An Anatomy of Monopsony: Search Frictions, Amenities and Bargaining in Concentrated Markets*

David Berger Kyle Herkenhoff Andreas R. Kostol Simon Mongey

April 12, 2023

Abstract

We contribute a theory in which three channels interact to determine the degree of monopsony power and therefore the wedge between a worker's spot wage and her marginal product (henceforth, the *wage markdown*): (1) heterogeneity in worker-firm-specific preferences (nonwage amenities), (2) firm granularity, and (3) off- and on-the-job search frictions. We use Norwegian data to discipline each channel and then reproduce novel reduced-form empirical relationships between market concentration, job flows, wages and wage inequality. Our main exercise quantifies the contribution of each channel to income inequality and wage markdowns. The markdowns are 21 percent in our baseline estimation. Removing nonwage amenity dispersion narrows them by a third. Giving the next-lowest-ranked competitor a seat at the bargaining table narrows them by half. Removing search frictions narrows them by two-thirds. Each counterfactual shows decreased wage inequality and increased welfare.

JEL codes: E2, J2, J42

Keywords: Monopsony, Inequality

*Berger: Duke University. Herkenhoff: University of Minnesota, Federal Reserve Bank of Minneapolis. Kostol: ASU WP Carey School of Business. Mongey: Kenneth C. Griffin Department of Economics, University of Chicago; Federal Reserve Bank of Minneapolis. This article is in preparation for the *NBER Macroeconomics Annual 2023*, and we thank the editors Martin Eichenbaum, Erik Hurst and Valerie Ramey. We also thank our discussants Giuseppe Moscarini and Katarína Borovičková. Any opinions and conclusions expressed herein are those of the author(s) and do not necessarily represent the views of the Federal Reserve Bank of Minneapolis or the Federal Reserve System. Thanks to Alexander Weinberg for excellent research assistance. This work was supported by the Norwegian Research Council grant numbers 227115 and 275123 and NSF Award $\#SES-2214431$.

1 Introduction

There is a growing consensus that imperfect competition in the labor market is pervasive[1]. Many local labor markets are dominated by a few firms, which gives them the ability to set wages and pay workers less than their marginal product. In his 2022 *AEA presidential address*, David Card argued that developing a tractable framework combining preference heterogeneity (in the *Daniel McFadden–industrial organization* tradition) and search-and-matching frictions with job ladders (in the Postel-Vinay and Robin, 2002, tradition) was key to understanding the importance of market power in the labor market. In this paper, we develop a theory of monopsony that incorporates the three paradigms of the modern monopsony literature: worker–firm-specific preference heterogeneity (Robinson, 1933), search frictions (Burdett and Mortensen, 1998), and firm granularity (Berger, Herkenhoff, and Mongey, 2022a; Jarosch, Nimcsik, and Sorkin, 2019). We then quantify our framework using Norwegian worker–firm data and use it to answer several pertinent questions: *How do the three sources of monopsony interact to shape wages, job flows, and welfare? Which sources of monopsony account for the wedge between a worker's pay and marginal product (henceforth, the wage markdown)? How does monopsony power affect wage inequality?*

We make three contributions. First, we use Norwegian administrative data to document that within an occupation, wage levels, wage inequality, and job flows vary systematically with local labor market concentration. These novel empirics motivate a number of our modeling assumptions and allow us to conduct overidentification tests on the role of concentration in our theoretical framework. Second, we develop a model of frictional labor markets with a finite number firms, as well as on-the-job-search, worker–firm-specific nonwage amenities, and vacancy posting. While some recent papers include the first two, we provide the first general equilibrium theory of search in granular markets, allowing us to examine meaningful counterfactuals. Third, we use the Norwegian data and the structure of our model to discipline and then quantify the wage, welfare, and job flow implications of each source of monopsony power. Our framework implies that amenities, duopsonistic wage bargaining, and search frictions account for one-third, one-half, and two-thirds of wage markdowns, respectively, where nonadditivity arises from the inherent nonlinearity of our model.

Empirics. We begin by using detailed data about workplace locations and workers' line of work to document the relationship between concentration levels, wages, wage inequality, and job flows. Using two separate fixed effect specifications that isolate within occupation–year, across-region variation and within occupation–region, across-time variation, we document a set of covariances between market concentration and labor market characteristics. More concentrated markets are associated with lower wages, less wage dispersion, lower employer-to-employer job flow rates, and lower job-finding rates. It is well-known that job-to-job transitions are a key source of wage growth (see, e.g., Postel-Vinay and Robin, 2002). These

[1] See Manning (2003) for a summary of the literature as well as and recent papers Dube, Jacobs, Naidu, and Suri (2019), Card, Cardoso, Heining, and Kline (2018), Lamadon, Mogstad, and Setzler (2022), Benmelech, Bergman, and Kim (2022), Azar, Marinescu, and Steinbaum (2022), Yeh, Macaluso, and Hershbein (2022), Schubert, Stansbury, and Taska (2022), and Berger, Herkenhoff, and Mongey (2022a), as well as Brooks, Kaboski, Li, and Qian (2019), Felix (2022) and Rubens (2023) outside of the U.S., among others.

strong links between job flows and concentration suggest that on-the-job search may be an important mechanism through which market structure affects the level and dispersion of wages.

Model. Motivated by these findings, we develop a theory that incorporates neoclassical sources of monopsony as well as frictional job flows and concentration. Our model features a finite number of firms, on-the-job search, worker–firm-specific nonwage amenities and strategic wage setting. The first two are necessary to replicate the fact that both employer-to-employer flows and wages are lower in more concentrated markets. Additionally, firm contact rates are determined in general equilibrium by optimal firm vacancy posting, given the endogenous distribution of workers across firms. This allows us to formulate counterfactuals and measure welfare.

Our framework substantially extends Postel-Vinay and Robin (2002) to accommodate a finite number of firms in each labor market and worker–firm-specific nonwage amenities. The number of firms in each market is given by $M < \infty$, and firms differ in their idiosyncratic but fixed productivity levels. Unemployed workers randomly meet vacancies of the M firms within their labor market, whereas employed workers randomly meet vacancies of the remaining $M-1$ firms, excluding their current employer. When an unemployed worker meets a firm, she draws a worker–firm-specific nonwage amenity, and the parties Nash bargain over surplus. When an employed worker meets a firm, she draws a new worker–firm-specific amenity, and the incumbent and poaching firms compete via alternating offers (e.g., Cahuc, Postel-Vinay, and Robin, 2006), yielding equilibrium values in which outside options are determined by Bertrand competition and the remaining surplus is split by Nash bargaining. This duopsony surplus-sharing protocol yields strategic complementarity between wage offers of incumbent and poaching firms. As we discuss below, the "hard-wired" duopsony in the Cahuc, Postel-Vinay, and Robin (2006) class of models is an important source of markdowns that operates through strategic wage setting. Last, we assume that firms optimally choose vacancies, taking match surplus and market contract rates as given. Modeling of a firm's vacancy posting decisions explicitly generates endogenous contact rates, which both closes the model and delivers a meaningful firm size distribution.

There are several caveats to our approach. Neoclassical models of monopsony focus on worker-firm specific amenities as generating a mechanism by which firms must increase wages to attract more workers. We take those same worker-firm-specific amenities and study them in a search and matching model where we find they are necessary to match features of the data and provide market power for firms, but we assume that they operate differently than in the neoclassical case. We allow firms to observe amenity draws and perfectly price discriminate. This differs with Robinson (1933) who treats amenities as unobserved and thus only allows for third degree discrimination. An important implication of our approach is that it does not necessarily yield inefficient allocations.

Quantification. We use linked employer–employee data from Norway to discipline the quantitative model. The administrative data offer several advantages over other data sources. The Norwegian data include information about the type of work that each employee is hired to do (i.e., her occupation) and the workplace location of every employment contract from 2006 to 2016. These complete records allow us to accurately measure job flows and thus better classify local labor markets.

We contribute a simple clustering algorithm to define local labor markets, which we apply to these data. Rather than working with connected sets of firms (which is computationally demanding), our approach uses the much lower-dimensional occupation-to-occupation flow transition matrix. We first isolate single-occupation markets with high self-flow rates. Among the remaining occupations, we K-means cluster the rows of the occupation-to-occupation flow transition matrix. The resulting groups are occupations with similar job flow patterns. We determine the optimal number of clusters using an objective function that is *increasing* in the self-flow rate but *decreasing* in the concentration of occupations in each cluster. This rewards the lowering of self-flow rates but penalizes the classification of all occupations in one large cluster and thus 'overfitting' of the data. We do not innovate on the dimension of geography; we simply define the boundary of a market to be the boundary of the commuting zone as computed in Bhuller (2009). This yields approximately 5,000 markets with a self-flow rate of 51 percent (among job transitioners, 51 percent transition back into the same market). Using the raw 3-digit occupation by commuting zone to define the market yields a similar number of markets but a lower self-flow rate of 45 percent. In an approach that intuitively groups connected sets of firms (stochastic blocks), Jarosch, Nimcsik, and Sorkin (2019) find 376 markets with a self-flow rate of 40 percent in Austria. Worker occupation data allow us to define markets with self-flow rates that improve upon the rate in Jarosch, Nimcsik, and Sorkin (2019) while significantly reducing computation time.

With market definitions in hand, we estimate the model to match key moments between 2006 and 2016 in Norway. First, we directly import market structures observed in the Norwegian data, including the number of firms in a market and the labor force in the market. Second, we discipline the role of amenities using the fraction of E-to-E job moves down the ladder, where rungs are defined by the poaching index following Bagger and Lentz (2019). Third, we discipline the role of search frictions based on the unemployment and E-to-E rates. Fourth, we discipline the bargaining power of workers using wage growth. We estimate the model on an overidentified set of moments to ensure that our model is consistent with observed covariances of market Herfindahl values and the level and standard deviation of wages. Despite its parsimony, the quantitative model fits nontargeted cross-market moments from our earlier empirics, generating lower E-to-E rates and U-to-E rates in more concentrated markets.

Results. In our main results, we use the model to generate five counterfactuals that highlight the importance of labor market competition for markdowns, welfare and wage inequality.

I. Concentration. We explore the role of concentration in depth since it is the newest element of our analysis. Our goal is to vary the number of firms (M) while holding the distribution of productivity (z) and bargaining protocol fixed. To implement this "idealized" *heuristic* experiment, we draw a vector of ten productivities from the ergodic distribution and duplicate this productivity vector 10 times to construct 10 markets. We then organize the productivity vectors into three counterfactual economies (from most to least concentrated): ten identical 10-firm markets, two identical 50-firm markets, and one 100-firm market. We hold the ratio of firms per worker fixed to remove mechanical effects from adding firms to a market. We find that workers' share of surplus monotonically increases as markets become less concentrated. A side effect of a higher surplus share is a reduction in compensating differentials;

i.e., the amenity wage penalty shrinks. With less concentration, wages w rise, markdowns narrow (i.e., w/z increases), and welfare increases. Inequality increases: across firms, more productive firms pay higher wages, and within firms, some workers are paid more due to better bargaining opportunities. Consolidating all 100 firms into a single market allows more workers to reach the highest-productivity firms, increasing output. Accordingly, unemployment falls, and E-to-E rates increase as more meetings result in job transitions.

We next explore the role of concentration in the actual Norwegian economy. We double the number of firms in the Norwegian economy by duplicating the existing productivity vector in every market (i.e., every firm's doppelgänger enters the market, leaving the job ladder rungs untouched) and double the number of workers in the market such that the number of firms per worker remains constant. While this experiment yields results that are qualitatively consistent with our idealized *heuristic* experiment, the quantitative effects of concentration on wages and welfare are limited. Markdowns narrow by 1 percentage point, average wages rise by 0.68 percent, and the standard deviation of log wages increases by 0.81 percent. Why are the effects of changing M so small? There are two reasons: (1) Approximately 70 percent of the labor force resides in markets with more than 150 firms, and hence, doubling the number of firms in these markets is irrelevant, and (2) the duopsony wage-setting assumption—i.e., that only two firms at the bargaining table strategically set wages—remains unchanged regardless of M.

