage inequality since the 1970s has been well documented. An enormous body of theoretical and empirical research has been conducted over the past two decades in an attempt to understand the causes of this trend.[1] While much has been learned from these analyses, several major questions remain unanswered. An important set of open questions concerns the link between wage inequality on the worker side to trends in the behavior of the firms and industries that employ these workers. A major difficulty with studying questions of this sort has been the lack of a comprehensive, matched employer-employee data set in the United States that covers the period of rising inequality, beginning with the 1970s.

In the absence of comprehensive evidence on wages paid by firms, it is frequently asserted that inequality within the firm is a driving force leading to an increase in overall inequality. For example, according to Mishel and Sabadish (2014), "a key driver of wage inequality is the growth of chief executive officer earnings and compensation." Piketty (2013) (p. 315) agrees, noting that "the primary reason for increased income inequality in recent decades is the rise of the supermanager." And he adds (p. 332) that "wage inequalities increased rapidly in the United States and Britain because U.S. and British corporations became much more tolerant of extremely generous pay packages after 1970."

To help address these questions, we use data on wage earnings for a one-sixteenth percent representative sample of U.S. workers, and the (100 percent) population of U.S. firms, between 1978 and 2012. Wage earnings in this data set have no top-coding, which allows us to study individuals at the top of the earnings distributions. Because it is based on administrative records, there is little measurement error, which is a pervasive problem in survey-based data. Additionally, because of the large sample size, we are able to obtain precise estimates using nonparametric methods.

Contrary to the assertions made by Mishel and Sabadish (2014), Piketty (2013), and others, we find strong evidence that *within-firm* pay inequality has remained

[1] For the basic facts about these trends, see, among others, Juhn et al. (1993a) and Autor et al. (2008), and for a survey of theoretical work, see, e.g., Acemoglu (2002) and Acemoglu and Autor (2011).

mostly flat over the past three decades. Between 1982 and 2012, the middle of the income distribution saw an increase in real wages of 18 log points (20 percent), while the top one percent saw an increase of 66 log points (94 percent). This change is roughly mirrored in their firms: individuals in the middle of the income distribution worked at firms with mean real wages 23 log points (25 percent) higher in 2012 than in 1982, but individuals in the top one percent worked at firms with mean real wages 72 log points (105 percent) higher. If we calculate the increase in individual inequality during that time period as the difference between the change at the top end with the change at the middle—a 48 log point difference—then virtually all of that increasing individual inequality is explained by the 49 log point difference between the firms of individuals at the top, versus firms of individuals in the middle. These trends are consistent across regions and industries, remain true when restricting by sex, age, and tenure, and are robust to various changes to the sample selection criteria.

There are several potential explanations for these findings. One possibility is increased sorting: that is, perhaps, in the 1980s firms were employing workers from a broader set of skill levels but have become increasingly specialized over time, so that now firms employ workers from narrower skills groups. Therefore, some firms pay much higher average wages than before because their average worker quality has increased. And vice versa for firms that are now paying lower than before.[2]

A second potential explanation (which is not necessarily mutually exclusive with the first one) is growing productivity differentials across firms. If the production technology delivers positive assortative matching and workers are mobile and then higher skill workers will flock into higher productivity firms (and vice versa for low productivity firms and lower skill workers). Therefore, increased productivity differences could trigger increased sorting. However, this productivity differential channel does not require sorting to work. If instead workers have strong attachments to their firms (perhaps due to firm-specific human capital), then workers will not reallocate across firms and their wage will reflect the diverging productivity levels across firms.

The evidence established above does not directly distinguish between these differ-

[2]Although this is a plausible hypothesis that can explain *some* of the rise in between-firm inequality, it seems that for this channel to generate the bulk of the rise it would require a substantial reorganization of firms during this period.

ent hypotheses. It is possible to use a regression framework to distinguish between these hypotheses, along the lines of Card et al. (2013), which is part of our ongoing research. (The paper will be updated with these new results once available).

More broadly, however, our results stress that any explanation for rising inequality must take into account an understanding of the nature of the firm and the economic motivations that led to its boundaries. They also suggest an explanation for why many do not feel that there has been an increase in inequality:[3] on average, individuals' inequality with their coworkers has changed little over the past three decades.

Several recent studies have attempted to answer similar questions. Abowd et al. (1999) employ a regression framework that allows them to disentangle firm effects, worker effects, and the nature of sorting by studying a longitudinal panel of French workers and firms. They find that firm effects, while important, are less important than individual effects. Card et al. (2013) use a similar technique to analyze a question more similar to ours: Using a matched employer-employee panel data set from West Germany, they find that increasing inequality is approximately equally explained by increased heterogeneity between workers, increasing heterogeneity between establishments, and increasing assortative matches between the two. Mueller et al. (2015) relate rising inequality to firm growth in the United Kingdom, finding that wages for high-skill jobs are diverging from wages for other jobs more at large firms than smaller firms, while the differential between wages for medium- and low-skill jobs is mostly unrelated to firm size. They also find evidence that rising inequality in developed countries may be driven by an increase in size of the largest firms.

Dunne et al. (2004) was the first paper to draw attention to the fact that rising inequality among workers was closely mirrored in rising inequality among (plants) establishments. However, these authors lacked data on wages within firms, which limited the scope of their analysis to between-firm data. Faggio et al. (2007) also found a similar link between rising worker and firm inequality in a sample of UK firms, particularly in the service sector, but again lacked a matched worker-firm database.

Closest to our work, Barth et al. (2014) use the Longitudinal Employer-Household

[3]One indicator of this is that the rise in inequality had mostly occurred by the early 2000s while the popular press did not focus on this unless the late 2000s , after the increased availability of data on income inequality.

Dynamics data spanning 1992 to 2007 as a source of U.S. employer-employee matched data. They also find a large share (about 2/3 in their analysis) of the rise in earnings inequality can be attributed to the rise in between-establishment inequality, but unlike our work they find an important role for cross-industry variation. One key difference between this study and our paper is that we focus on the population of firms as the measure of employer—identified by Employer Identification Numbers (EINs)—as opposed to establishments (plants) in these studies. This is important as it allows us to study a variety of questions about pay structure within the firm, including in corporate headquarters. Second, our data is not top-coded which allows us to examine CEOs and other executive pay, since we can look at divisions up to the top 99.99%. Finally, Barth et al. (2014)'s LEHD data analysis covers the 16 year period 1992–2007 while our data set spans 35 years from 1978 to 2012, including the Great Recession.

The paper is organized as follows. Section 2 describes the data set and the construction of the matched employer-employee data set, presents summary statistics from the sample, and discusses the methodology. Section 3 presents the main results and Section 4 concludes.

2 Empirical Analysis

2.1 Data and Sample Selection

The main source of data used in this paper is the confidential Master Earnings File (MEF), which is compiled and maintained by the U.S. Social Security Administration (SSA). The MEF has previously been used in Guvenen et al. (2014b), which contains a more detailed description of the data set as well as the steps of the sample selection. Therefore, here we provide a brief overview and refer the reader to that paper for more details.

The MEF contains a separate line of record for every individual that has ever been issued a Social Security Number. In addition to basic demographic information (sex, race, place of birth, date of birth, etc.), the MEF contains labor earnings information for every year from 1978 to 2012. Earnings data in the MEF is based on Box 1 of Form

W-2, which is sent directly from employers to the SSA. Data from Box 1 is uncapped, and includes wages and salaries, bonuses, exercised stock options, the dollar value of vested restricted stock units, and other sources of income.

The Matched Employer-Employee Dataset

Because earnings data are based on the W-2 form, the data set includes one record for each individual, for each firm they worked for in each year. Crucially for our purposes, the MEF also contains a unique employer identification number (EIN) for each W-2 earnings record. Because MEF is a population sample and has EIN records for each job of each worker, we can use worker side information to construct firm-level variables. In particular, we assign all workers who received wage earnings from the same EIN in a given year to that firm. The fact that a worker can hold multiple jobs and/or can transition from one job to another in a given year creates some complications in this assignment procedure, which we deal with as explained in Appendix A. The resulting matched employer-employee data set contains information on the wage distribution within each firm as well as the distributions of workers by gender, age, job tenure, as well as the total employment and wage bill by firm. We analyze the full sample of firms in this matched data set. In the baseline sample we restrict attention to firms with at least 10 full-time equivalent (FTE) employees in a given year, and if they are not in the Educational Services or Public Administration industries (although results are robust to relaxing these restrictions). We conduct robustness analyses with different cutoff levels (from 1+ FTE to 1000+ FTE).

Turning to workers, it is both challenging to analyze the universe of all workers given the substantial sample size and it is not necessary given that the results are unlikely to change if we were to work with a subsample. Therefore, we select a one-sixteenth representative sample of individuals. We select individuals into our base sample if they have relatively strong labor market attachment, defined by earning at least the equivalent of 40 hours per week for 13 weeks at that year's minimum wage. Furthermore, an individual is only included in our sample if his/her employer is in the sample. All wage earnings observations are capped (Winsorized) at the 99.999th percentile, and all dollar values are adjusted for inflation using the Personal

Consumption Expenditures (PCE) price index. Further details on how we process the data are available in Appendix A.

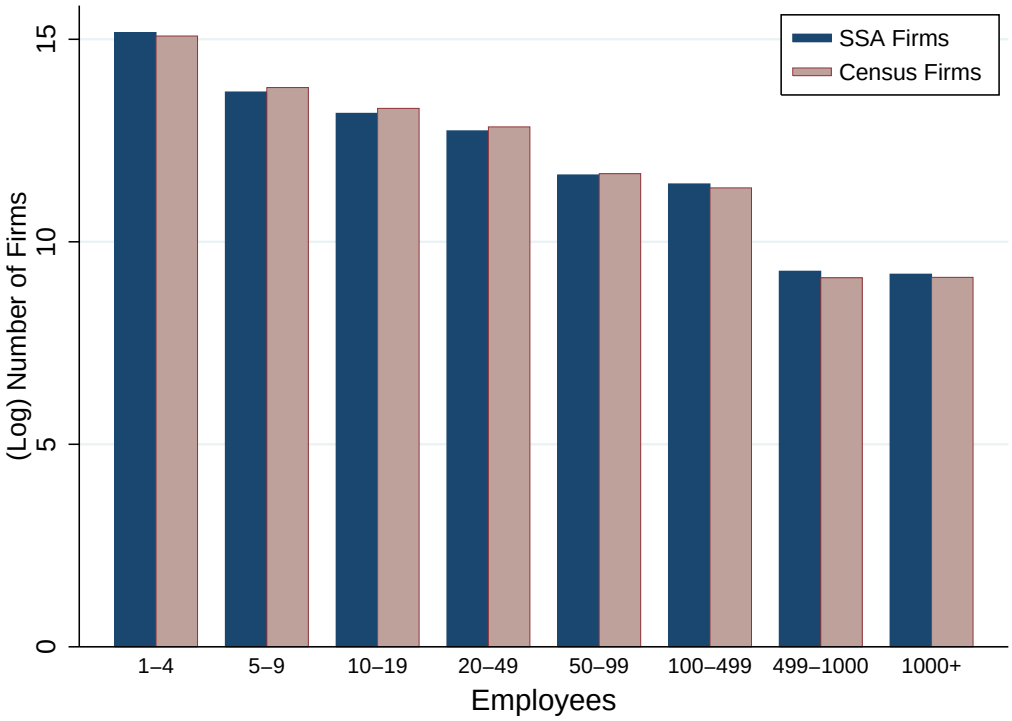
2.2 What is a Firm?