II. Exclusion. To explore the effect of firms removing themselves from future E-to-E contacts with the worker, we allow a worker at a given firm to meet that firm again, rebargain and thus extract all the surplus. This mechanism is related to the approaches in Zhu (2012) and Jarosch, Nimcsik, and Sorkin (2019)[2]. We find that this leads to an economically small change in observed markdowns, narrowing them by approximately 1 percentage point compared to the 21-percentage-point markdown in the baseline economy. Again, the reason is that the bulk of the labor force resides in markets with many firms, limiting the impact of self-exclusion.

III. Preference heterogeneity. To quantify the effects of preference heterogeneity (i.e., nonwage, worker–firm-specific amenities), we eliminate all variations in amenities and set them to a single value, resulting in a uniform level of amenities across the economy, while we maintain the same level of aggregate amenities as in our baseline economy. In this counterfactual, we find that the Herfindahl index more than triples. Workers now flow to the highest-productivity firm and stay there, unlike in our baseline economy, where differences in amenities can cause workers to leave the highest-productivity firms. High-productivity firms also post more vacancies, understanding that workers are less likely to leave due to idiosyncratic tastes. As a result, output, productivity, and welfare increase substantially. Without amenity dispersion, wage inequality falls, but wage levels rise as workers flow to—and stay at—more productive firms. Despite the level of amenities being the same, markdowns narrow by 7 percentage

[2]In our baseline model, firms do not exclude themselves from future U-to-E transitions, only contemporaneous E-to-E meetings. Thus, this particular experiment holds fixed the set of possible U-to-E transitions, differentiating what we do from the approach in Jarosch, Nimcsik, and Sorkin (2019).

points[3]. This represents a 30 percent reduction in markdowns from the level in the baseline economy.

IV. Search frictions. To quantify the effects of search frictions, we increase worker contact rates to 100 percent per period, leading workers to always meet a firm in every period regardless of their employment status. Markdowns narrow by 14 percentage points, which represents a 60 percent reduction from the level in the baseline economy. Greater contact with lower-ranked firms allows workers at higher-productivity firms to bid up their share of surplus, while they are also more likely to meet the highest-ranked firms. The Herfindahl index rises to 0.75 (an eightfold increase), as workers rapidly climb the job ladder. Wage inequality falls dramatically, as a majority of workers work at the highest-productivity firm and quickly negotiate the highest-possible surplus share.

V. Bargaining. As in Cahuc, Postel-Vinay, and Robin (2006), in our model, there are only ever two firms at the bargaining table. Duopsony is a feature of the economic environment regardless of the number of firms in a given market. We remove this "hard-wired" role for duopsony by assuming that whenever a worker meets a firm whose surplus rank is K, she also meets the next-best firm (i.e., the rank $K+1$ firm). Now three firms are always at the bargaining table. Holding vacancies fixed, this protocol does not alter allocations. It simply redistributes surplus from firms to workers, increasing wages. In general equilibrium, however, vacancies adjust (since the firm's share of surplus is now lower), shifting the allocation of workers to firms. We additionally assume a single amenity value calibrated to deliver the same aggregate amenities as in our baseline economy, which allows a monotonic ranking of firms. With a single amenity, compensating differentials factor less into wages, and markdowns narrow to 16 percentage points. Our main result is that when we additionally alter the bargaining protocol, markdowns narrow down to 3 percentage points. This explains approximately half of the baseline economy's markdowns. This experiment suggests that future work on the precise structure of strategic wage setting is valuable.[4]

Our counterfactuals have implications for policymakers who may seek to address inefficiencies arising from labor market power. Our wage decomposition results point to a significant role for policies that alleviate markdowns due to amenities and strategic wage setting. Merger policy—primarily focused on the number of firms in a market, M—may have more moderate effects on wage markdowns absent multiparty bargaining. The extent to which bargaining is between two, three or more parties is an empirical question that deserves more attention in light of our findings.

Further research can use our framework to study the distributional consequences of policies in granular labor markets. It is tractable enough to incorporate realistic policies (e.g., minimum wages, taxation, and antitrust), richer theories of the household and firm (e.g., costly human capital accumulation), and alternative contractual environments (e.g., noncompetes, as in Shi (2023)).

We review the literature and then proceed as follows. Section 2 describes the Norwegian administrative data and offers motivating empirics. Section 3 describes the model and defines the equilibrium. Section 4 provides details on model calibration and fit. Section 5 decomposes wages in the steady state to analyze the mechanisms through which concentration shapes wages. Section 6 conducts the main

[3] Roussille and Scuderi (2022) find a similarly large role for amenities in empirical analysis of online job-board wage postings.

[4] Early progress on this question is being made by Flinn and Mullins (2021), among others.

counterfactual exercises and discusses potential policy implications of our findings.

Related literature. We contribute to a growing theoretical and quantitative literature by integrating the three existing monopsony paradigms into one framework: search frictions, nonwage amenities, and granularity. There are two main classes of monopsony models, each with two subgroups: (i) models in which frictional markets generate monopsony power with a continuum of firms (e.g., Burdett and Mortensen, 1998; Manning, 2003; Engbom and Moser, 2022; Hurst, Kehoe, Pastorino, and Winberry, 2022) and a finite number of firms (Burdett, Shi, and Wright, 2001; Zhu, 2012; Jarosch, Nimcsik, and Sorkin, 2019; Bagga, 2022; Bloesch and Larsen, 2023) and (ii) models in which neoclassical markets in the presence of amenities generate monopsony power with a continuum of firms (e.g., Robinson, 1933; Card, Cardoso, Heining, and Kline, 2018; Taber and Vejlin, 2020; Kroft, Luo, Mogstad, and Setzler, 2020; Lamadon, Mogstad, and Setzler, 2022) and a finite number of firms (e.g., Bhaskar and To, 1999; Bhaskar, Manning, and To, 2002; Berger, Herkenhoff, and Mongey, 2022a; Azkarate-Askasua and Zerecero, 2022; Berger, Herkenhoff, and Mongey, 2022b). Unlike these existing frameworks, our model simultaneously features (1) search and matching frictions, (2) neoclassical non-wage amenities, (3) price discrimination within the firm, and (4) a finite number of firms. Additionally, we model vacancy posting, and thus our general equilibrium model links employer concentration to both prices and quantities. This allows us to discuss welfare and conduct normative counterfactual exercises.

We contribute to a growing empirical literature that explores the relationship between worker and firm outcomes and market granularity. Recent work has documented cross-sectional relationships between standard measures of concentration (Herfindahl index values) and wages or employment (Benmelech, Bergman, and Kim, 2022; Rinz, 2022; Yeh, Macaluso, and Hershbein, 2022) and vacancies (Azar, Marinescu, Steinbaum, and Taska, 2018; Azar, Berry, and Marinescu, 2022; Azar, Marinescu, and Steinbaum, 2022). To our knowledge, we are the first to (i) document reduced-form, cross-market relationships between Herfindahl values, job flows, and various measures of wage inequality, including within- and between-firm wage inequality, and (ii) combine occupational data and clustering techniques to define markets.

2 Empirical analysis

This section presents new evidence on market structure, job mobility, and wage-setting behavior in Norwegian labor markets.

2.1 Data and measurement

While the use of linked employer–employee data covering the universe of firms, establishments and employees is now common among researchers, the Norwegian data have the key advantage that employers must record the type of work that each employee is hired to do and the workplace location of every employment contract from 2006 to 2018. These bibliographic records allow us to define labor markets using geography and occupation rather than industry (as in Berger, Herkenhoff, and Mongey, 2022a)[5]. We

[5]For the US, economists have used Burning Glass data for occupation and wage information (Schubert, Stansbury, and Taska, 2022). These data pose serious issues for analyses such as the one here. First, the data do not contain information on

can then count the employers within each occupation–location market and track changes to individuals' wages when they move between employers.

Data. Data collection consists of two steps. In the first step, we use the information about the work contract that the employer submits to the employment agency (NAV). To comply with labor laws, the employer must enter a specific position from a list of more than 6,000 possible job titles and the workplace location, wage income, and work hours[6]. Job titles are then grouped into 354 four-digit occupations by Statistics Norway based on similarity of work[7]. We cluster these occupations using novel techniques to compute the *occupational scope* of labor markets. In the second step, we combine linked employer–employee data with socioeconomic variables from longitudinal population registers. These include demographic information (e.g., sex, age, residential municipality, and education). We can therefore determine commuting distances between residence and workplace, which facilitates computation of the *geographical scope* of labor markets. We can therefore allow labor markets to cross administrative borders of municipalities and counties[8].

Institutional detail. Norway has a population of 5 million, and Oslo, the capital, accounts for approximately one-fifth. The labor force aged 25 to 66 is some 2 million, and the labor share of income is approximately 70 percent[9]. In 2016, unemployment was approximately 4.5 percent. There were 176,019 firms and 234,941 establishments with workers on payroll.

Firms can hire employees on either fixed-term or permanent contracts and can dismiss workers if they underperform relative to their peers or if the firms are operating at a loss. Employment protection in Norway ranks near the median among OECD countries and is comparable to that in France and Sweden[10]. Wages and typical working hours, in turn, tend to be set by collective bargaining at the industry level, after which wages are supplemented by local adjustments or wage drift, bargained over at the worker–

the universe of employees, employers, and jobs or wages paid to employees. Second, data on advertisements lack information on the quantity of positions and hence cannot be used to compute market shares. Third, only 6 percent of the advertisements scraped and collated by Burning Glass have wage, employer and occupation information. Table A4 of Hazell, Patterson, Sarsons, and Taska (2022) shows that while the 2010 to 2019 data contain 239 million ads, dropping those without wages or a range of wages posted and without firm, county, sector, or occupation data leaves only 15 million ads, which is 6.27 percent of the initial sample. Further screening reduces their analysis sample to less than 1.6 percent of all ads.

[6]The 4-digit occupational classification is based on the International Standard Classification of Occupations (ISCO) adapted to Norwegian labor markets by Statistics Norway and the employment agency. There are strong incentives for correct reporting. First, the Norwegian labor law stipulates that firms undergoing a mass layoff, defined as laying off more than ten workers over 30 days, must follow the last-in, first-out principle. The ordering is typically defined within position and establishment. Second, the employment agency uses information about occupation and the workplace location for targeted job search assistance. In practice, employers report positions by a 7-digit system (see https://www.ssb.no/klass/klassifikasjoner/145/, in Norwegian), with new job titles added at regular intervals.

[7]For example, the occupational code for "economics and business" includes consultants, controllers, junior and senior credit analysts, and research and chief economists, to name a few. The codes also cluster unskilled positions, such as maintenance workers and janitors, and different levels of management positions into distinct groups.

[8]See Data Appendix A for a more detailed description of the sources and variables and Bhuller (2009) for commuting patterns.

[9]The petroleum sector accounts for a large fraction of income but is excluded from the calculation of the labor share.

[10]Union membership in Norway is high relative to that in other countries in the OECD and the US but fell from 58 to 53 percent from 1992 to 2013 (link: OECD Statistics Trade Union Statistics). Unions play an important role in ensuring that firms comply with labor law, stating, for example, that downsizing requires a one-month notification to employees, with the dismissal time varying from one to six months, depending on age and tenure. Wrongful discharge can end with a lawsuit, where firms must compensate the dismissed employees for lost income.

firm and collective agreement level, which may vary by occupation within a firm (see, e.g., Bhuller, Moene, Mogstad, and Vestad, 2022). This two-tier framework gives rise to a relatively compressed wage structure. The Norwegian safety net covers lost income from unemployment. The primary insurance source is unemployment insurance (UI) benefits, which begin after a three-day waiting period, replacing approximately two-thirds of workers' past earnings net of tax (see, e.g., Røed and Zhang, 2003)[11].