Throughout the paper, we use Employer Identification Numbers (EINs) as the boundary of a firm. Although there is no precise economic definition of a firm, the EIN corresponds closely to what many economists consider to be the firm's boundaries. All employers must have an EIN, and many firms use only one EIN. Corporations, for example, should have different EINs for each subsidiary, but use the same EIN for all divisions.[4] Because of this, Walmart stores has only one EIN but employs over 1 million people in more than 4,000 locations. One alternative measure of the boundaries of the firm, as used by Barth et al. (2014), is the establishment which is a physically distinct business location (such as an individual Walmart store). Establishments are valuable to study, but the drivers of wage decisions may be more frequently made at a higher level—for example, Walmart has a national salary policy with very limited regional variation.

Indeed, the Bureau of Labor Statistics, which also uses EINs to define the boundaries of the firm, notes that "the firm level is more consistent with the role of corporations as the economic decision makers than each individual establishment".[5] In discussing small businesses, the Census Bureau—which uses EINs, as well as other information not available in our data set, to define firms—also notes that "most scholars prefer to define small business in terms of the size of the entire company or firm, not individual establishments".[6] Figure 1 shows that the number of firms of various sizes in the SSA data set roughly correspond to the number of firms in the Census data.

[4] See IRS Publication 1635, "Understanding Your EIN."
[5] http://www.bls.gov/bdm/sizeclassqanda.htm#q3
[6] http://www.census.gov/econ/smallbus.html

FIGURE 1 – Number of Firms, by Size



Notes: Natural log of the number of firms in each size category are shown. Census firm data is from http://www2.census.gov/econ/susb/data/2012/us_state_naicssector_small_emplsize_2012.xls and http://www2.census.gov/econ/susb/data/2012/us_state_naicssector_large_emplsize_2012.xls. Census numbers count the number of employees at a point in time, while the SSA numbers count the number of FTEs over the course of a year.

2.3 Summary Statistics

We begin by providing some broad statistics on the MEF data to see how it aligns with aggregate measures from NIPAs and the BLS. First, aggregating wages and salaries from W-2 records over all individuals in the MEF yields a total wage bill of $6.8 trillion dollars in 2012. The corresponding figure from NIPAs is $6.9 trillion, which is quite close.[7] In fact, as shown in Figure 2, the two series track each other closely

[7]For this particular statistic, it is not obvious that the NIPA measure is the more accurate one. This is because the NIPA statistic comes from BLS's Quarterly Census of Employment and Wages (QCEW) and although this is a comprehensive survey of employers, it covers about 97% of employment in the US. Most notably, the QCEW excludes employment that is not covered by the unemployment insurance system—most agricultural workers on small farms, all members of the

year by year.[8]

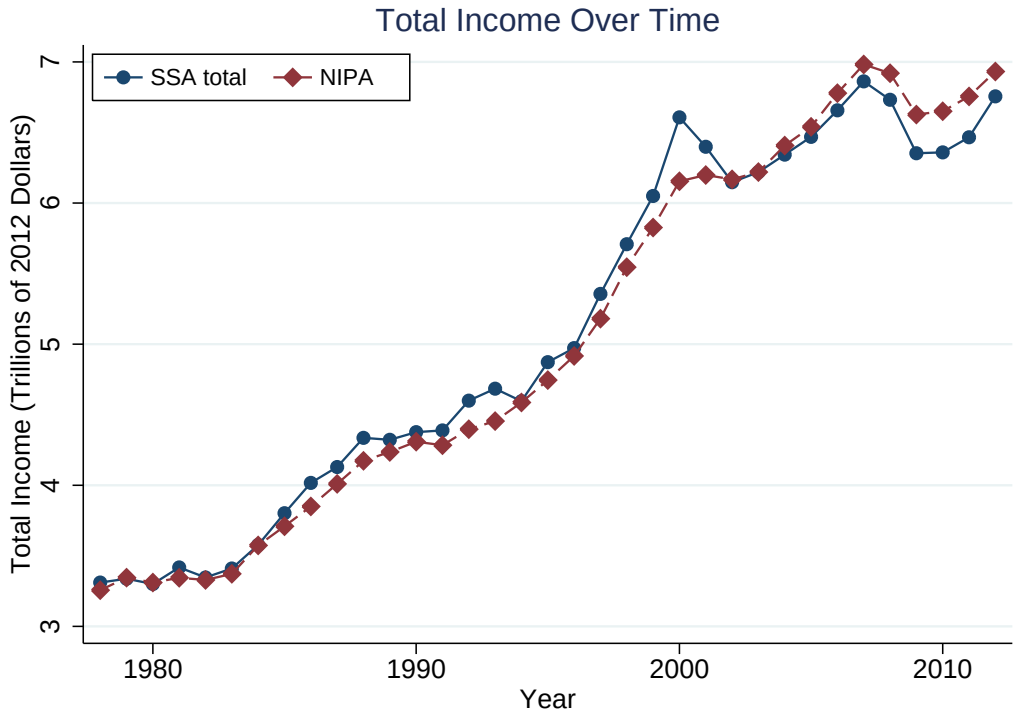
Second, the total number of individuals in the MEF who received W-2 income in a given year (our measure of total employment) also closely tracks total employment in the Current Population Survey (CPS). In 2012, for example, the MEF measure contains 153 million workers, while the CPS indicated that, on average, 142 million individuals were employed at any given time. The difference may be largely due to the fact that the CPS is a point-in-time estimate; if people cycle in and out of employment, they may be missed in the CPS data but will be included in the MEF (which is aggregate over the year). It is also possible that individuals who earned a very small amount of money in a given year may not report themselves as being employed in the CPS even if they did receive a W-2. Figure 3 shows total employment in the MEF and CPS; these two series generally track each other well over time.

Third, there are 6.1 million unique EINs in the MEF in 2012; as discussed above, we are assuming that each EIN represents a firm. This number is slightly higher than the 5.7 million firms identified by the Census Bureau's Statistics of U.S. Businesses data set. In addition, as shown in Figure 4, the trends in each of these data sets are similar over time (at least since 1988, when the Census data begins). There are somewhat more firms in the SSA data than in the Census data in every year. This may be due to the fact that the Census is able to aggregate some EINs up to the firms that contain them, or because they may not count firms with very small total wage bills.

Armed Forces, elected officials in most states, most employees of railroads, some domestic workers, most student workers at schools, and employees of certain small nonprofit organizations. All these employees will be included in the MEF as long as they are paid by W-2s. The definition of wage earnings used is essentially the same as ours, with the exception that employer contributions to 401K accounts are treated as wage income by some states and hence included in BLS measure for those states (but are not included in our W-2 measures). See http://www.bls.gov/cew/cewfaq.htm for details.

[8]One difference between the two measures comes from the recording of deferred compensation: the NIPA measure is based on payroll data where stock options and restricted stock units (RSUs) are reported in the year they are granted and at the current value (or the Black-Scholes value in the case of stock options). In contrast, the MEF records the W-2 values, where RSUs are reported at the time of vesting and stock options at the time of exercise—typically *several years after they are granted*—and at the price on the vesting/exercise day. This difference can be seen in the discrepancy in the late 1990s when the SSA measure exceeds the NIPA measure (due to the stock price boom); the opposite pattern is seen during the Great Recession for the same reason.

FIGURE 2 – Total wage bill

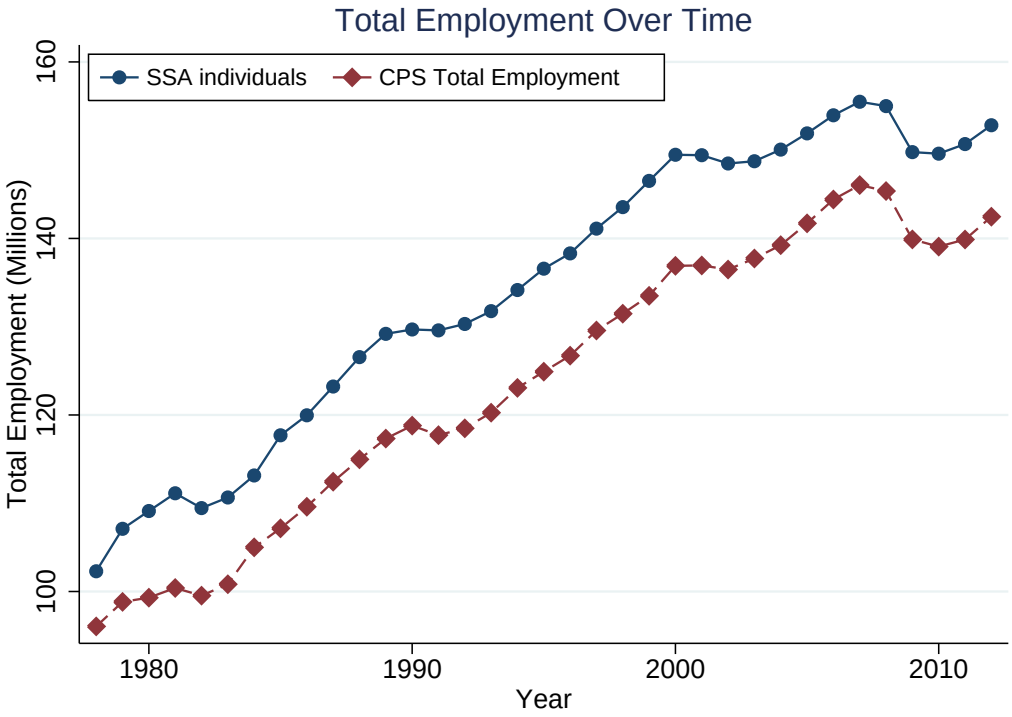


Notes: SSA data includes all entries in the MEF. National Income and Product Accounts (NIPA) data is from the St. Louis Federal Reserve Bank's FRED service, series A576RC1, "Compensation of Employees, Received: Wage and Salary Disbursements." All data are adjusted for inflation using the PCE price index.

Table I reports the aggregate numbers from the MEF and the benchmarks discussed above as well as totals from the base sample used in this paper.

The remaining statistics in this section are based on the sample of the MEF that we use for analysis in the rest of the section. In particular, this only includes full-time workers, and restricts the analysis to firms with at least 10 FTE that are not in public administration or educational services. The median individual in this sample earned $33,600 in 2012, up from $28,000 (in 2012 dollars) in 1982. The median firm in our 2012 sample contained 21 full-time equivalent employees (FTEs), but the median individual was at a firm with 983 FTEs. These statistics, as well as others describing the sample, are shown in Table II.

FIGURE 3 – Total employment



Notes: SSA data includes all entries in the MEF. Current Population Survey (CPS) total employment shows the yearly average of the monthly employment numbers in the CPS. This data is from the Bureau of Labor Statistics Table LNS12000000.