2.2 Defining labor markets

In our empirical analysis, we build markets from the ground up using individual data. To avoid issues of entry and exit from the labor force, we focus on residents aged 25 to 60. We define a *local labor market* as a group of 4-digit occupations within a commuting zone (CZ) region indexed by r, where the CZs are taken from Bhuller (2009). Below, the *self-flow rate* is the fraction of job-to-job transitions from one group of occupations back into the same group[12]. We then group occupations as follows:

1. First, we isolate single-occupation markets with high self-flow rates (e.g., those with rates of more than 50 percent, such as the market for dentists)
2. Among the remaining occupations, we K-means cluster the rows of the occupation-to-occupation flow transition matrix
3. For each commuting zone region indexed by r, we compute the Herfindahl of employment across clusters (HHI_r^K).[13] We then determine the optimal number of K clusters by maximizing an objective function that is (i) increasing in the self-flow rate but (ii) decreasing in HHI_r^K, such that we penalize the classification of all occupations in one large cluster:

$$\underbrace{\frac{\text{Average self-flow rate of all } K \text{ clusters}}{\text{Standard deviation self-flow rate of all } K \text{ clusters}}}_{\text{(i) Reward on fit}} \times \underbrace{\left[1 - HHI_r^K\right]}_{\text{(ii) Penalty on overfitting}}$$

Table 1 summarizes our market definitions. Our procedure yields a self-flow rate of 51 percent. We obtain approximately 103 clusters per commuting zone, and the average market has 404 workers. The unweighted number of employees per firm–market is 6, and the average firm operates in 2.3 markets.

To ground our definition of markets, consider the fictitious example of a dental care firm in Oslo named *ABC Dental*. *ABC Dental* is an 8-person firm that hires workers in three occupations corresponding to its 6 dentists, 1 janitor and 1 groundskeeper. The commuting zone region of *ABC Dental* is Oslo. Suppose dentists are in a single dentist cluster, and janitors and groundskeepers are both in a manual low-skill service cluster.[14] *Markets* are cluster–CZ pairs and indexed by j (e.g., 'manual low-skill services-

[11] Payroll taxes finance the UI system, and there is no experience rating on the firm. The potential benefit period is 52 weeks for workers who have earned less than twice the National Insurance basic amount for the last three years. The "basic amount" of benefits is currently approximately USD 1,000 per month. UI benefits are capped at a maximum level of previous earnings, currently six times the basic amount, which creates a kink in the benefit formula. To remain eligible for the cash benefits, work hours must have fallen by at least 50 percent, and recipients must be actively looking for work and willing to take any employment.

[12] In Appendix A.3, we describe the construction of self-flow rates and plot the employment distribution of occupations by their self-flow. Approximately 50 percent of the workforce has a rate above 50 percent.

[13] $HHI_r^K = \sum_{k=1}^{K}(s_r^k)^2$ where s_r^k is the employment share of occupation cluster $k \in \{1, \ldots, K\}$ in commuting zone region r.

[14] Note the clustering algorithm does not produce 'labels.' We only use the label ' manual low-skill services' for heuristic purposes.

Moment	Value
Fraction of EE flows within market 2006–2016 (%)	51.40
Number of markets per region	102.8
Average firm employment per market	6.20
Average labor force per market	404.7
Average markets per firm	2.30
Total number of markets	4783

Table 1: Market definition summary statistics

Notes: Summary statistics are unweighted. All rows except top row are calculated from December 2016.

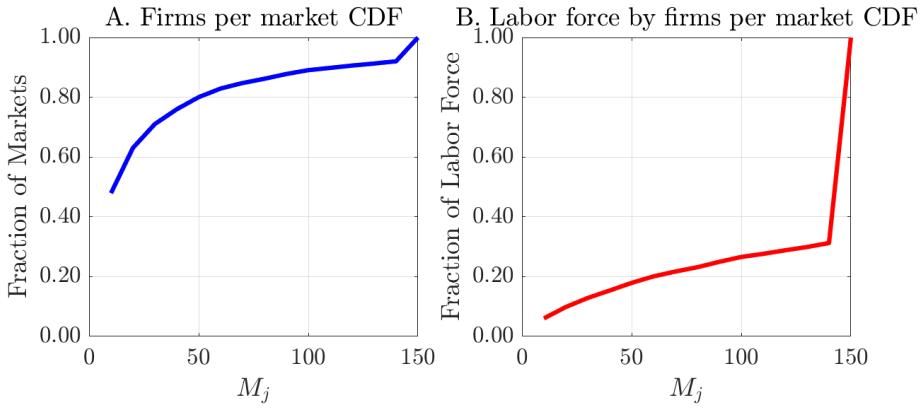


Figure 1: CDFs of market size and labor force

Notes: Panel A is the cumulative distribution function (CDF) of the number of firms per market. We winsorize the data at 10 and 150 firms per market. Panel B is the CDF of the labor force by number of firms per market.

Oslo' is one market and 'dentists-Oslo' is another). *Firms* are firm–market pairs (e.g. 'manual low-skill services-*ABC Dental*-Oslo' is treated as a different firm than 'dentists-*ABC Dental*-Oslo'). *Occupations* are the 4-digit raw occupation (e.g. dentist, janitor, and groundskeeper).

Figure 1 plots the distribution of firms and employment across markets, ordered by the number of firms in each market M_j. As we do in the calibration, we truncate the graph at $M_j = 150$ firms in a market. Only 10 percent of markets have more than 150 firms (Panel A), but these markets employ more than 70 percent of the Norwegian labor force (Panel B).

Table 2 provides summary statistics on key labor market outcomes for 2006 to 2016. The economy-wide unemployment rate averaged 4 percent. The monthly job-to-job transition rate was 0.7 percent, the job finding rate was 8 percent per month, and the layoff rate was 0.4 percent. Relative to those in US data, worker flows in the Norwegian labor market are noticeably lower. We note that all flow rates in the main body of the text are computed within markets (flows within markets divided by workers remaining within markets) to be consistent with the model definition of a market. Including job flows outside of the market yields similar statistics and covariances (see Appendix D). We then rank firms by the fraction of their hires coming from other firms—i.e., the poach rank index (see Bagger and Lentz, 2019)—and find that 15 percent of job-to-job transitions are down this ladder. We use this moment to discipline the role of nonwage amenities.

Moment	Value	Moment	Value
Unemployment rate (%)	4.01	Average firms per market	75.2
E-to-E rate (monthly %)	0.65	HHI wage bill (wage bill weighted)	0.09
U-to-E rate (monthly %)	8.08	HHI employment (employment weighted)	0.08
E-to-U rate (monthly %)	0.35	Standard deviation of log wages	0.63
Fraction E-to-E moves down poach index	0.15	Fraction of log wage variance within firms	0.61

Table 2: Summary statistics – Worker flows, concentration, wage inequality

Notes: Flow rates are computed within markets (flows within markets divided by workers remaining within markets) to be consistent with the model definition of a market. Similar statistics are obtained when we include job flows outside of the market.

The average number of firms in a market is large: 75.2. However, markets are concentrated. The average wage bill Herfindahl index (weighted by the wage bill[15]) is 0.09. This is the same concentration as in a market with only 11 equally sized firms[16]. Similarly, the average employment Herfindahl index (employment weighted) is 0.08. The wage bill HHI is higher due to larger firms paying higher wages. The standard deviation of log wages is 63 percent, with the bulk of this (61 percent) accounted for by dispersion in wages within firms.

2.3 Regression framework

Our goal is to study the relationship between employer concentration wage levels, the wage distribution, and job-to-job transitions. We provide a set of covariances between concentration and labor market outcomes that any theory of concentration and labor market dynamics should replicate. We do not attempt to attribute causality. We, along with the existing literature, lack credible instruments for measuring concentration. For example, take the change in local concentration due to a national firm exit used by Azkarate-Askasua and Zerecero (2022). Through the lens of our theory in Section 3, this shock alters the rungs of the job ladder. Thus, it is not a pure "concentration" shock, ceteris paribus.[17] The same is true of instrumenting changes in local exposure with changes in national concentration (e.g., Azar, Marinescu, and Steinbaum, 2022) since the second stage still implies changes in the number of rungs on the local job ladder.

Instead, we consider two different dimensions of variation in the Herfindahl values by using across-region variation within occupation–years and across-time variation within occupation–regions. Importantly, however, the unit of observation remains four-digit occupations in a commuting zone, allowing us to use occupation fixed effects. Each approach differs in its interpretation. Comparisons across regions may reflect sorting across space on unobservables. Comparisons across time may reflect changing demand patterns. We do not take a stance on what drives these covariances. However, both approaches provide similar negative correlations between Herfindahl values, wages, job flows, and wage inequality.

[15] See Berger, Herkenhoff, and Mongey 2022a for why weighting by the wage bill is appropriate.

[16] A market with M identically sized firms has an HHI of $HHI_M = \sum_i (1/M)^2 = 1/M$. Hence, an HHI of 0.09 is what one would obtain from a market with $1/0.09 \approx 11$ equally sized firms.

[17] These natural experiments can only be interpreted through a structural model, as Azkarate-Askasua and Zerecero (2022) adeptly do.

Let o denote the occupation, r the region (commuting zone), t the time (the data are monthly, and we denote the corresponding year as $\tau(t)$), and $m(o,r)$ the market to which the occupation–region was assigned by our algorithm[18]. Let γ_{FE} denote either (1) occupation–year fixed effects ($\gamma_{o\tau(t)}$), thus isolating across-region variation, or (2) occupation–region fixed effects γ_{or}, thus isolating across-time variation. Given our focus on market-level outcomes, we do not weight our regressions[19]. We estimate the following equation using ordinary least squares:

$$y_{o,r,t} = \gamma_{FE} + \beta HHI_{m(o,r),t} + X_{o,r,t} + \epsilon_{o,r,t} \tag{1}$$

We include a vector of controls, $X_{o,r,t}$, that vary at the occupation–region–time level. As we discuss in Section 4, our model removes mechanical variation in the number of firms per worker. We therefore control for lagged quintiles of firms per worker measured at the market–time level. We also control for month of the year to hold seasonal fluctuations fixed, lagged labor force growth and age, gender, and education composition at the *ort* level.

2.4 Empirical results

Across regions. Figure 2 provides a graphical representation of our regression evidence using across-region, within-occupation–year variation. The x-axis (Herfindahl index) and y-axis (labor market outcome) variables are residualized on occupation–year fixed effects and the controls $X_{o,r,t}$. This leaves across-region variation (e.g., *Oslo* vs. *Bergen*, for dentists in 2008). We normalize the residualized Herfindahl value by its standard deviation and subtract its mean (i.e., convert it to a Z-score) to ease interpretation. We also perform inference by clustering standard errors at the market level.

We find a statistically significant negative relationship between employment-to-employment transition rates and the market Herfindahl index (Panel A). To interpret the relationships, we note that the unweighted employment Herfindahl index has a mean of 0.28 and a standard deviation of 0.27. The slope of the bin-scatter implies that a one-standard-deviation increase in the market employment Herfindahl index is associated with a 0.06-percentage-point reduction in the E-to-E rate ($= -0.00214 \times 0.27$), which corresponds to approximately 10 percent of the sample average E-to-E rate (see Table 2). Panels B and C show similar negative relationships for U-to-E rates and E-to-U rates. On net, these yield a negative relationship with the unemployment rate (Panel D). Below, we find that this negative unemployment–HHI relationship is not robust to the choice of fixed effects.

There is also a negative relationship between wages and the Herfindahl index (Panel E). This relationship is significant with occupation-year fixed effects alone, but insignificant with occupation-year fixed effects and controls. Nonetheless, with occupation-year fixed effects and controls, a one-standard-deviation increase in the market employment Herfindahl index is associated with a 0.27-percentage-point reduction in the wage ($=-.0104 \times 0.27 \times 100$). Last, Panel F illustrates a strong negative relationship between concentration and wage inequality (the standard deviation of log wages).

[18] Recall that our clustering approach has potentially clustered 4-digit occupations into different groupings in different regions due to heterogeneity across regions in the occupation-to-occupation flow matrix. Hence, our unit of analysis is the occupation, but statistics such as concentration or wage inequality are measured at the level of the market (cluster–CZ) to which the occupation–CZ is assigned by our algorithm.

[19] Small concentrated markets vs. large less concentration markets is precisely the comparison that we want to study and thus should not be downweighted by labor force size.

One concern is that these covariances reflect the sorting of better workers into less concentrated markets. We think that this is unlikely for two reasons. First, we control for education level and show that the main conclusions remain unaltered if we omit these controls in Table 3. Second, high-skill workers are more likely to have high UE rates but lower EU rates. Nevertheless, we consider an alternative specification that isolates within-occupation–region across-time variation, thereby mitigating concerns regarding spatial sorting on unobservables.

Across time. Figure 3 repeats the exercise with both the Herfindahl index and labor market outcome residualized on occupation–region fixed effects and the controls X_{ort}. This leaves across-time variation (e.g., *2006* vs. *2007*, for dentists in Oslo). Again we observe the negative correlation between concentration and labor market flows (Panels A, B, and C), wages (Panel E) and wage inequality (Panel F). Here, however, the relationship with unemployment is flipped (Panel D). However, with occupation–region fixed effects, the relationship between wages and concentration is robustly negative, regardless of controls. Panel E implies that a one-standard-deviation increase in the market employment Herfindahl index is associated with a 4.5-percentage-point reduction in the wage ($= -0.166 \times 0.27 \times 100$).

Regressions. In a regression setting, the significant, negative relationship between concentration and all three worker flows and wage inequality is robust across all specifications (Table 3 A, B, C, and F). Table 3 provides the regression tables corresponding to Figures 2 and 3. We estimate (1) with and without controls and for both sets of fixed effects (occupation–year, denoted O-Y, and occupation–region, denoted O-R). The dependent variable in Column (1) is the monthly E-to-E transition rate (not expressed as a percent). The coefficient can be interpreted as follows: a one-standard-deviation increase in the HHI is associated with 0.05 percent ($= -0.00194 \times 0.27 \times 100$) reduction in the employment-to-employment transition rate. The relationship between concentration and unemployment rates is sometimes insignificant, negative, or positive depending on the fixed effects and inclusion of controls (Panel D). Log wages are significantly negatively related to concentration in all specifications except for that with occupation–year fixed effects with controls (Panel E).

Robustness to alternative labor market definitions. We explore the robustness of our empirical results to defining labor markets in alternative ways in Appendix C. Rather than clustering occupations by K-means using the occupational flow matrix within each commuting zone, we follow a recent literature (Lindenlaub and Postel-Vinay, 2021) and extract the relevant clusters of occupations within a CZ using modularity maximization following the work of Schmutte (2014). Reassuringly, the patterns that we find are similar to those obtained when we use our baseline clustering algorithm.

Lastly, an alternative approach to defining markets is to simply use raw 3-digit occupations and commuting zones. However, the self-flow rate would be 45 percent. We used this market definition in an earlier draft of this paper and found quantitatively similar results.

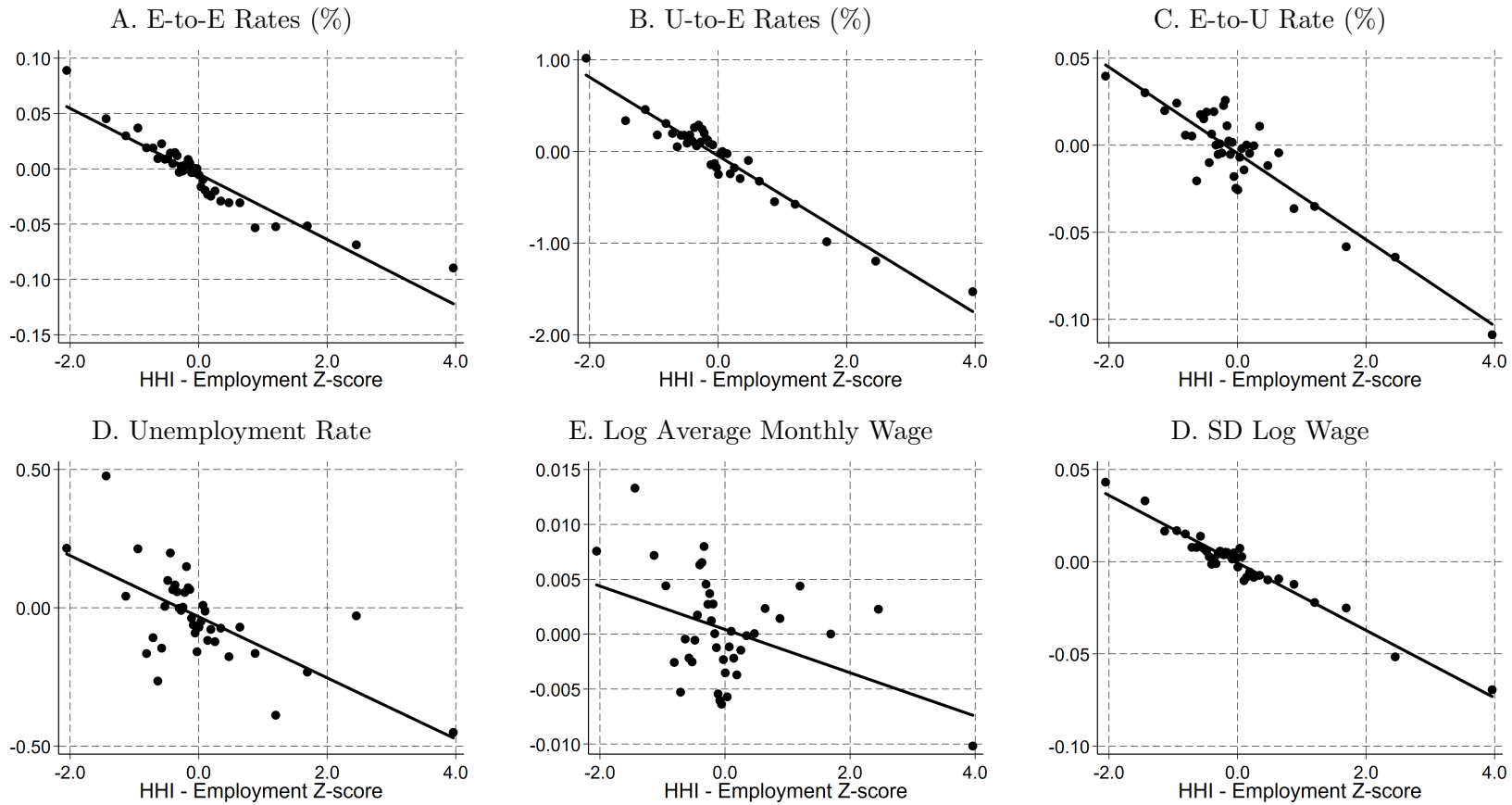


Figure 2: Concentration and labor market outcomes residualized on **occupation–year FEs**, leaving across-region variation

Note: For each market (where a market is defined, as in Section 2, as a cluster of occupations within a commuting zone), we compute the employment Herfindahl index (HHI). For each 4-digit occupation–commuting zone–year, we compute the average of the dependent variable within 40 centiles of the market HHI, unweighted. We then residualize all x and y variables on occupation–year fixed effects (FEs), age composition, gender composition, education composition, lagged firms-per-worker ventiles, lagged labor force growth, and month-of-year dummies. The average NOK/USD in 2021 was 9.

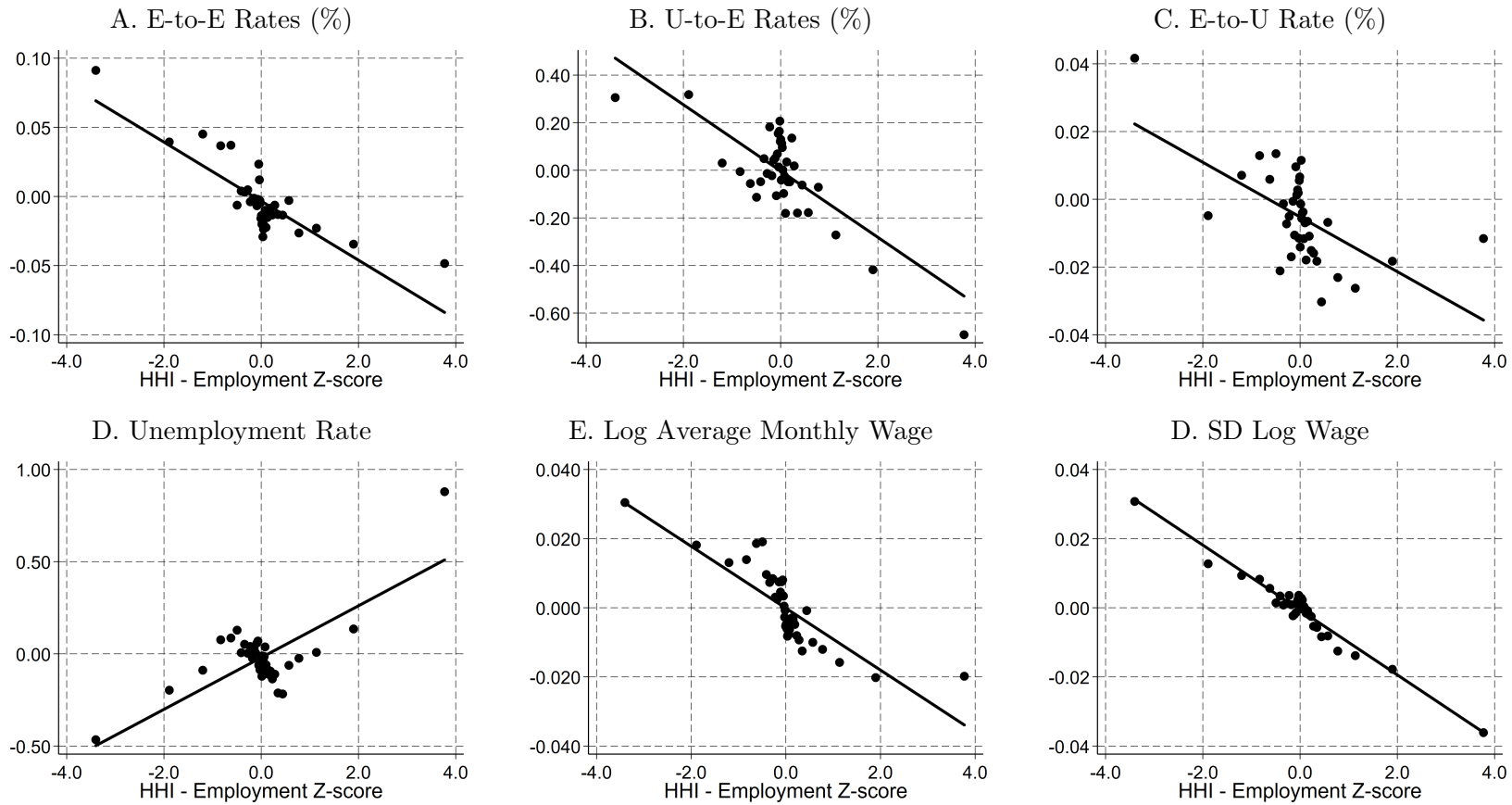


Figure 3: Concentration and labor market outcomes residualized on **occupation–region FEs**, leaving across-time variation

Note: For each market (where a market is defined, as in Section 2, as a cluster of occupations within a commuting zone) we compute the employment Herfindahl index (HHI). For each 4-digit occupation–commuting zone–year, we compute the average of the dependent variable within 40 centiles of the market HHI, unweighted. We then residualize all x and y variables on occupation–region fixed effects (FEs), age composition, gender composition, education composition, lagged firms-per-worker ventiles, lagged labor force growth, and month-of-year dummies. The average NOK/USD in 2021 was 9.

	A. EE rate				B. UE rate				C. EU rate			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
HHI	-0.00194***	-0.00214***	-0.00361***	-0.00475***	-0.0356***	-0.0330***	-0.0231***	-0.0284***	-0.00191***	-0.00177***	-0.000825**	-0.00172***
	(7.58e-05)	(9.11e-05)	(0.000232)	(0.000312)	(0.00208)	(0.00234)	(0.00443)	(0.00520)	(0.000154)	(0.000177)	(0.000325)	(0.000415)
FE	O-Y	O-Y	O-R	O-R	O-Y	O-Y	O-R	O-R	O-Y	O-Y	O-R	O-R
Controls	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y
Obs.	1,035,450	892,774	1,035,382	892,731	628,097	553,705	627,933	553,544	1,035,450	892,774	1,035,382	892,731
R^2	0.066	0.181	0.047	0.164	0.065	0.075	0.094	0.101	0.069	0.082	0.095	0.110

	D. Unemployment rate				E. Log wage				F. Standard deviation of log wage			
	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)
HHI	0.00251	-0.00781***	0.0631***	0.0420***	-0.0328***	-0.0104	-0.167***	-0.166***	-0.126***	-0.134***	-0.226***	-0.216***
	(0.00302)	(0.00298)	(0.00650)	(0.00668)	(0.00956)	(0.00970)	(0.0197)	(0.0211)	(0.00892)	(0.00960)	(0.0166)	(0.0198)
FE	O-Y	O-Y	O-R	O-R	O-Y	O-Y	O-R	O-R	O-Y	O-Y	O-R	O-R
Controls	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y
Obs.	1,043,861	892,927	1,043,790	892,885	1,043,861	892,927	1,043,790	892,885	1,042,160	892,176	1,042,089	892,134
R^2	0.408	0.425	0.509	0.528	0.793	0.817	0.784	0.825	0.382	0.402	0.518	0.537

Table 3: Regression analysis: Concentration and labor market outcomes

Note: Standard errors clustered at the market level. For the FEs, (i) O-Y refers to occupation–year fixed effects, and (ii) O-R refers to occupation–region fixed effects. For the controls, Y indicates controls for quintiles of firms-per-worker measured at the market–month level, month-of-the-year, lagged labor force growth, age, gender, and education composition at the *ort* level. In this table (unlike the figures), EE, UE, EU and Unemployment rates take values between 0 and 1 and are thus not expressed in percentage points.