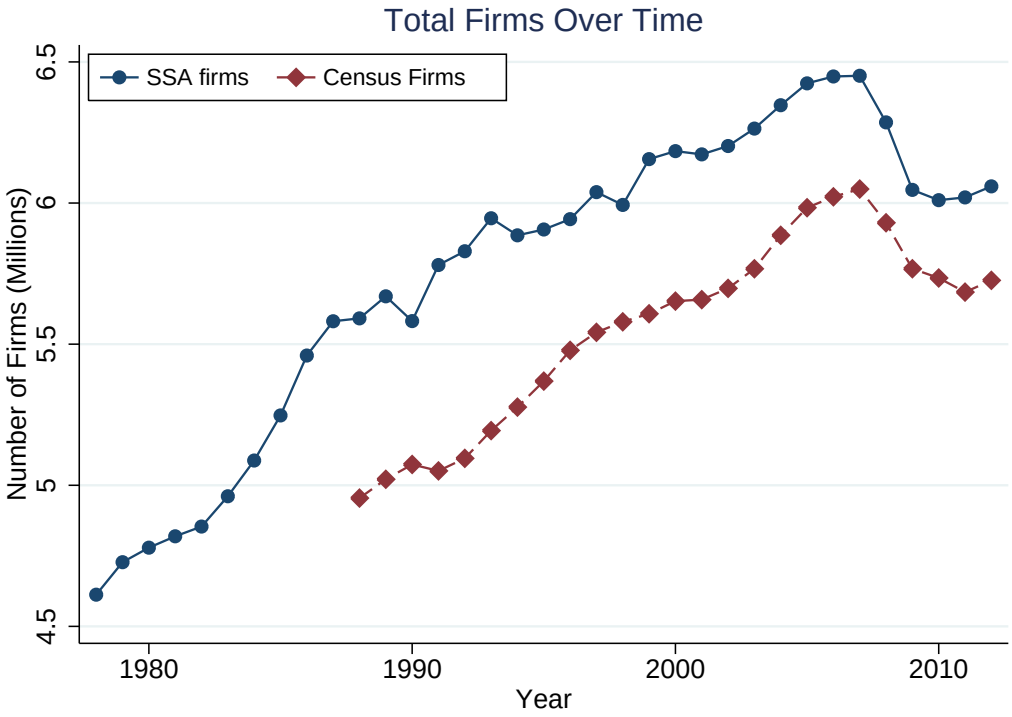
2.4 Empirical Method

We are now ready to state the main decomposition we are interested in. Let $w_t^{i,j}$ be the log wage of worker i employed by firm j in period t. Consider the identity:

$$w_t^{i,j} \equiv \overline{w}_t^A + \left[\overline{w}_t^j - \overline{w}_t^A\right] + \left[w_t^{i,j} - \overline{w}_t^j\right], \tag{1}$$

where $\overline{w}_t^A$ is the average wage earnings in the economy, and $\overline{w}_t^j$ is the average wage earnings paid by firm j. Let N_j denote total employment in firm j (in period t, which

FIGURE 4 – Total number of firms



Notes: SSA data includes all entries in the MEF. Census firms shows the total number of firms reported by the Census Bureau's Statistics of U.S. Businesses data set, available at http://www.census.gov/econ/susb/historical_data.html.

is suppressed). Taking the variance of both sides yields:[9]

$$\sum_{j=1}^{J}\sum_{i}^{N_j}\left(w_t^{i,j}-\overline{w}_t^{A}\right)^2=\sum_{j=1}^{J}\sum_{i}^{N_j}\left[\overline{w}_t^{j}-\overline{w}_t^{A}\right]^2+\sum_{j=1}^{J}\sum_{i}^{N_j}\left[w_t^{i,j}-\overline{w}_t^{j}\right]^2, \tag{2}$$

$$\Rightarrow N\text{var}_i(w_t^{i,j})=\sum_{j=1}^{J}N_j\left[\overline{w}_t^{j}-\overline{w}_t^{A}\right]^2+\sum_{j=1}^{J}N_j\text{var}(w_t^{i,j}|i\in j) \tag{3}$$

Dividing both sides by N and letting $P_j=N_j/N$ denote the employment share of

[9]The covariance term that should appear on the right hand side of (2), $2\sum_{j=1}^{J}\sum_{i}^{N_j}\left[\overline{w}_t^{j}-\overline{w}_t^{A}\right]\left[w_t^{i,j}-\overline{w}_t^{j}\right]$, is zero by construction, and is hence omitted.

TABLE I – Total Values From the Data

Year	Statistic	Benchmark	MEF	Our Sample
1982	Number of Employees	99.5	109	66.6
1982	Number of Firms	.	4.85	.826
1982	Total Wage Bill	3.33	3.35	2.49
1992	Number of Employees	118	130	84.3
1992	Number of Firms	5.1	5.83	1.03
1992	Total Wage Bill	4.4	4.6	3.55
2012	Number of Employees	142	153	103
2012	Number of Firms	5.73	6.06	1.09
2012	Total Wage Bill	6.93	6.76	5.38

Notes: Number of employees benchmark is from the National Income and Product Accounts (NIPA) data from the St. Louis Federal Reserve Bank's FRED service, series A576RC1, "Compensation of Employees, Received: Wage and Salary Disbursements." These data are adjusted for inflation using the PCE price index. Total employment is from the Current Population Survey (CPS)'s yearly average of the monthly employment numbers; this data is from the Bureau of Labor Statistics Table LNS12000000. Census firms shows the total number of firms reported by the Census Bureau's Statistics of U.S. Businesses data set, available at http://www.census.gov/econ/susb/historical_data.html. MEF data includes all observations in the Master Earnings File. Sample statistics are from the subset of the MEF that we use in this paper. Restrictions are described above; this sample only includes full-time workers in firms with at least 10 FTE that are not in public administration or educational services. Sample values are multiplied by sixteen in order to be comparable to the population. Number of employees and number of firms are in millions; total wage bill is in trillions of 2012 dollars.

firm j, we can write

$$\text{var}_i(w_t^{i,j}) = \underbrace{\text{var}_j(\overline{w}_t^j)}_{\text{Between-firm dispersion}} + \sum_{j=1}^{J} P_j \times \underbrace{\text{var}_i(w_t^{i,j} | i \in j)}_{\text{Within-firm } j\text{dispersion}} . \quad (4)$$

This equation provides a simple way to decompose total wage dispersion in the economy into (i) between-firm dispersion in average wages paid by each firm, and (ii) a second component which is within-firm dispersion in pay weighted by employment share of each firm. Computing the terms in equation (4) for two different time periods and differencing then provides a decomposition for the change in total variance in terms of the changes in between-firm and within-firm dispersion terms. Below we are going to use several decompositions that are all based on this general idea. We will often study other measures of dispersion, such as percentile differentials, but the main

TABLE II – Percentiles of various statistics from the data

Year	Group	Statistic	25%ile	50%ile	75%ile
1982	Firm	FTE	13.5	20.3	39.7
1982	Firm	Mean Wage /$1,000	17.4	26.2	37.5
1982	Firm	Total Wage /$1,000	317	577	1236
1982	Indiv.	Age	25	34	46
1982	Indiv.	FTE at firm	71.2	659	8758
1982	Indiv.	Total Wage /$1,000	14.7	28.0	47.8
1982	Indiv.	Wage/Firm Avg	0.54	0.84	1.20
2012	Firm	FTE	13.6	20.8	41.7
2012	Firm	Mean Wage /$1,000	21.0	33.9	51.3
2012	Firm	Total Wage /$1,000	394	758	1805
2012	Indiv.	Age	29	41	52
2012	Indiv.	FTE at firm	99.1	983	12630
2012	Indiv.	Total Wage /$1,000	17.1	33.6	59.9
2012	Indiv.	Wage/Firm Avg	0.50	0.80	1.18

Notes: Values indicate various percentiles for the data for individuals or firms. All dollar values are in thousands and are adjusted for inflation using the PCE price index. Only firms and individuals in firms with at least 10 full-time equivalent employees are included. Firm statistics are based on mean wage at firms and are *not* weighted by number of employees. Only employed individuals are included in all statistics, where employed is defined as earning the equivalent of minimum wage for 40 hours per week in 13 weeks. Individuals and firms in public administration or educational services are not included.

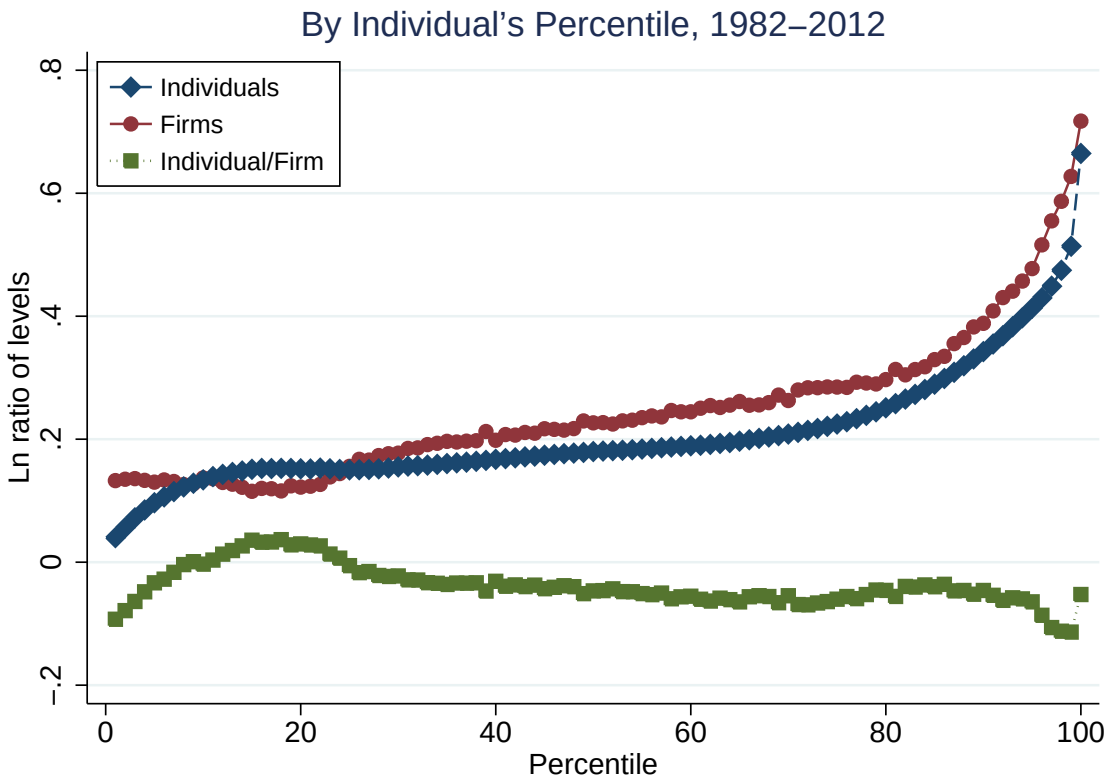
idea is the same.