Summary. In the rest of the paper, we ask whether a benchmark theory of frictional labor markets with on-the-job-search and bargaining (Cahuc, Postel-Vinay, and Robin, 2006) can quantitatively replicate these empirical relationships when extended to accommodate (i) concentrated markets, (ii) a vacancy-posting equilibrium, and (iii) worker–firm-specific tastes. We then use the model to decompose the role of different economic forces in determining wage markdowns, employment and inequality.

3 Model

In our model, time is discrete and runs forever. We assume that there are J markets indexed by $j \in \{1, \ldots, J\}$. Within each market, there are $\overline{N}_j$ workers and M_j firms. Firms are indexed by $i \in \{1, \ldots, M_j\}$. The measure of firms is exogenous, and workers are assumed to be immobile across markets. For the remainder of the section, we suppress the market index j.

Workers. Workers have linear utility and maximize the net present value of discounted utility. Both workers and firms discount the future at a rate β. Workers are either employed or unemployed. Among the employed, a measure n_i are employed at firm i. Total employment in a labor market is therefore $n = \sum_{i=1}^{M} n_i$, and thus, the measure of unemployed individuals in a labor market is $u = \overline{N} - n$.

A worker's per-period utility is the summation of income and a worker–firm-specific taste shock, ε. This taste shock is a stand-in for commuting times, how well a worker gets along with her boss and colleagues, and any other worker–firm-specific nonwage characteristics. We assume that the taste shock is independently drawn from the distribution $F(\varepsilon)$ when a worker contacts a firm, and we assume that ε is constant throughout a worker–firm match. Both the firm and worker observe and contract on the amenity draw. Thus, we allow firms to *first-degree* price discriminate on amenities[20].

Search is random. Each period, a random fraction ϕ of unemployed individuals search for job openings. Unemployed individuals receive utility from home production, b. Employed workers search on the job with probability $\xi\phi$ and do not apply to jobs at their existing firm. We refer to this search protocol as *partially directed search*.

Firms. Firm i's productivity is fixed and is denoted z_i. Posting v_i vacancies costs $c(v_i, M, \overline{N})$, where c is convex in vacancies. Empirically, unemployment rates vary relatively little across markets that vary widely in terms of the number of firms per worker (see Appendix Table 9). For the model to scale and achieve this stylized fact, we remove variation in firms per worker by scaling the vacancy costs accordingly. Anticipating the calibration, we assume that vacancy costs are given by $c(v_i, M, \overline{N}) = \left(\frac{M}{\overline{N}}\right)^{\gamma} \frac{1}{1+\gamma} v_i^{1+\gamma}$ The scaling factor achieves neutrality of the unemployment rate with respect to firms per worker $\left(M/\overline{N}\right)$. For convenience, we suppress the market employment and firm arguments of $c(\cdot)$ and write $c(v_i)$ for vacancy costs.

We assume that firms post vacancies nonstrategically. In an earlier version of this paper, we solved the strategic vacancy posting decision for markets with few firms and found that strategic vacancy posting motives yield no discernible effects on aggregates while simultaneously making the model less

[20]Note that Robinson (1933)'s class of models does not allow for wage discrimination based on amenity draws, thus yielding inefficient allocations. First-degree price discrimination, on the other hand, yields the efficient allocation, with zero consumer surplus and wages set below marginal products.

tractable. Strategic vacancy posting, when combined with amenities, also gives rise to complex hiring rules[21]. We assume away this behavior and leave it to future researchers to find tractable solutions to this problem.

Meeting rates. When workers apply for jobs, only a fraction of those applications actually result in a meeting with a prospective employer via a vacancy. After meeting, the worker and prospective firm observe the nonwage amenity draw ε upon which they base their matching decision. A match occurs if the worker moves to the prospective employer. Matches end through on-the-job-search as well as at an exogenous rate δ.

From the worker's perspective, let λ_{ik} denote the rate at which a firm i worker meets a firm k vacancy. From the firm's perspective, let λ_{ik}^{f} denote the rate at which firm i's workers meet a firm k vacancy. Our convention with subscripts is origin first (i), then destination (k).

To describe the meeting process, we must keep track of the origin and destination of job applicants. Let x_{ik} (x_{uk}) denote the measure of firm i workers (unemployed workers) who randomly apply for jobs at firm k. Then, $x_i = \sum_{k \neq i} x_{ki} + x_{ui}$ is the total measure of workers who randomly apply for jobs at i.

Throughout, we assume that meetings *at each firm* are governed by a constant-returns-to-scale meeting function $\overline{m}(v_i, x_i) \leq \min\{v_i, x_i\}$. In the absence of on-the-job search, firm-specific constant-returns-to-scale matching functions and the usual pooled random search model of Diamond–Mortensen–Pissarides are equivalent. Let A denote match efficiency. In practice, we use a Cobb–Douglas matching function:

$$\overline{m}(v, x) = \min\{A v^{\alpha-1} x^{\alpha}, 1\}$$

We let $f(\theta_i)$ denote the job-finding rate at firm i, where the tightness at firm i is defined to be $\theta_i = v_i / x_i$. Constant returns to scale imply $\overline{m}(v_i, x_i) = x_i f(\theta_i)$.

From each firm i, there is a measure $\xi \phi n_i$ of workers who engage in *partially directed* on-the-job search. Because employed workers randomly apply only for vacancies posted by firms other than the one where they are currently employed, the measure of workers at firm i who apply to firm k is given by

$$x_{ik} = \frac{v_k}{\sum_{j \neq i} v_j} \xi \phi n_i.$$

A fraction ϕ of unemployed individuals apply to all M firms randomly. Therefore, the measure of workers who are unemployed and apply to firm k is given by

$$x_{uk} = \frac{v_k}{\sum_{j} v_j} \phi u.$$

We can now derive the probability that a worker at firm i meets firm k into three terms. The first term is the probability that a worker searches, $\xi\phi$. The second term is the conditional probability that the worker applies to a vacancy at firm k, $v_k / \sum_{j \neq i} v_j$. The third term is the conditional probability that a meeting occurs, $\overline{m}(v_k, x_k)/x_k$. These yield the worker meeting rate:

$$\lambda_{ik} = \xi\phi \times \left(\frac{v_k}{\sum_{j \neq i} v_j} \right) \times \left(\frac{\overline{m}(v_k, x_k)}{x_k} \right) = \xi\phi \times \left(\frac{v_k}{\sum_{j \neq i} v_j} \right) \times f(\theta_k). \tag{2}$$

The contact rate of unemployed workers is defined similarly, except that unemployed workers may meet

[21] For example, a firm may turn down all hires with the lowest amenity draw so as to wait until it meets a worker with a better amenity draw, to whom it can pay lower wages via a compensating differential. In a market with 150 firms and 3 potential amenity draws, 3^{150}, such possible complex hiring rules exist.

any firm and, hence, $\lambda_{uk} = \phi(v_k/\sum_j v_j) f(\theta_k)$.

We similarly divide the probability of a meeting a firm k's vacancy with a worker at firm i into two terms: i) the probability of a meeting between a worker and a vacancy, $\overline{m}(v_k, x_k)/v_k$, and ii) the probability that the worker originated from firm i, x_{ik}/x_k. These yield the firm meeting rate:

$$\lambda_{ik}^f = \frac{\overline{m}\left(v_k, x_k\right)}{v_k} \times \left(\frac{x_{ik}}{x_k}\right). \tag{3}$$

The probability that firm k meets an unemployed worker is given by $\lambda_{uk}^f = \overline{m}(v_k, x_k)/v_k(x_{uk}/x_k)$.

Flow balance holds. Using the definitions of x_{ik} and x_k, one can check that $n_i \lambda_{ik} = v_k \lambda_{ik}^f$: firm i workers' rate of meeting firm k vacancies equals firm k's rate of meeting vacancies with firm i workers.

Bargaining over promised values. We follow Cahuc, Postel-Vinay, and Robin (2006). When a worker meets a new firm, the incumbent and poaching firm propose sequential offers. We assume that firms offer workers promised values that they are committed to. There are three possible cases: the worker meets a firm that can deliver a maximum promised value that is (i) less than the value promised to the worker by her current employer, (ii) greater than the value promised to the worker by her current employer but less than the maximum promised value of the worker's current employer, or (iii) greater than the maximum promised value that the worker's current employer can offer.

In case (i), we assume no change to the worker's wage, and the worker remains with the incumbent firm. In case (ii), we assume that the incumbent firm offers a new promised value that delivers the full joint value of the match with the poaching firm. In case (iii), the worker moves to the poaching firm, and Nash bargaining determines the split of the remaining surplus between the poaching firm and worker, where the full joint value of the match with the incumbent firm constitutes the worker's outside option. Let $\theta \in [0, 1]$ denote the worker's Nash bargaining parameter. Let $\sigma \in [\theta, 1]$ denote the worker's share of the match surplus. Because of cases (ii) and (iii), σ may increase above the worker's Nash bargaining weight θ.

Wage determination. The bargaining protocol pins down the promised values, but the wage that delivers the promised values is indeterminate. We assume that firms deliver the promised values to workers as a constant wage unless the worker receives an credible outside option. This is a common assumption and delivers wage dynamics consistent with the empirical evidence in Cahuc, Postel-Vinay, and Robin (2006), Jarosch (2014), and Lise and Robin (2017).

3.1 Forward-looking decisions

Let the continuation value of a worker at firm i with bargained surplus share σ and taste shock ε be given by $W_i(\sigma, \varepsilon)$. Likewise, let firm i's continuation value of a match with bargained surplus share σ and taste shock ε be given by $J_i(\sigma, \varepsilon)$. The continuation value of an unemployed individual is given by U. We will frequently work with both the joint value of a match, $P_i(\varepsilon) := W_i(\sigma, \varepsilon) + J_i(\sigma, \varepsilon)$, and the match surplus, $S_i(\varepsilon) := W_i(\sigma, \varepsilon) - U + J_i(\sigma, \varepsilon) \equiv P_i(\varepsilon) - U$. Because firms commit to the promised values and workers and firms have linear utility, it can be shown that the match surplus and joint value are independent of the division of surplus, σ.

We assume that once workers and firms separate, the job position is destroyed. To hire again, then,

a firm needs to post new vacancies. Hence, implicit in these definitions is that the value of an unfilled vacancy is zero, similar to the setup in Lise and Robin (2017).

Before we exposit the continuation values of the worker, we note that Appendix E provides a full derivation of the main equations in the text, including the joint value of a match, the surplus of a match, and the wage equation. Additionally, Appendix E shows how one may solve for surplus using a simple matrix inversion.

Unemployed workers. Unemployed workers enjoy home production b and meet with firm k with probability λ_{uk}. When they meet with firm k, they draw a taste $\varepsilon' \sim F$ for working at firm k. They receive a share θ of surplus. The continuation value of an unemployed worker is therefore

$$U = b + \beta \left[U + \theta \int \sum_k \lambda_{uk} \max \left\{ S_k \left(\varepsilon' \right), 0 \right\} dF \left(\varepsilon' \right) \right].$$

Employed workers. The worker value is the value of unemployment plus some share σ of the match surplus (following Lise and Postel-Vinay, 2020). In other words, σ is defined to be the number that satisfies the following equation:

$$W_i \left(\sigma, \varepsilon \right) = U + \sigma \Big[P_i(\varepsilon) - U \Big]$$

The Cahuc, Postel-Vinay, and Robin (2006) bargaining protocol implies that when a worker at firm i with amenity ε meets a vacancy at firm k and draws amenity ε', there are three possible outcomes:

(i) If $P_i(\varepsilon) > W_i(\sigma, \varepsilon) > P_k(\varepsilon')$, the worker stays at firm i with promised value $W_i(\sigma, \varepsilon)$.