A Graphical Construct for Empirical Analysis

We present the results of our analysis in the forms of graphs that can communicate a lot information in an effective way. To this end, we shall focus on the percentiles or quantiles of various distributions. Three key variables defined above will appear in much of the analysis, and they are: a given individual's wage earnings, $w_t^{i,j}$; average wage earnings paid by a firm, $\overline{w}_t^j$; and finally the difference between the two variables, $w_t^{i,j} - \overline{w}_t^j$, which is, loosely speaking, the residual earnings of worker i relative to his/her firm average wage.

It is instructive to refer to Figure 5 to explain how each line is constructed. We

FIGURE 5 – High paid individuals now work at higher-paying firms, but are not higher paid relative to their firms.



Notes: For each percentile, statistics are based on individuals in that percentile of income in each year: their firms' average incomes and their own incomes. All values are adjusted for inflation using the PCE price index. Only firms and individuals in firms with at least 10 full-time equivalent employees are included. Firm statistics are based on mean wage at firms and are weighted by number of employees. Individual/firm is based on the individual's income as a fraction of firm mean income. Only employed individuals are included in all statistics, where employed is defined as earning the equivalent of minimum wage for 40 hours per week in 13 weeks. Individuals and firms in public administration or educational services are not included.

first group all individuals who satisfy the sample selection criteria described in Section 2.1 into percentile bins on the basis of their income in 1982.[10] Then we calculate the average of log real wages (in 2012 dollars) for each percentile bin. Let $P_t xy$ denote this average for percentile bin xy in year t. We then repeat the same procedure for

[10] Although we have data going back to 1978, we start our analyses in 1982, except where otherwise indicated, because of some minor data completeness issues in the first two years, and in order to use data points separated by 30 years. Results are very similar if we vary the starting point for these graphs.

2012 (now for individuals who satisfy sample selection in that year). The blue line marked with diamonds (labeled "Individuals") in Figure 5 plots $P_{2012}xy - P_{1982}xy$ for all percentile groups $xy = 1, 2, ..., 99, 100$ against the percentile number xy on the horizontal axis. For example, we calculate $P_{1982}50 = 10.23$ (corresponding to about \$27,700), whereas the comparable number in 2012 was $P_{2012}50 = 10.41$ (corresponding to about \$33,200). The difference of 0.18 (or 18 log points) is plotted on the graph at the 50th percentile. Note that this measure does not use any of the panel structure of the data; individuals in the 50th percentile in 1982 are almost certainly different from those in the 50th percentile in 2012.

For the red line marked with circles (labeled "Firms"), we put individuals into percentile bins based on their own wage earnings in 1982—just as we did for the "Individuals" line above—but for each percentile bin, we calculate the average of the log of mean real wages at each individual's *employer (or firm)*. We repeat the same procedure for 2012. For example, in 1982, individuals in the 50th percentile of individual wages were employed in firms with average log mean real wages of 10.42 (corresponding to about \$33,600); in 2012, individuals at the 50th percentile were in firms with an average log mean real wage of 10.65 (corresponding to about \$42,200). The difference of 0.23 is plotted on the graph at the 50th percentile.

Finally, the green line marked with squares (labeled "Individual/Firm") is based on the residual wage measure, $w_t^{i,j} - \overline{w}_t^j$. Specifically, we compute the average of $w_t^{i,j} - \overline{w}_t^j$ across all workers within a percentile in each year.[11] We then plot the difference between this statistic in 1982 and 2012. For example, in 1982, individuals in the 50th percentile of individual wages had average log wages 0.19 lower than their firms' mean wages (corresponding to about 82% of their firms' mean wage). In 2012, individuals in the 50th percentile had average log wages 0.24 lower than their firms' mean wages (corresponding to about 79% of their firms' mean wages). We plot the difference of –0.05 at the 50th percentile. Note that this "Individual/Firm" line will be mechanically equal to the difference between the "Individuals" line and the "Firms" line.

For all these graphs, results should be interpreted similarly. A flat line indicates

[11] Notice that in all likelihood, the workers we average over are employed in different firms and each residual is computed with respect to a worker's own employer.

that inequality for that statistic has not changed over the time period, because the statistic for those at the top and the bottom have changed by the same amount. An upward-sloping line indicates that inequality has increased, because the statistic for those at the top has increased more than the statistic for those at the bottom; and by the same logic, a downward-sloping line indicates that inequality has decreased. This graphical construct thus allows us to detect changes in inequality that might be confined to one part of the wage distribution and may not be very visible in broad inequality statistics.

Particular care should be given to the interpretation of the green "Individual/Firm" line. The *level* of this line indicates the extent to which a particular demographic group gains or loses relative to the firm average. When we examine the whole population, or subsets of the population that, for each firm, include either everyone or no one at that firm, the green line's weighted average level must be near zero (except for small differences due to, for example, Jensen's inequality and data for individuals who work at multiple jobs). However, the interpretation is different when we look at demographic subsets of the population, as in Subsection 3.2, where we examine only a subset of each firm. For these analyses, there is no presumption that the level for any group must have an average of zero; instead, we interpret the average level as the extent to which the group has gained or lost relative to firm average. Next, in addition to level, the *slope* of the green line indicates the extent to which inequality has increased (if there is an upward slope) or decreased (downward slope) for the specified subset of the population.

3 Results

Except at the very top, the highest-paid individuals now work at higher-paying firms, but are not higher paid relative to those firms

We are now ready to discuss the first (and main) substantive finding of this paper, shown in Figure 5. The blue "individuals" line shows the well-documented trend: inequality in individuals' wage earnings has increased markedly between 1982 and 2012. For example, wages in the 50th percentile increased by 18 log points in those

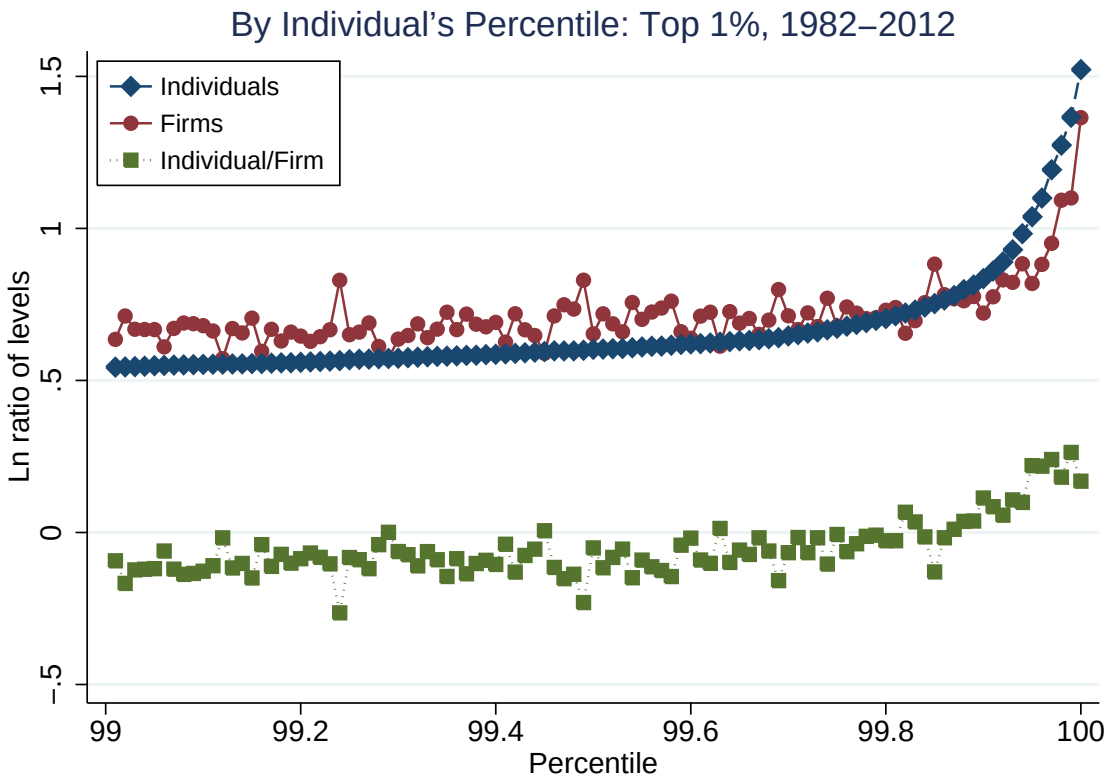
three decades, while wages in the top one percent increased by an average of 66 log points. The red line shows a similar trend for the firms in which these employees work. Mean wages at the firms that workers in the 50th percentile worked in increased by an average of 23 log points, while mean wages at the firms that the top one percent worked in increased by an average of 72 log points. This indicates that 101 percent ($= \frac{72-23}{66-18}$) of the increasing inequality for individuals is explained by rising inequality among their firms. Meanwhile, as shown by the green "individual/firm" line, individuals throughout the income distribution are faring similarly, relative to their firms, in 2012 as they did in 1982. Individual incomes as a fraction of mean firm incomes for those in the 50th percentile decreased by 5 log points. Income as a fraction of the firm's mean income for individuals in the top one percent also decreased by the same 5 log points.

As shown in Figure 6, the increase in inequality for the top 1% of incomes is much greater—consistent with findings from Piketty and Saez (2003) and Guvenen et al. (2014a) who point out the increasing inequality at the top of the earnings distribution. But, again most of this is accounted for by differences in firm level incomes. Individuals in the top 0.01 percent in 2012 are earning 152 log points more than in 1982. They are also at much higher-paying firms: mean income at their firms is 136 log points higher in 2012 than the firms of the top .01 percent in 1982. These individuals therefore now earn 17 log points more, relative to their firms, than the top .01 percent did in 1982.

Wage dispersion between firms is increasing, while dispersion within firms has been stable

Figure 7 conveys a similar story but using a slightly different calculation. The blue "Individuals" line is exactly the same as in Figure 5. However, the red "Firms" line now just ranks firms into (weighted) percentiles by the average pay of each firm, and plots the average change of log mean real wage in each percentile. So, for example, the 50th percentile change of 0.23 shows that the median wage firm in 2012 pays 23 more log-points than then median-wage firm in 1982. Finally, the green "Individual/Firm" line first calculates the ratio of individual income to firm mean income for each individual,

FIGURE 6 – At the very top, individuals are paid more with respect to their firms