(ii) If $P_i(\varepsilon) > P_k(\varepsilon') > W_i(\sigma, \varepsilon)$, the worker stays at firm i but is now delivered a promised value $W_i(\sigma', \varepsilon) = P_k(\varepsilon')$, where $\sigma' = S_k(\varepsilon') / S_i(\varepsilon)$.

(iii) If $P_k(\varepsilon') > P_i(\varepsilon)$, the worker moves to firm k and Nash bargains over the gains from trade $(P_k \left(\varepsilon' \right) - P_i \left(\varepsilon \right))$, with the full joint value at firm i, $P_i(\varepsilon)$, as her outside option.

Under this protocol, the worker policy function is to move to the firm with the greatest surplus. Note that workers may move down the productivity ladder if the amenity draw increases surplus above that associated with the incumbent firm.

In case (iii), the Nash bargaining solution delivers a worker continuation value that maximizes:

$$\max_{\widehat{W}} \left(P_k(\varepsilon') - \widehat{W} \right)^{1-\theta} \left(\widehat{W} - P_i(\varepsilon) \right)^{\theta}$$

The resulting promised value is equal to the entire joint value between the worker and firm i plus a fraction θ of the gains from trade:

$$\widehat{W} = P_i \left(\varepsilon \right) + \theta \left[P_k \left(\varepsilon' \right) - P_i \left(\varepsilon \right) \right],$$

which is convenient to express as a fraction of the match surplus:

$$W_k(\sigma', \varepsilon') = \widehat{W} = U + \sigma' [P_k \left(\varepsilon' \right) - U], \quad \text{where} \quad \sigma' = \theta + (1 - \theta) \frac{S_i(\varepsilon)}{S_k(\varepsilon')}$$

Given the above, it can be verified that the worker's share of surplus evolves according to:

$$\sigma' = \begin{cases} \left(\frac{\theta S_k(\varepsilon') + (1-\theta) S_i(\varepsilon)}{S_k(\varepsilon')} \right) & \text{if } S_k(\varepsilon') > S_i(\varepsilon) \\ \max \left\{ \sigma, \frac{S_k(\varepsilon')}{S_i(\varepsilon)} \right\} & \text{if } S_k(\varepsilon') \leq S_i(\varepsilon) \end{cases}$$

As discussed above, we assume that the promised values are delivered as a constant wage $w_i(\sigma,\varepsilon)$ unless the worker has a meeting with an employer that triggers renegotiation. Employed workers at firm i meet with firm k with probability λ_{ik}. Under the bargaining protocol of Cahuc, Postel-Vinay, and Robin (2006), the continuation value of the worker can be written as (see Appendix E):

$$\begin{aligned} W_i(\sigma,\varepsilon) &= w_i(\sigma,\varepsilon)+\varepsilon+\beta\Big[W_i(\sigma,\varepsilon)-\delta\sigma S_i(\varepsilon) \\ &\quad +\int\sum_{k\neq i}\lambda_{ik}\max\{0,\min\{\theta[S_k(\varepsilon')-S_i(\varepsilon)],S_k(\varepsilon')-S_i(\varepsilon)\}+(1-\sigma)S_i(\varepsilon)\}dF(\varepsilon')\Big]. \end{aligned} \tag{4}$$

Joint value. Rather than exposit the value of a firm directly, we focus on the joint value of a match between worker and firm i with taste shock ε. They jointly produce z_i unless (i) the worker receives an outside offer at a firm that generates greater surplus or (ii) the match exogenously dissolves. Thus, the joint value takes into account the worker's future value of a new match or unemployment:

$$P_i(\varepsilon)=z_i+\varepsilon+\beta\left[P_i(\varepsilon)+\theta\int\sum_{k\neq i}\lambda_{ik}\max\{S_k(\varepsilon')-S_i(\varepsilon),0\}dF(\varepsilon')-\delta S_i(\varepsilon)\right]$$

Surplus. The surplus of a match relative to unemployment can be expressed similarly. The costs include the flow value of unemployment, b, and the option value of unemployment forfeited by being employed at i: $\theta\int\sum_k\lambda_{uk}\max\{S_k(\varepsilon''),0\}dF(\varepsilon'')$. The benefits are production, amenities, and potential gains from on-the-job search, $\theta\int\sum_{k\neq i}\lambda_{ik}\max\{S_k(\varepsilon')-S_i(\varepsilon),0\}dF(\varepsilon')$:

$$\begin{aligned} S_i(\varepsilon) &= \Big(z_i+\varepsilon\Big)-b+\beta\Big[(1-\delta)S_i(\varepsilon)+\theta\int\sum_{k\neq i}\lambda_{ik}\max\{S_k(\varepsilon')-S_i(\varepsilon),0\}dF(\varepsilon') \\ &\quad -\theta\int\sum_k\lambda_{uk}\max\{S_k(\varepsilon''),0\}dF(\varepsilon'')\Big]. \end{aligned} \tag{5}$$

Wage equation. Combining the worker's value (4) and surplus (5), we can compute the wage based on surplus values alone. The wage function $w_i(\sigma,\varepsilon)$ delivers a surplus share σ at firm i:

$$\begin{aligned} w_i(\sigma,\varepsilon) &= \sigma z_i-(1-\sigma)\varepsilon+(1-\sigma)\left[b+\beta\theta\int\sum_k\lambda_{uk}\max\{S_k(\varepsilon'_u),0\}dF(\varepsilon'_u)\right] \\ &\quad -\beta\int\sum_{k\neq i}\lambda_{ik}\max\left\{0,\min\left\{(1-\sigma)\theta(S_k(\varepsilon')-S_i(\varepsilon)),(S_k(\varepsilon')-S_i(\varepsilon))\right\}+(1-\sigma)S_i(\varepsilon)\right\}dF(\varepsilon') \end{aligned} \tag{6}$$

The wage equation includes four terms: (i) workers obtain σ of production, (ii) the firm can offer a lower wage to workers with higher taste shocks to deliver any given promised value, (iii) workers obtain $(1-\sigma)$ of their outside option, and last, (iv) there is backloading since firms that offer greater future pay prospects can initially pay less.

Optimal vacancy posting. The firm vacancy posting problem requires knowledge of the distribution of workers across amenity values and employers. The probability that a worker at firm k has amenity draw ε is given by the endogenous ratio $n_k(\varepsilon)/n_k$. As discussed above, we assume that the vacancy posting decision is nonstrategic. Thus, the firm chooses v_i to maximize the following objective, taking

all contact rates, worker stocks and surplus values as given[22]:

$$\max_{v_i} \; -c(v_i) + \underbrace{(1-\theta)v_i \int \lambda_{ui}^f \max\Big\{S_i(\varepsilon'),0\Big\} dF(\varepsilon')}_{\text{Hire from unemployment}} + \underbrace{(1-\theta)v_i \iint \sum_{k\neq i} \lambda_{ki}^f \left(\frac{n_k(\varepsilon)}{n_k}\right) \max\Big\{S_i(\varepsilon') - S_k(\varepsilon),0\Big\}\, d\varepsilon\, dF(\varepsilon')}_{\text{Hire from employment}} \tag{7}$$

This yields the following optimality condition for firms:

$$v_i = c'^{-1}\left((1-\theta)\int \lambda_{ui}^f \max\Big\{S_i(\varepsilon'),0\Big\} dF(\varepsilon') + (1-\theta)\iint \sum_{k\neq i} \lambda_{ki}^f \left(\frac{n_k(\varepsilon)}{n_k}\right) \max\Big\{S_i(\varepsilon') - S_k(\varepsilon),0\Big\} d\varepsilon dF(\varepsilon')\right)$$

Laws of motion for employment. The laws of motion for employment and unemployment are given by the following equations, where primes denotes next-period values[23]:

$$\begin{aligned} n_i'(\varepsilon) &= \left(1-\delta-\sum_{k\neq i}\lambda_{ik}\int \mathbf{1}_{[S_k(\varepsilon')\geq S_i(\varepsilon)]} dF(\varepsilon')\right) n_i(\varepsilon) + \lambda_{ui}\mathbf{1}_{[S_i(\varepsilon)>U]} f(\varepsilon) u + \sum_{k\neq i}\lambda_{ki}\int \mathbf{1}_{[S_i(\varepsilon)\geq S_k(\varepsilon')]} f(\varepsilon) n_k(\varepsilon') d\varepsilon' \\ u' &= \left(1-\sum_i \lambda_{ui}\int \mathbf{1}_{[S_i(\varepsilon')\geq U]} dF(\varepsilon')\right) u + \delta\Big(\overline{N}-u\Big) \quad , \quad n_i = \int n_i(\varepsilon)\, d\varepsilon \quad , \quad u = \overline{N} - \sum_i n_i. \end{aligned} \tag{8}$$

Equilibrium. Since markets do not interact, it suffices to define the equilibrium for a single market. We continue to suppress the market j index, and note that in the quantitative model, all J markets satisfy the following equilibrium definition.

In a given market with a mass of firms M and labor force $\overline{N}$, a *stationary equilibrium* is a stock of vacancies and employment $\{v_i, n_i(\varepsilon)\}_{i=1}^M$, an unemployed value U, surplus values $\{S_i(\varepsilon)\}_{i=1}^M$ (which implicitly define the worker's mobility policy function), meeting rates $\{\lambda_{ui}^f, \lambda_{ui}\}_{i=1}^M$ and $\{\{\lambda_{ki}^f, \lambda_{ik}\}_{k\neq i}\}_{i=1}^M$ such that

1. *Worker optimality*: Given surpluses and contact rates, worker mobility decisions are optimal (i.e., mobility decisions are consistent with surpluses (5) and delivers the worker value (4)).
2. *Firm optimality*: Given surpluses, contact rates, and worker stocks, v_i solves (7).
3. *Market clearing*: Worker mobility decisions (implicitly defined by $\{S_i(\varepsilon)\}_{i=1}^M$ and U) and optimal firm vacancy postings deliver a stationary distribution of workers given by equation (8) consistent with $\{n_i(\varepsilon)\}_{i=1}^M$.

4 Calibration and model fit

This section describes our calibration approach. We then explore the model fit relative to the data moments in Section 2.

[22] Specifically, for example, firm i does not internalize that its vacancies affect the contact rates of workers at firm k and hence affect the surplus $S_k(\varepsilon')$, which then affects the cost of hiring a worker from firm k.

[23] Note that search efficiency is accounted for by λ_{ik} and λ_{ui} and that the density of $F(\varepsilon)$ is denoted $f(\varepsilon)$.

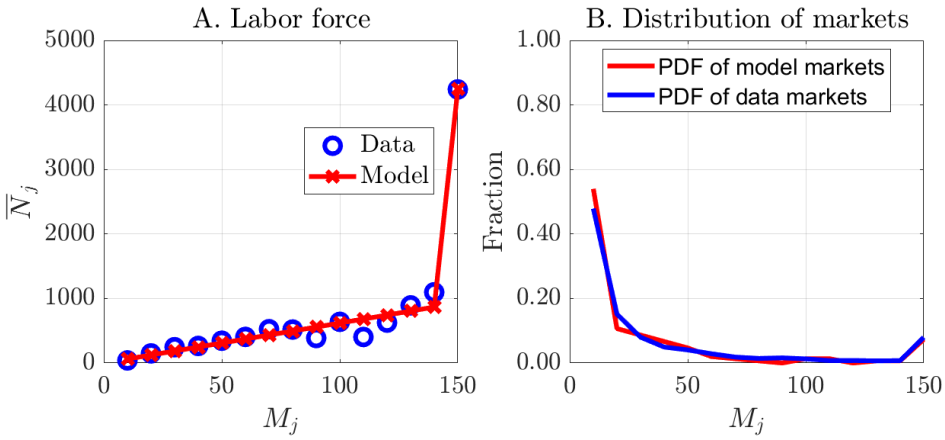


Figure 4: Calibration of market size and labor force

Notes: Panel A plots the relationship between labor force size $\overline{N}_j$ and market size M_j. Panel B plots the probability density function of market sizes in the model vs. the data.

4.1 Calibration

Markets. We adopt the clustering method described in Section 2 to define markets. With these market definitions in hand, we can compute the labor force per market, $\overline{N}_j$, and the distribution of the number of firms per market, $G(M_j)$. Rather than parameterize the joint distribution of M_j and $\overline{N}_j$, we note that a linear relationship between labor force and firms per market fits the data quite well (see Figure 4 and the discussion in the following paragraph). Thus, we impose $\overline{N}_j = a\overline{M}_j$ so that once we draw the market size M_j, we know $\overline{N}_j$. We truncate the data at $M_j = 150$, above which we find very little difference with respect to equilibrium as we raise M_j. At the final truncated market size, for the labor force to add up to that of Norway, we deviate from the linear relationship and simply compute the necessary $\overline{N}_j$ to match the size of the Norwegian labor force (in particular, the characteristic that 70 percent of workers are in $M_j = 150$ markets).

Figure 4 A illustrates the relationship between $\overline{N}_j$ and M_j, as well as our fitted values. We assume there are $J = 200$ markets (despite there being 3,700 in our data), and for each market, we draw the number of firms $M_j \sim G(M_j)$ where $G(\cdot)$ is the empirical cumulative distribution function (CDF) of firms per market. Figure 4 B plots the distribution $G(M_j)$ in the data and in our $J = 200$ markets, illustrating the limited Monte Carlo simulation error in our quantitative experiments.

Preferences and technology. A period is one month. On an annual basis, the discount rate is 4 percent ($\beta = 0.96^{1/12}$). We assume that the matching function elasticity is given by $\alpha = 0.50$ and that unemployed workers search every period, $\phi = 1$. Employed workers' intensity of on-the-job search $\xi = 0.32$ delivers the job-to-job transition rate shown in Table 2. The match efficiency parameter $A = 0.18$ then delivers the unemployment rate. The home production parameter $b = 0.85$ is estimated to yield a 66 percent replacement rate as in Kostol (2017).

We calibrate the amenity distribution to match the fraction of employment-to-employment moves that occur down the poaching index ladder. We assume that amenities are distributed uniformly $\varepsilon \sim U[0, \bar{\varepsilon}]$

Parameter		Value	Moment	Model	Data
Match efficiency	A	0.18	Unemployment rate	0.03	0.04
OJS intensity	ξ_e	0.32	Aggregate EE rate	0.01	0.01
Home production	b	0.85	Replacement rate	0.86	0.66
Vacancy cost elasticity	γ	1.16	Employment HHI	0.09	0.08
Bargaining power	θ	0.18	Average log wage growth	0.01	0.01
Upper bound amenity	$\overline{\varepsilon}$	0.76	Fraction of EE moves down poach index ladder	0.21	0.15
Standard deviation of productivity	σ_z	0.14	Standard deviation of log wages	0.69	0.63
			Regression β: Log market wage on HHI	-0.09	-0.09
			Regression β: Standard deviation of log wages on HHI	-0.18	-0.18

Table 4: Calibration

and calibrate $\overline{\varepsilon} = 0.76$. As discussed in Section 2, the poaching index is simply a firm's share of hires who are poached from competitors rather than hired from unemployment[24]. We construct the same index in our model to map this moment to the data. If $\overline{\varepsilon} = 0$, so that there were no amenities, then *all* job-to-job transitions would be to firms with a higher poaching index. Through the lens of our model, the fact that only 85 percent of moves are up this ladder provides evidence for idiosyncratic tastes and disciplines $\overline{\varepsilon}$. The bargaining power $\theta = 0.18$ is calibrated to match the average wage growth rate in Norway[25].

The vacancy cost convexity parameter $\gamma = 1.16$ is estimated to match the employment Herfindahl index. Intuitively, greater values of γ compress vacancy postings at the most productive firms, reducing dispersion in employment. Last, we assume that the dispersion of firm productivity is log normal: $\log z \sim N\left(-\frac{1}{2}\sigma_z^2, \sigma_z\right)$. The standard deviation $\sigma_z = 0.14$ is estimated to match the standard deviation of log wages.

To further discipline the parameters that govern concentration (γ, σ_z, b, θ) and its relationship with wages, we also target the regression coefficients of wages and the standard deviation of wages on the Herfindahl values in Table 3 (columns (18), (20), (22), and (24)). Thus, our estimation is overidentified. We take a simple average of the regression coefficients based on occupation–year and occupation–region fixed effects. The results of the estimation are shown in Table 4 along with the corresponding data moment that identifies each parameter.

4.2 Model fit

Table 5 compares the model's fit to the remaining reduced-form estimates from Section 2. The first two rows are the wage moments that we explicitly target. These coefficients are negative, but this is not by construction. The model can generate a counterfactually positive relationship between concentration and wages. How? With a sufficiently high productivity dispersion, the most concentrated markets are

[24] See Bagger and Lentz (2019) Table 2 for more detail.

[25] In this class of models, low values of θ can deliver negative wages, or near-zero wages, generating what look like fat-tailed wage growth distributions (e.g., in reasonable calibrations, workers can move from wages of 0.001 to a wage of 1, for instance, yielding 1,000 percent growth rates). Higher values of θ remove negative and/or near-zero wages, thus bringing the wage growth rates in line with the data.

Dependent variable	**Targeted in**	**Model**	**Data**	
Y_{ort}	**estimation**		Occ–year FE	Occ–region FE
		(1)	(2)	(3)
Log wage	Yes – Target average of (2) and (3)	-0.0915	-0.0104	-0.166***
Standard deviation of log wage	Yes – Target average of (2) and (3)	-0.1799	-0.125***	-0.230***
E-to-E rate	No	-0.0027	-0.00214***	-0.00475***
U-to-E rate	No	-0.0103	-0.0330***	-0.0284***

Table 5: Model fit relative to regression estimates in Section 2

Notes: Regression estimates taken from Table 3. Log wages correspond to columns (18) and (20). The standard deviation of wages corresponds to columns (22) and (24). EE rate estimates correspond to columns (2) and (4). EU rate estimates correspond to columns (6) and (8).

the markets in which one firm has drawn an outlier draw of productivity, z_i. Workers bargain over a share of surplus that is proportional to z_i, and thus concentrated markets can offer higher wages.

In terms of nontargeted moments, the model naturally generates the negative relationship between employment-to-employment transition rates and concentration. In the extreme, a market with a solo monopsonist $N = 1$ has zero employer-to-employer job transitions. The model also generates the negative observed relationship between job-finding rates and concentration. Note that the model can generate a counterfactually positive relationship between concentration and the job finding rate. If concentrated markets are also the most productive (i.e., a firm in such a market drew an extremely high z_i), then surplus and vacancy postings reflect the high surplus value, and hence, high job-finding rates and high concentration occur simultaneously. That the model correctly generates the right negative relationship between the UE rates and Herfindahl values is thus a positive contribution of the model.

In summary, we have a quantitative model that matches the cross-sectional empirical relationship between concentration and (i) wages, (ii) worker flows, and (iii) wage inequality.

5 Wage decomposition and mechanisms

Before we discuss our model counterfactuals, we first provide a decomposition of wages. Rewriting equation (6), we can express the wage as the sum of four components: the output share, the opportunity cost, the amenity discount, and the quit/promotion discount. In our calibrated steady state, the sum of the output share components over the sum of total wages is 91.4 percent. The opportunity cost is approximately half as important, and the amenity discount and quit/promotion discounts lower wages

by approximately 20 percent. We summarize these results below:

$$
\begin{aligned}
w_k(\sigma,\varepsilon) \quad = \quad & \qquad (9)\\
\text{1. Output share (91.4\%)} \qquad & \sigma z_k \\
\text{2. Opportunity cost (42.0\%)} \qquad & + (1-\sigma)\left(b + \beta\theta \sum_{k'}^{M} \lambda_{uk'} \int \max\left\{S_{k'}(\varepsilon'), 0\right\} dF(\varepsilon')\right) \\
\text{3. Amenity discount (-18.8\%)} \qquad & -(1-\sigma)\varepsilon \\
\text{4. Quit/promotion discount (-14.6\%)} \qquad & - (1-\sigma)\beta \int \sum_{k' \neq k}^{M} \lambda_{kk'} \mathbf{1}_{[S'>S]} \Big[S_k(\varepsilon) + \theta\Big(S_{k'}(\varepsilon') - S_k(\varepsilon)\Big)\Big] dF(\varepsilon') \\
& - \beta \int \sum_{k' \neq k}^{M} \lambda_{kk'} \mathbf{1}_{[\sigma S \leq S' < S]} \Big[S_{k'}(\varepsilon') - \sigma S_k(\varepsilon)\Big] dF(\varepsilon').
\end{aligned}
$$

The wage equation provides insight into how wage concentration, duopsonistic bargaining, and search frictions affect wages. First, concentration directly limits the surplus share σ. The finiteness of the number of firms implies that σ is bounded below one at the highest productivity–amenity match (i.e., the highest-ranked/highest-surplus firms). Consider the economy without amenities, and denote the highest attainable surplus share at the most productive firm $\overline{\sigma}_1$. In a single-firm market, workers coming out of unemployment can meet only the most productive firm, and thus, $\overline{\sigma}_1 = \theta < 1$. With two firms in a market where the $k = 1$ firm is the most productive, the highest possible share of surplus is $\overline{\sigma}_1 = \frac{\theta S_1 + (1-\theta) S_2}{S_1}$, which occurs when a worker at the lower-ranked firm meets the top-ranked firm via on-the-job-search. A similar argument holds with amenities.

Concentration also manifests itself through future contact rates. Similar to Jarosch, Nimcsik, and Sorkin (2019), we assume that workers cannot meet their current employer[26]. Hence, for a worker at firm k, the quit/promotion discount sums only over $k' \neq k$. Without nonwage amenities, this does not affect the surplus value but does affect the split of the surplus. With nonwage amenities, this affects the surplus value since the worker is restricted from drawing a new amenity value at her current employer (i.e., working for a new boss, changing departments), potentially creating surplus. Therefore, firms' ability to exclude themselves from future job-to-job transitions generates lower total surplus and thus lowers wages.

Neoclassical monopsony forces working through nonwage amenities contribute negatively to wages and thus drive a gap between a worker's wage and marginal product due to compensating differentials. Note that when $\sigma = 1$, workers are paid all the way up to their output z_i, and the amenity discount disappears. That is, a worker could obtain a high amenity value from working at firm i, but in a dynamic labor market with a high rate of outside offers, σ would quickly increase, and this idiosyncratic taste would no longer reduce pay. Hence, search frictions and amenities interact; furthermore, the mere finiteness of firms in the market bounds $\overline{\sigma}_1 < 1$, generating positive amenity discounts.

Given the wage-setting protocol in Cahuc, Postel-Vinay, and Robin (2006), the worker's bargaining power θ limits wage payments. Perhaps more subtly, the duopsonistic bargaining protocol—i.e., only

[26]Note that Jarosch, Nimcsik, and Sorkin (2019) further exclude the unemployed from meeting their former employer.

two firms ever compete simultaneously for a worker's services—also lowers the share of worker surplus σ. We explain this more in detail below when we allow more than two firms to bargain for a worker's services.

6 Counterfactuals

We now use the model to investigate a number of counterfactuals that isolate the roles of (1) concentration, (2) amenities, and (3) search frictions for wage markdowns, welfare, and inequality. As discussed, there are many facets of concentration. Concentration affects the set of possible meetings, attainable surplus shares, and outside options. In what follows, we attempt to thoroughly explore the various dimensions of concentration, as how we model the granularity of firms is arguably the most novel aspect of our framework. We then proceed to isolate the effects of amenity dispersion and search frictions on wages and welfare.

6.1 Isolating the effect of the number of firms, M

A. An ideal experiment. Our first counterfactual exercise aims to isolate the role of firms per market, M. We first consider an idealized experiment in which we solve an economy with ten identical 10-firm markets—i.e., with the same vector of productivities in each market—and then combine these into two identical 50-firm markets and, finally, one 100-firm market. When we combine markets, we combine the labor force as well, ensuring that $M/\overline{N}$ is constant across exercises, thus removing any mechanical changes to firms per worker. The vector of productivities $\boldsymbol{z} = (z_1, \ldots, z_{10})$ that we consider is evenly spaced between the 10^{th} and 90^{th} percentiles of the ergodic distribution of z_i. When we combine five of the 10-firm markets to produce a single 50-firm market, our exercise keeps the rungs of the productivity ladder fixed but reduces concentration.