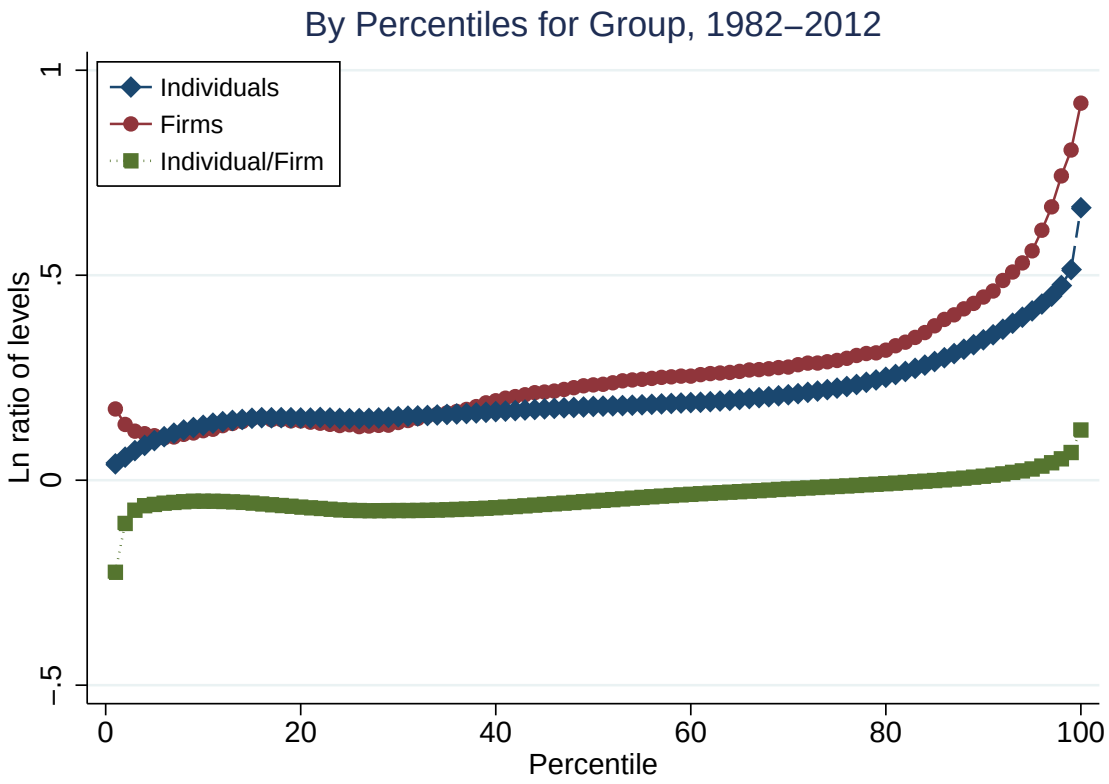
Notes: See the notes for Figure 5.

and arranges individuals into percentiles on the basis of that statistic. In other words, the red line shows the change in inequality between firms, without regard to where their employees are in the income distribution; while the green line shows change in inequality within firms, without regard to where those employees are in the absolute distribution. We find very similar results—almost the entirety of the increase in individual inequality can be accounted for the by the increase in cross-firm inequality.

Results are consistent within sub-periods

Figure 8 shows how the results from Figure 5 vary across time. Data values for each graph show what values the indicated percentiles would take in Figure 5, if that graph still began in 1982 but ended in the indicated year. Despite some variations

FIGURE 7 – Increasing inequality for individuals is mirrored by inequality between firms, but not for individuals within firms

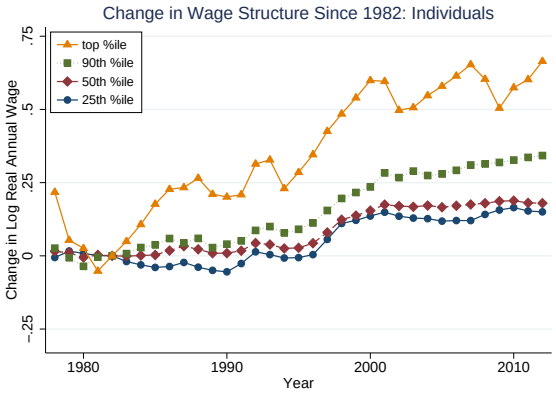


Notes: For each percentile, change in the natural log of that quantile of the parameter is shown. All values are adjusted for inflation using the PCE price index. Only firms and individuals in firms with at least 10 full-time equivalent employees are included. Firm statistics are based on mean wage at firms and are weighted by number of employees. Individual/firm is based on the individual's income as a fraction of firm mean income. Only full-time workers are included in all statistics, where full-time is defined as earning the equivalent of minimum wage for 40 hours per week in 13 weeks. Individuals and firms in public administration or educational services are not included.

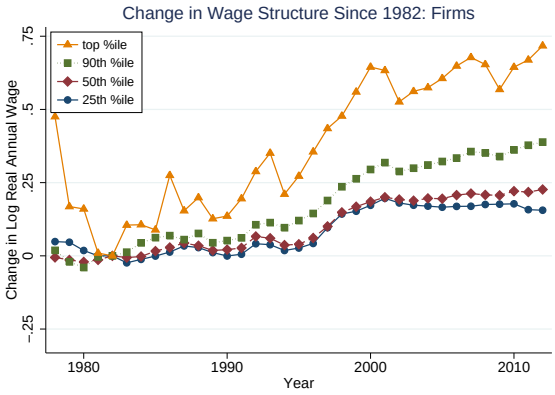
due to business cycles and other factors, Figure 8a shows that wage dispersion for individuals has been gradually increasing over time, as the higher percentiles steadily increase faster than the lower percentiles. Similarly, Figure 8b shows that mean real wages at the firms of individuals at the top of the income distribution have increased rapidly over time, while wages at firms of individuals lower in the income distribution have increased less. Figure 9c, on the other hand, shows a starkly different picture for individuals as a fraction of their firms. This ratio has changed little over three decades

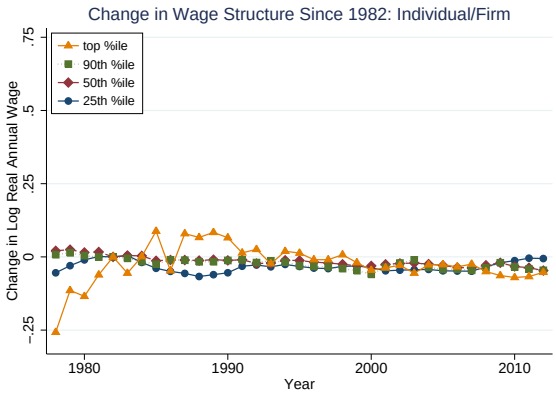
FIGURE 8 – Time-Series Variation in Total-, Between-Firm, and Within-Firm Wage Inequality

(A) Individuals



(B) Firm





(C) Individual/Firm

Note: For each percentile, statistics are based on individuals in that percentile of income in each year: their firms' average incomes and their own incomes. For other data notes, see 5.

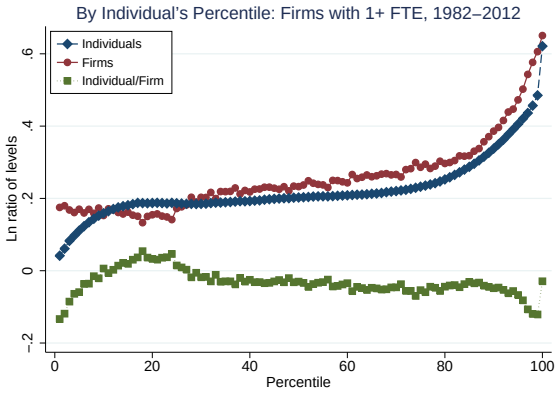
for individuals at the top, middle, or bottom of the income distribution. Throughout the last three decades, as the top-paid individuals are paid more, their coworkers have seen similar increases.

3.1 Inequality Patterns by Firm Type

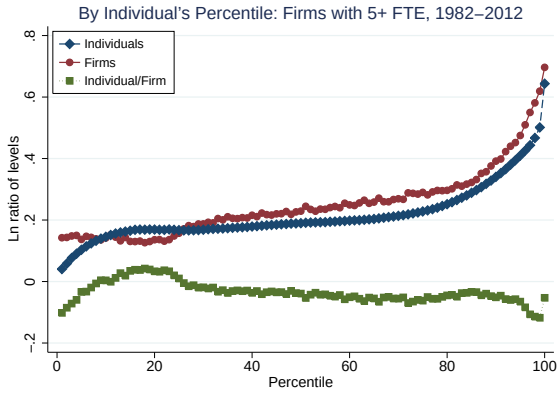
As noted earlier, the baseline analysis conducted so far is based on a sample of firms with at least 10 FTEs. Figure 9 shows that our results are not sensitive to changes in this threshold. When we look only at firms, and the individuals in firms, with at

FIGURE 9 – Trends are similar with different size thresholds

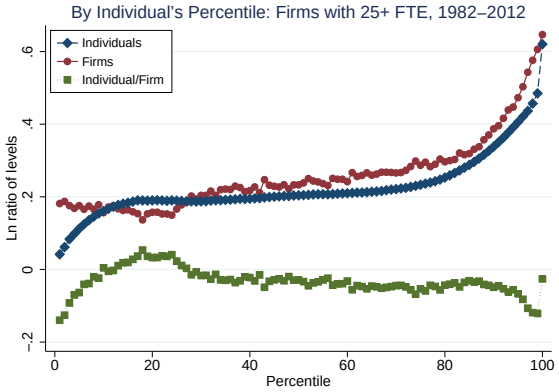
(A) 1+ FTE



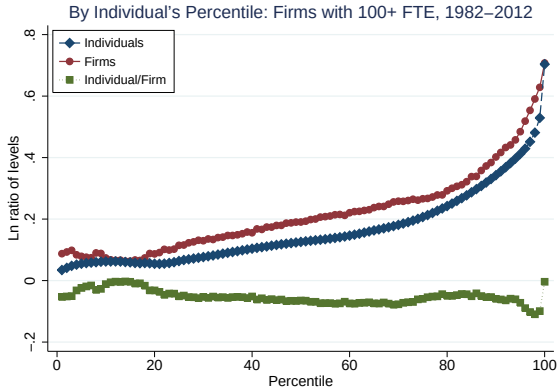
(B) 5+ FTE



(C) 25+ FTE



(D) 100+ FTE

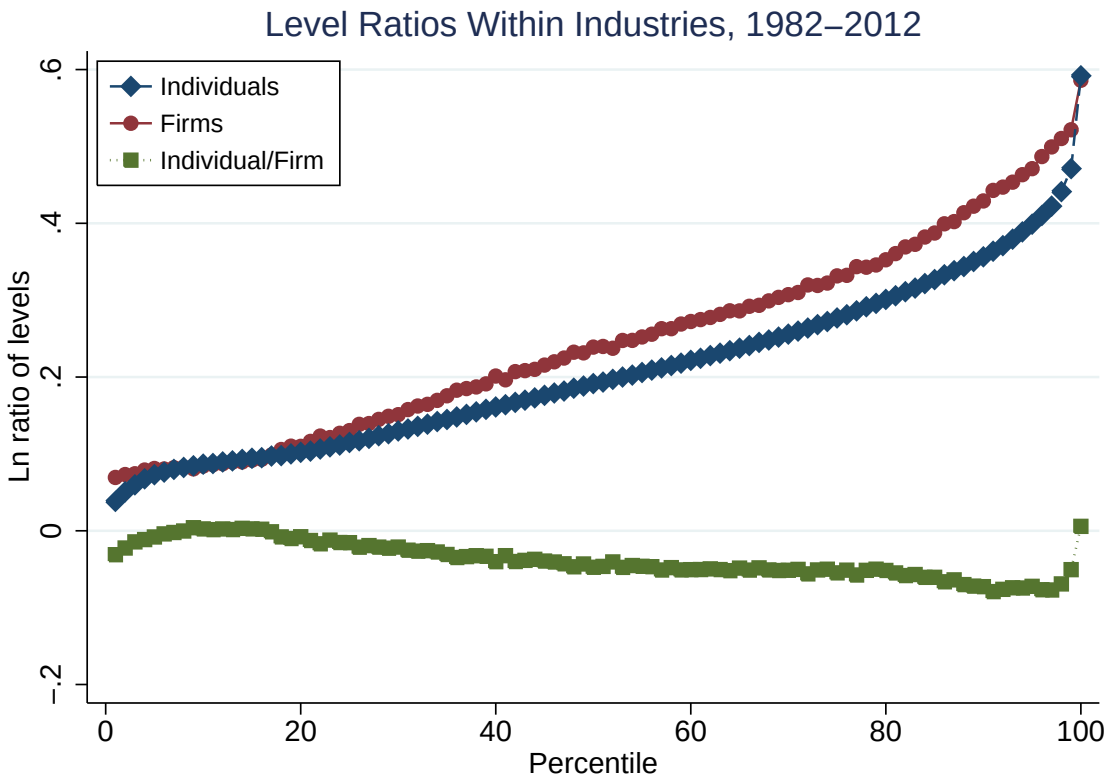


Notes: See the notes for Figure 5.

least 1 FTE, or at least 5, or 25, or 100 FTEs, we continue to see that workers at the top of the income distribution are at much better-paying firms in 2012 than they were in 1982, and earn a similar amount relative to their firms in the two years. For each of these size restrictions, between 89 percent and 103 percent of the change in individual income is explained by firms.

Figures 10 and 11 show that increasing dispersion between firms is not explained by increasing dispersion between industries. Figure 10 essentially creates one of the canonical graphs for each 4-digit SIC code, then averages the results. Thus, for example, individuals in the 50th percentile earn around the median wage for their

FIGURE 10 – Trends in Inequality, Controlling for Industry Fixed Effects



Notes: For each percentile, statistics are based on individuals in that percentile of income in each year: their firms' average incomes and their own incomes, within their 4-digit SIC industry in a given year. All values are adjusted for inflation using the PCE price index. Only firms and individuals in firms with at least 10 full-time equivalent employees are included. Firm statistics are based on mean wage at firms and are weighted by number of employees. Individual/firm is based on the individual's income as a fraction of firm mean income. Only full-time workers are included in all statistics, where full-time is defined as earning the equivalent of minimum wage for 40 hours per week in 13 weeks. Individuals and firms in public administration or educational services are not included.

industry. As before, we then calculate the average of individuals' wages within a percentile; average mean wages for the firms these individuals are at; and averages of the ratio between individual wages and mean firm wages for all individuals. As in other analyses, individuals who are well-paid relative to their industries are now at much higher-paying firms than individuals who are not as well compensated relative to their industries. Meanwhile, individuals at all values of income relative to their industries have incomes relative to their firms that are similar now to the ratios faced by comparable individuals three decades ago. Even within 4-digit industries, 87 percent of rising inequality is explained by firms.

Figure 11 instead restricts the sample to firms, and the individuals at firms, in a few broadly-defined industries. The amount of increasing inequality within each of these industries varies, but the same trend holds in each: individuals at the top of the income distribution, much more than those in the middle and bottom, are at firms with higher average wages in 2012 than top individuals were in 1982; and these high-paid individuals do not have higher wages relative to their firms now than they did three decades ago. Within these industries, firms explain between 34 percent of rising inequality (for Services) and 133 percent of rising inequality (for Utilities).

Figure 12 shows that similar trends hold in different geographies across the country. These graphs each restrict the sample to firms, and individuals at firms, with headquarters each of the four Census regions. The Northeast, South, Midwest, and West all show the same trends described above for the national sample, with between firms explaining between 74 percent of rising inequality (in the Midwest) and 112 percent (in the West).

3.2 Who Gains and Who Loses Within A Firm?

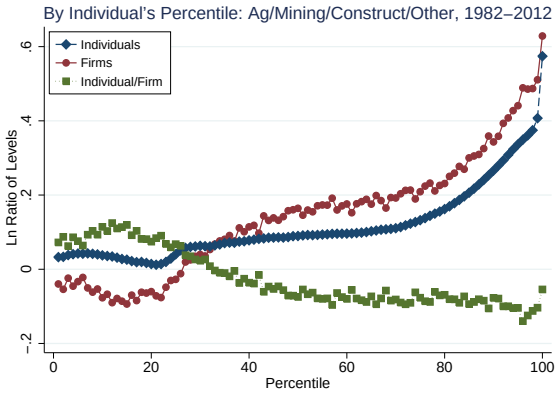
The fact that the overall earnings dispersion within firms did not increase does not mean that the pay structure within firms remained static during this period. It is entirely possible that the pay of workers of different types (by age, gender, tenure, etc.) shift relative to each other, even when the overall level of dispersion remains more or less unchanged. To understand whether there might have been such changes in the within-firm pay structure, we now investigate how the wages of workers with different observable characteristics changed inside a firm.

Gender Structure within A Firm

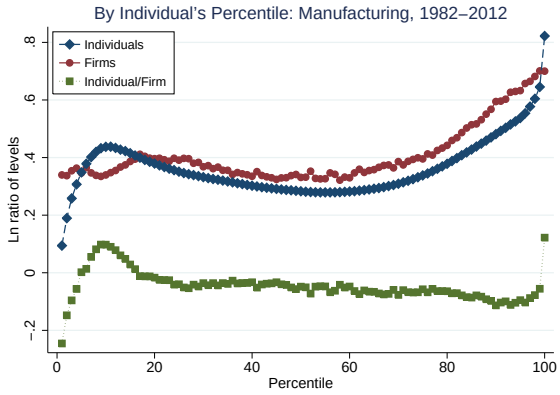
First, we examine the gender structure of pay within a firm. It is well documented that the gender gap shrank significantly, especially during the 1980s, was relatively stagnant in the 1990s, and shrank further in the 2000s. However, this fact alone is not sufficient to know what happened within firms. It is possible, for example, that female employment grew especially strongly in firms and industries (such as services)

FIGURE 11 – Trends in Inequality, Within Selected Industries

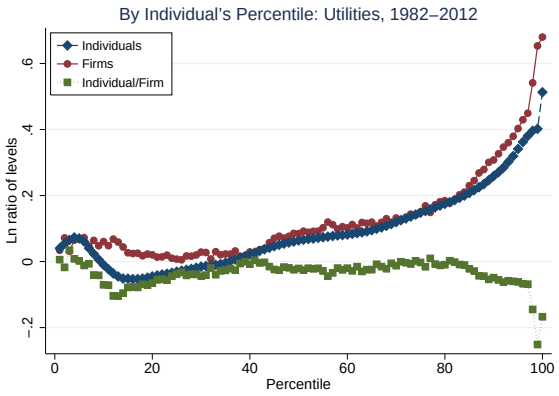
(A) SIC: 0100 to 1799 and 9900 to 9999



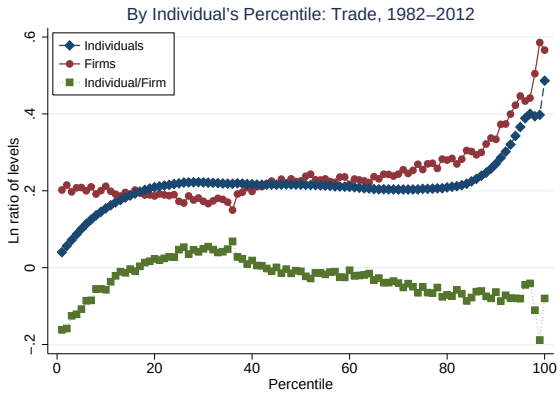
(B) SIC: 2000 to 3999



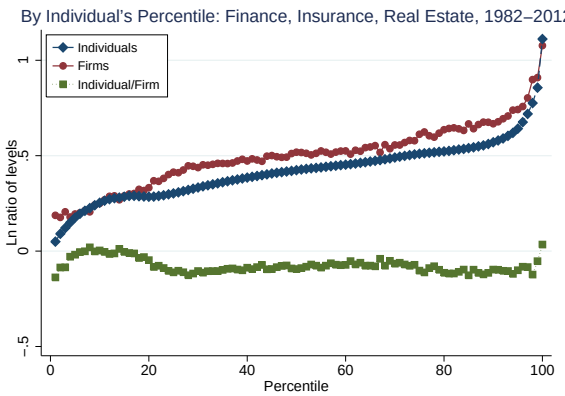
(C) SIC: 4000 to 4999



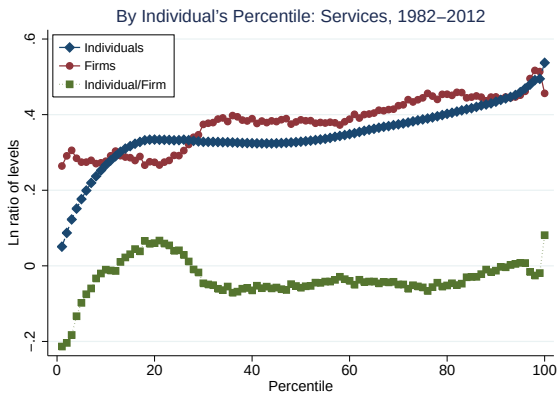
(D) SIC: 5000 to 5999



(E) SIC: 6000 to 6799



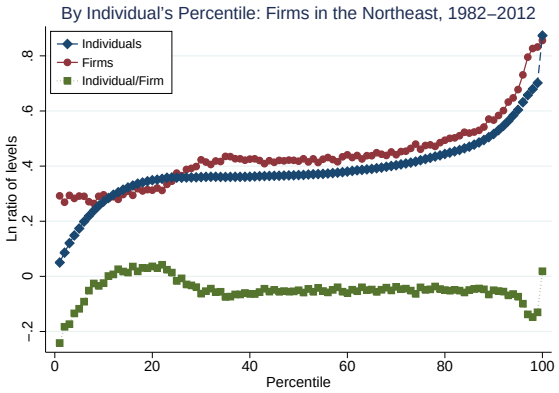
(F) SIC: 7000 to 8999



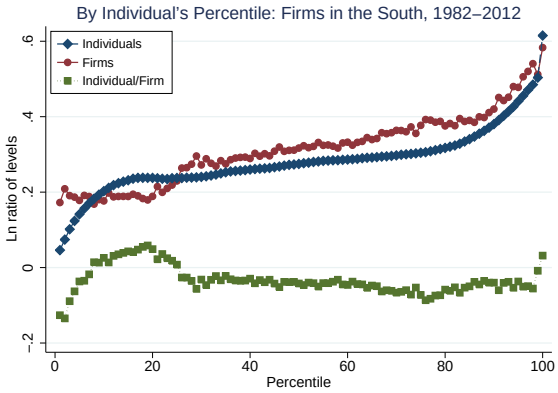
Notes: See the notes for Figure 5.

FIGURE 12 – Trends in Inequality, By US Geographical Regions

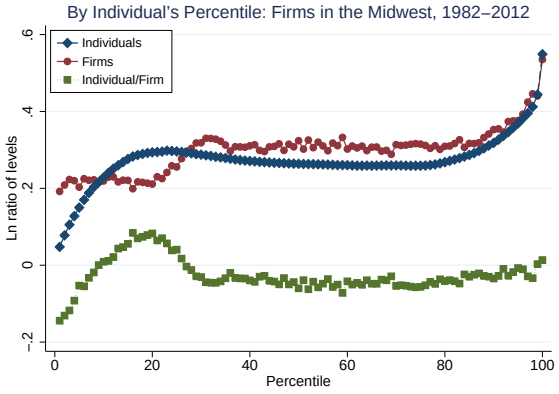
(A) Northeast



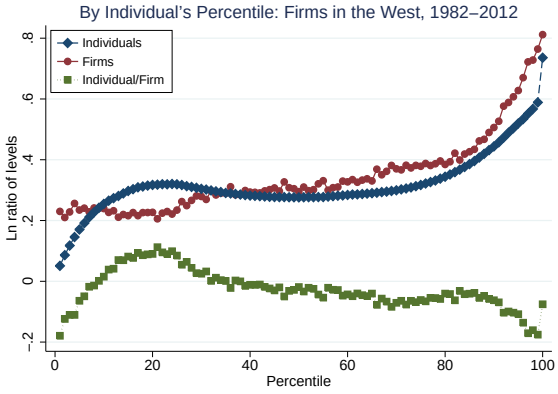
(B) South



(C) Midwest



(D) West



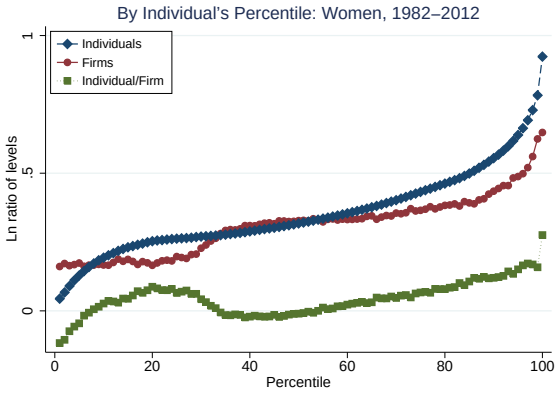
Notes: See the notes for Figure 5.

where wages grew more strongly, generating the closing gender gap, but there was no or little change in gender gap within each firm. In this section, we examine this question.

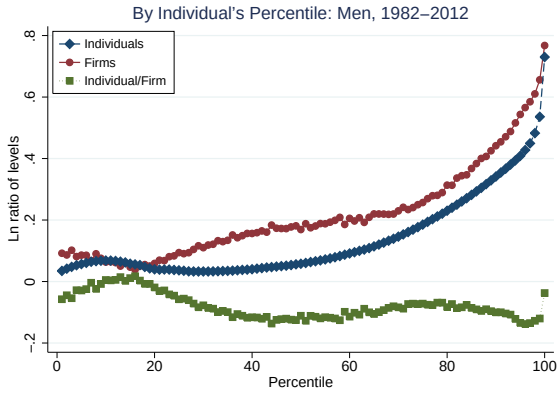
Figure 13 constructs the same graphs as before but now separately for male and female workers. The key line in each panel is the green one. It is above the zero line and increasing with the earnings level for women, whereas the opposite pattern (below zero and declining with earnings) is seen for men. The interpretation is that even within each firm, especially for workers above the median wage, women gained relative to men, and this gain has been larger the higher wage level we focus on. This

FIGURE 13 – Same trends By Gender

(A) Females



(B) Males



Notes: See the notes for Figure 5.

suggests that the gender gap has been closing within firms, particularly at higher pay levels—so skilled employees, managers and executives female workers have seen rising wages in particular.

The fact that women are earning more in 2012 than they were in 1982 is not new; however, we are not aware of other research showing that they are now earning more, relative to their firms, than they did in the 1980s. More research may be needed to understand this fact.

Returns to Age

Next, we turn to the pay structure across age groups. Figure 14 plots our canonical graph for four age groups: workers who are younger than 34, those aged 35 to 44, those aged 45 to 54, and those aged 55 and up. Starting with the first group—young workers—we see that the green line is flat near zero for percentiles below the 40th but then slopes downward reaching –42 log points for the top 1 percent. The implication is that highly-paid young workers have lost ground within the firm, relative to low-paid young workers, as well as relative to the firm average. We see a similar but less dramatic picture for the 35–44 year-old group. Here, the green line is flat between the 20th percentile and the 80th percentile, but those whose earnings were in the top 20 percent suffer a loss relative to the average. The 45–54 year-old group is similar and

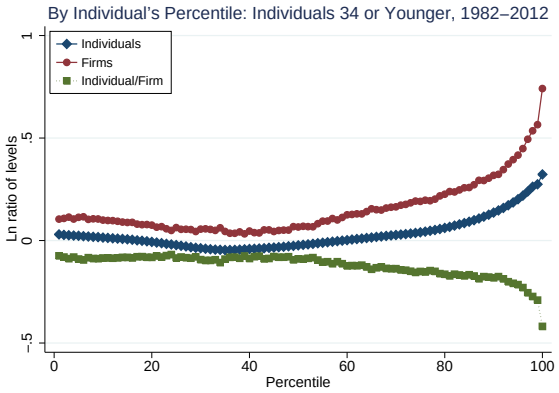
the group of oldest workers seem to have lost the least relative to the firm average. Moreover, inequality within this group did not seem to decline, unlike what we have seen for other age groups. Because of these trends, between 96 percent (for ages 45–54) and 195 percent (for individuals under 35) of the change in inequality for these age groups is accounted for by firms. The fact that results are stronger within age groups may be because we are controlling for that portion of inequality that is caused by changes in the age distribution as the population gets older. On the other hand, we also should note that these are the only subgroups in which none of the individuals in the sample in 1982 are in the same sample in 2012 (except a very small number of individuals who were over 55 and working in both years).

Returns to Firm Tenure

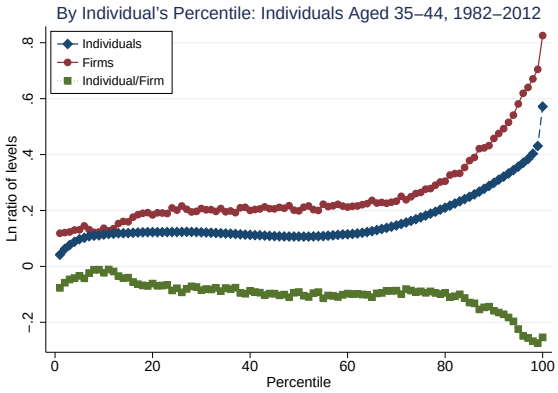
Finally, we examine changes in income inequality among individuals with similar levels of tenure at the same job. Figure 15 plots changes in inequality, restricting in both years to those in their first year at a firm; those with between 2 and 4 years of tenure at their firm; those with between 5 and 9 years of tenure; and those with at least 10 years of tenure. These graphs use 1992 as the starting point so that we can tell how long each individual has been at their firm. The results within these subgroups are similar to those for the whole population. For new employees at the 50th percentile, for example, wage as a fraction of firm wage increased by 3 log points; for new employees in the top percentile, this statistic decreased by 10 log points. Meanwhile, for employees with at least 10 years at the same firm, wage as a fraction of mean wage for 50th percentile individuals decreased by 2 log points; for the top paid among these long-tenured employees, that statistic decreased by 13 log points. Overall, between 128 percent (for those with 2 to 4 years of tenure) and 226 percent (for new employees) of rising individual inequality for these tenure groups is explained by rising firm inequality.

FIGURE 14 – Same trends By Age Groups

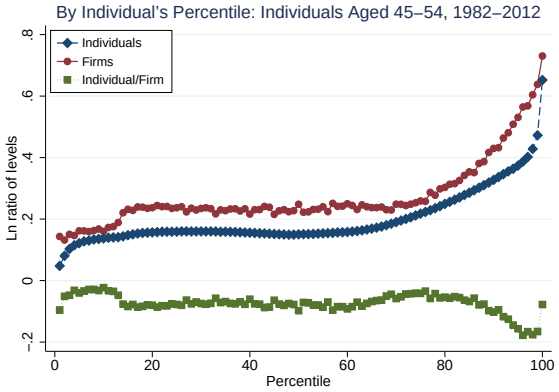
(A) Less than 34



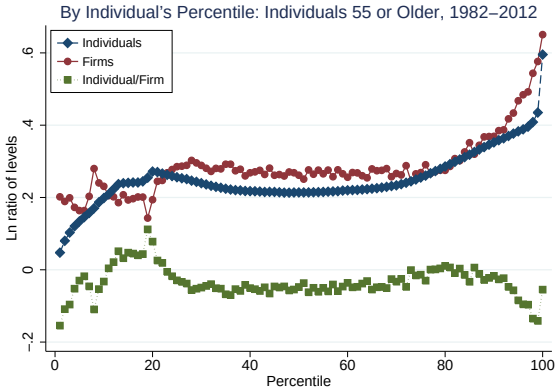
(B) Ages 35 to 44



(C) Ages 45 to 54



(D) Ages 55 and up



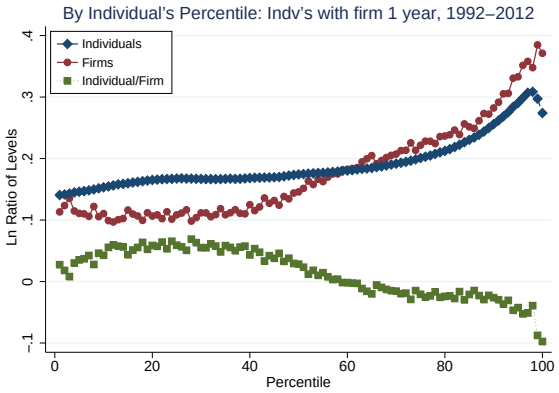
Notes: See the notes for Figure 5.

4 Conclusions

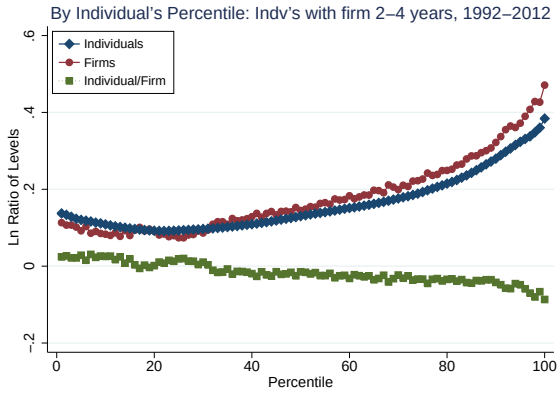
Most Americans earn the vast majority of their income as wages in a firm, but there is little research attempting to understand rising income inequality within the context of these firms. This paper attempts to bridge that gap. Firms appear to matter. Although individuals in the top one percent in 2012 are paid much more than the top one percent in 1982, they are now paid less, relative to their firms' mean incomes, than they were three decades ago. Instead of top incomes rising within firms, top-paying firms are now paying even higher wages. This may tend to make inequality

FIGURE 15 – Trends by Tenure at the Same Job

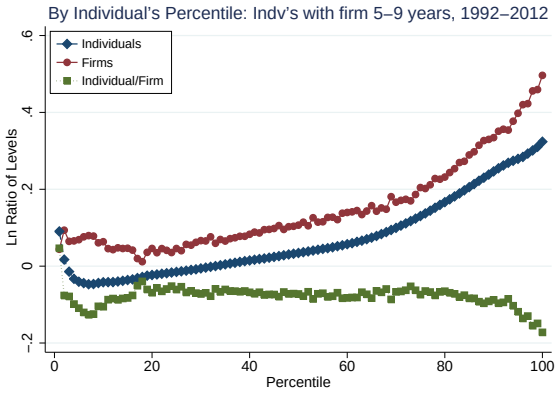
(A) New Employees



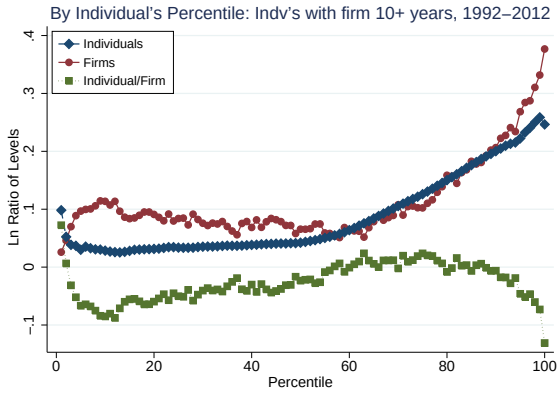
(B) Employees with 2-4 years at the same firm



(C) Employees with 5-9 years at the same firm



(D) Employees with 10+ years at the same firm



Notes: See the notes for Figure 5.

more invisible, as individuals do not see rising inequality among their peers. More research needs to be done to understand why inequality between firms has increased so much more than inequality within them. But this fact of stable inequality within firms should inform our understanding of the great increase in inequality within the United States over the last three decades.

Next steps in this research project are to undertake a more detailed analysis of worker movement across firms following Card et al. (2013), and to use Census and Compustat data to try to understand the movement of plant and firm level differences in pay and productivity.

References

Abowd, John M., Francis Kramarz, and David N. Margolis, "High Wage Workers and High Wage Firms," *Econometrica*, March 1999, *67* (2), 251–334.

Acemoglu, Daron, "Technical Change, Inequality, and the Labor Market," *Journal of Economic Literature*, March 2002, *40* (1), 7–72.

_ **and David Autor**, "Skills, Tasks and Technologies: Implications for Employment and Earnings," in "Handbook of Labor Economics, Vol 4B," Elsevier, 2011, chapter 12, pp. 1043–1166.

Autor, David, Lawrence Katz, and Melissa S. Kearney, "Trends in U.S. Wage Inequality: Revising the Revisionists," *Review of Economics and Statistics*, 2008, *90* (2), 300–23.

Barth, Erling, Alex Bryson, James C. Davis, and Richard Freeman, "It's Where You Work: Increases in Earnings Dispersion across Establishments and Individuals in the U.S.," NBER Working Papers 20447, National Bureau of Economic Research, Inc September 2014.

Card, David, Jörg Heining, and Patrick Kline, "Workplace Heterogeneity and the Rise of West German Wage Inequality," *The Quarterly Journal of Economics*, 2013, *128* (3), 967–1015.

Dunne, Timothy, Lucia Foster, John Haltiwanger, and Kenneth R. Troske, "Wage and Productivity Dispersion in United States Manufacturing: The Role of Computer Investment," *Journal of Labor Economics*, April 2004, *22* (2), 397–430.

Faggio, Giulia, Kjell Salvanes, and John Van Reenen, "The Evolution of Inequality in Productivity and Wages: Panel Data Evidence," NBER Working Papers 13351, National Bureau of Economic Research, Inc 2007.

Guvenen, Fatih, Greg Kaplan, and Jae Song, "The Glass Ceiling and The Paper Floor: Gender Differences Among Top Earners, 1981–2012," Working Paper, University of Minnesota 2014.

_ , **Serdar Ozkan, and Jae Song**, "The Nature of Countercyclical Income Risk," *Journal of Political Economy*, 2014, *122* (3), 621–660.

Juhn, Chinhui, Kevin M Murphy, and Brooks Pierce, "Wage Inequality and the Rise in Returns to Skill," *Journal of Political Economy*, June 1993, *101* (3), 410–42.

_ , **Kevin M. Murphy, and Brooks Pierce**, "Wage Inequality and the Rise in Returns to Skill," *The Journal of Political Economy*, June 1993, *101* (3), 410–442.

Mishel, Lawrence and Natalie Sabadish, "CEO Pay and the top 1%: How executive compensation and financial-sector pay have fueled income inequality," EPI Issue Brief 331, Economic Policy Institute May 2014.

Mueller, Holger M., Paige P. Ouimet, and Elena Simintzi, "Wage Inequality and Firm Growth," NBER Working Papers 20876, National Bureau of Economic Research, Inc 2015.

Piketty, Thomas, *Capital in the Twenty-First Century*, Harvard University Press, 2013.

_ **and Emmanuel Saez**, "Income Inequality In The United States, 1913-1998," *The Quarterly Journal of Economics*, February 2003, *118* (1), 1–39.

A Appendix: Data procedures

As noted in Subsection 2.1, this paper uses data from the Social Security Administration's Master Earnings File. We begin with an extract from this file that includes one observation for each year, for each individual, for each firm that this individual worked for. (For self-employed individuals, the data set also contains these earnings from the IRS as reported in Schedule-SE tax form by the individuals. Because our focus is on firms with employees, we exclude these earnings from our analysis.) For example, if Alice worked at Alpha, Inc. in 2001, 2002, and 2003, and Beta, Inc. in 2002, our file would include four observations with her information—three based on her job at Alpha, and one based on her job at Beta. For each observation, this file includes the year; a transformation of that individual's Social Security Number, along with the associated sex and date of birth; and the EIN, along with the associated 4-digit SIC code and state.

The first step we take with this data is to exclude individuals who did not have a reasonably strong labor market attachment in a given year from the analysis for that year. More concretely, we consider an individual to be "employed" in a given year and include in the analysis if, summing across all jobs, he/she earns at least the equivalent of 40 hours per week for 13 weeks at that year's minimum wage; in 2012, that wage would have amounted to $3,770. (We also conducted robustness checks with other threshold levels, which show similar results.) This condition ensures that we are focusing on data about individuals with a reasonably strong labor market attachment, and that our results are comparable to other results in the wage inequality literature, such as Juhn et al. (1993b) and Autor et al. (2008). The data from any individual earning below this threshold in a given year is excluded from all results for both firms and individuals in that year.

We then calculate statistics for each firm. To do this, we calculate the fraction of each individual's earnings that come from each employer in a given year; this fraction at a given firm represents the number of *full-time equivalent* (FTE) employees that work at that firm. For example, suppose that, in 2002, Alice earned $15,000 at Alpha and $5,000 in Beta. She would then count as 0.75 of an FTE in Alpha and 0.25 of an FTE in Beta. Using this, we calculate the total wage and total employment at each

firm. Suppose that, in 2002, Bob earned $20,000 from Alpha (and no money from any other firm), Carol earned $100 from Alpha (and no money from any other firm), and no one other than Alice, Bob, or Carol earned income from Alpha. We would exclude Carol's earnings because she did not have a strong labor market attachment. The total wage bill at Alpha would be $35,000 ($15,000 from Alice and $20,000 from Bob), while total employment at Alpha would be 1.75 FTE (0.75 FTE from Alice and 1.0 FTE from Bob). Average wage at Alpha would then be calculated as $20,000 (= $35,000/1.75). Next, we assign firm statistics to individuals based on the firm at which that individual had the largest earnings. In our running example, Alice would be assigned firm-wide statistics based on Alpha rather than Beta in 2002, because she had more earnings at Alpha than at any other firm. Thus she would be noted as being in a firm with total wage bill $35,000, employment of 1.75 FTE, average wage of $20,000, and the location and industry of Alpha.

In order to analyze a representative sample of individuals in a computationally feasible way, we analyze a one-sixteenth representative sample of all US individuals from 1978 to 2012. The sample is organized as a longitudinal panel, in the sense that once an individual is selected into the sample, he/she remains in the sample until he/she dies. In particular, an individual is in our sample if the MD5 hash of a transformation of their Social Security Number begins with a zero; because MD5 hashes are hexadecimal numbers, this will select one in sixteen individuals. MD5 is a cryptographic algorithm that deterministically turns any string into a number that is essentially random. It is designed so that a slightly different input would lead to a completely different output in a way that is essentially impossible to predict. Because it took cryptographic researchers several years to figure out a way that, under certain circumstances, MD5 is somewhat predictable, this algorithm is certainly random enough for our purposes. Thus whether one individual is included in our sample is essentially independent of whether some other individual is included, regardless of how similar their SSNs are.

Where our results analyze the same firm over multiple years, we include a correction to ensure that firms that change EINs are not counted as exiting in one year and entering in the next. We define an EIN in Year 1 as being the same firm as a different EIN in Year 2 if the following conditions are met. First, Year 1 must be the last

year in which the original EIN appears, while Year 2 must be the first year that the new EIN appears in our data. Next, more than half of the individuals who worked in each firm must have also worked in the other firm. Finally, to ensure that our results aren't influenced by a few individuals switching companies, we only include EINs that employ at least 10 individuals.

Firms are only included in our sample if they have at least 10 FTEs in a given year to ensure that firm-wide statistics are meaningful; for example, comparing an individual to the mean wage at their two-person firm may not be a good way to characterize inequality within firms in a given year (though our results are robust to changing this threshold, as shown in Figure 9). We also exclude firms in the Educational Services (SIC Codes 8200 to 8299) and Public Administration (SIC Codes 9000 to 9899) industries, as employers in these industries are frequently not what we would consider firms. Finally, we exclude employers with EINs that begin with certain two-digit codes that are associated with Section 218 Agreements, or other issues that may not be handled consistently in the data across years. Individuals whose primary job is with a firm in one of these excluded categories are also dropped from the data in that year. Thus Alice and Bob, in the example above, would not a part of any individual-level statistics for 2002, because Alpha, the primary firm for both, did not have at least 10 FTEs. (However, Alice's $5,000 income at Beta, and the 0.25 FTE she counts for there, would be included in Beta's statistics. Excluding Alice's earnings and employment from Beta's totals because Alpha is below the size threshold could in turn force Beta below the size threshold, and then excluding Beta's employees could force several other firms below the size threshold. To avoid this chain reaction, we simply include Alice in Beta's numbers so long as her total earnings are above the "strong labor market attachment" threshold.)

To avoid potential problems with outliers and to address privacy concerns, we cap (Winsorize) observations above the 99.999th percentile. Winsorized variables are firms' total employment; firms' total wage bill; firms' average wage; individuals' wage; and individual wage as a fraction of average firm wage. Variables are Winsorized immediately before analysis. For example, suppose that Gamma, Inc. is the largest firm in the data set, with 1 million employees and a total wage bill of $30 billion, while the 99.999th percentile for employees is 100,000, and for total wage is $10 billion. Then

Gamma, Inc. would be analyzed as though it had 100,000 employees and $10 billion in wages, but average income of $30,000.

Finally, we adjust all dollar values in the data set to be equivalent to 2012 dollars with the Personal Consumption Expenditure (PCE) price index[12].

[12] http://research.stlouisfed.org/fred2/series/PCEPI/downloaddata?cid=21