Figure 5 A uses equation (9) to plot the effect of M on the average wage and its four components. We plot percent changes relative to the 10-firm benchmark, reducing concentration as we move from left to right. We find that average wages increase by 2 percent (black). As M increases from 10 to 100, the output share component increases by 4 percent, being responsible for more than all of the gains (blue). The output share component consists of $\sigma \times z$, and both increase. The share parameter σ increases by 3 percent alone due to a higher inflow of outside offers. The average productivity level z increases as workers flow to higher z firms.

When shares of surplus are higher, the opportunity cost of employment is lower, which reduces the wage (green). Notably, this nearly offsets all gains due to reallocation to higher-productivity firms (blue). Once these offsetting effects are accounted for, the dominant force in increasing the wage is the change in the amenity discount. As the worker's surplus share increases due to more competitive outside offers in a denser labor market, the reduction in wages due to worker–firm-specific tastes is reduced. Wage discrimination on the basis of idiosyncratic factors is impossible when competition is tight, as workers receive more and better outside offers. The wage penalty due to amenities declines, pushing up wages by 2 percent (red).

With more competition, workers are now more likely to meet with higher z firms in the future, which

leads to more backloading of wages (purple). In the 50-firm and 100-firm markets, workers at z_{10} can meet with another z_{10} firm that they might like more due to personal preferences. As the worker will gain some surplus from that later transition, a lower wage is required today to deliver the promised values. Quantitatively, we find that this effect is small, contributing only a small negative effect on wages as we increase the number of firms per market by a factor of ten.

Figure 5 B plots four other moments of interest, including wage inequality, welfare, and markdowns. First, wage inequality rises as we move from the 10-firm market to the 100-firm market. Workers at the top of the ladder now have many more possibilities in terms of competitive outside options, leading to higher wages to retain workers. Lucky workers receive many outside offers from good firms, fanning out the wage distribution, and increasing wage inequality.

Total welfare—the sum of household utility and firm profits net of vacancy posting costs—rises by 0.75 percent. As worker surplus shares rise, firm profits net of vacancy costs fall. As a result, household welfare, ignoring firm profits, rises by over 1 percent.

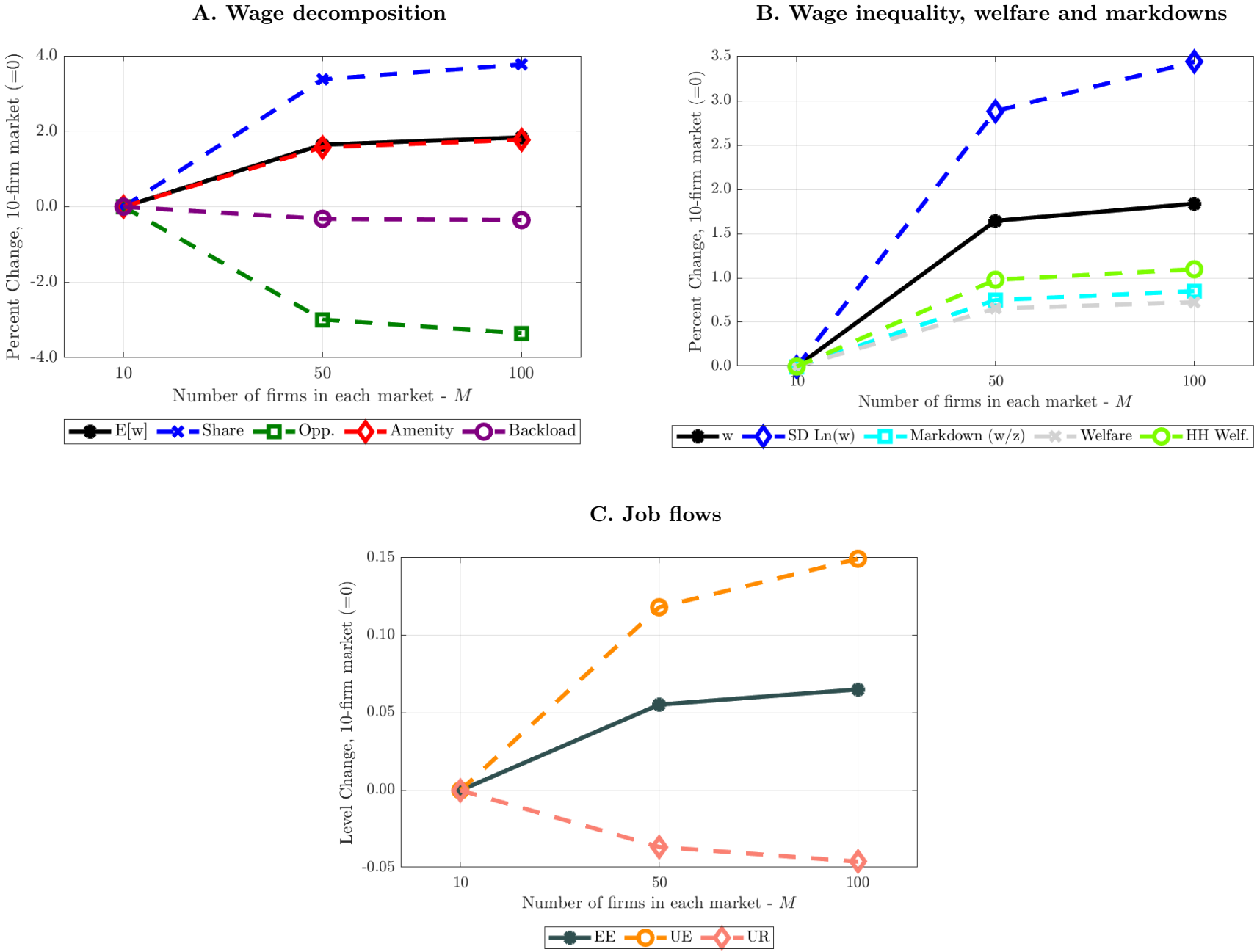
A central focus of the literature on monopsony is the static wage markdown (e.g., Robinson (1933) and Berger, Herkenhoff, and Mongey (2022a), among many others). In static economies, the definition of the markdown is universally agreed upon in the literature—it is the ratio of the spot wage to the marginal revenue product of the worker. In the dynamic context, markdowns can be defined in a number of ways. For example, the markdown could be logically defined as σ or perhaps via a comparison of the net present value of wages with the net present value of productivity. To stay as close as possible to the literature, we define markdowns following the static definition, here defined at the worker–firm level:

$$\textit{Markdown defintion:} \qquad \mu := w/z$$

Thus, in our dynamic setting, μ reflects a variety of forces, including imperfect competition and its interaction with amenities, bargaining, and backloading. As a conceptual point, we are able to decompose what researchers would estimate in static models into its various components including backloading, which may not necessary reflect noncompetitive behavior. We aggregate by taking the average across matches.

Figure 5 B shows that markdowns narrow, with workers' spot wages relative to their marginal product increasing by 1 percent in the 100-firm market over the wages the 10-firm markets. Importantly, the majority of the reductions in markdowns occur between the ten 10-firm markets and the two 50-firm markets. There is very little impact on markdowns between the two 50-firm markets and the single 100-firm market. This foreshadows our counterfactuals in Norway. The bulk of employment in Norway is in markets with 150 or more firms. Doubling the number of firms in an $M_j = 150$ firm market does little to markdowns, with the rungs of the productivity ladder held fixed. Conversely, markdowns are extremely sensitive to the number of firms in highly concentrated markets.

Figure 5: Altering M in isolation: Ideal experiment



Last, Figure 5 C shows the increase in mobility that underpins the improvement in workers' shares of output. The effect of the ability to meet with a same-rank firm of higher personal preference is that job-to-job transition rates rise (orange). A knock-on effect is that, in equilibrium, poaching workers becomes easier, incentivizing vacancy posting. Higher vacancies per worker support an increase in the U-to-E rate (black) and hence a fall in the unemployment rate (pink). Importantly, with endogenous recruiting effort—here via vacancies—it is not only the split of surplus that is impacted by the density of firms in the market but also quantities.

B. Application to Norway. We now extend the idealized experiment to the full Norwegian economy. Limited by computational resources, we conduct the following experiment: we duplicate the productivity vector in every simulated Norwegian market while simultaneously doubling the number of workers in each market. As in the ideal experiment, this preserves the rungs of the job ladder (i.e., the vector of z's is duplicated but not altered otherwise), and the number of firms per worker remains the same. Table 6 column (A) contains statistics for our baseline economy, while column (B) considers the counterfactual economy. With the doubling of M in each market via duplication, the Herfindahl value is approximately halved from 0.09 to 0.05, which is mostly mechanical. As tighter competition leads workers to obtain a greater surplus share, the wage markdown narrows by 1 percentage point. The employer-to-employer job transition rate increases marginally, while labor productivity improves by 0.42 percent relative to that in our baseline economy. Similarly to in the ideal experiment, welfare increases, the average wage increases, and wages become more dispersed. The welfare and labor productivity effects are nonnegligible: welfare increases by 0.25 percent. Throughout, the mechanisms are identical to those in the ideal experiment.

These results suggest a limited role of concentration in shaping Norwegian markdowns, but importantly, the exercise captures only one facet of concentration. There are two reasons that the doubling of M yields small effects on markdowns: (1) approximately 70 percent of the labor force resides in markets with more than 150 firms, and hence, doubling the number of firms in those markets is irrelevant (Figure 5 B), and (2) the duopsonistic wage setting remains unchanged when we double M (i.e., there are still only ever two firms at the bargaining table that are strategically setting wages). To dig deeper, we conduct two exercises designed to measure the effects of concentration and granularity operating through outside options and duopsonistic strategic wage setting.

6.2 Additional sources of monopsony power

A. Granularity. In our model, similar to the setups in Zhu (2012) and Jarosch, Nimcsik, and Sorkin (2019), firms are able to partially exclude themselves from future meetings when bargaining with workers. Such an assumption makes sense in a granular market. In our model, this is operationalized by a restriction that workers cannot meet with their current employer through job-to-job transitions in the next period. Jarosch, Nimcsik, and Sorkin (2019) go one step further and also exclude the current employer from the next period's possible U-to-E transitions during the bargaining process, which we do not consider.

Table 6 C reports what happens when we remove this assumption. Now workers redraw their amenity value at the firm. One can interpret this as life events that change the utility of a firm. Even if they

draw the same amenity value, the firm now bargains against itself. One can interpret this as a worker changing departments or bosses. We find that wage dispersion increases substantially. Since the firms may have to compete internally for workers, wage dispersion increases at the top of the ladder since surplus shares can now go all the way to one. Two offsetting forces limit the narrowing of markdowns, which narrow by only one percentage point. On the one hand, workers' ability to meet the same firm leads to higher surplus shares, especially at the top, which increases wages. On the other hand, their ability to redraw amenity values leads to large amenity wage penalties, which lower markdowns. The latter improves welfare, however, which increases by 0.57 percent.

B. Amenities. Similarly to Robinson (1933), we include a role for nonwage amenities in driving the employment and wage setting decisions of firms. Unlike Robinson (1933), however, we allow firms to first-degree price discriminate over the amenity value. This puts greater downward pressure on wages but, in many settings, does not necessarily result in inefficient allocations.

We find that our removing amenity dispersion yields a large reduction in the wage markdown. In Table 6 D, we restrict the amenity distribution to one value, calibrated to deliver the same aggregate amenity per capita as in our baseline economy. In other words, we study a mean-preserving removal of amenity dispersion. Workers' spot wages as a ratio to their productivity are now 7 percentage points greater than in the baseline economy. This represents a 33 percent narrowing of markdowns compared to the initial 21-percentage-point markdown observed in our baseline economy. With no heterogeneity in the amenity penalty, wage dispersion declines by a tenth relative to its level in the baseline economy.

Workers' sorting due to the combination of firm productivity and idiosyncratic tastes reduces the level of output in the economy relative to that in an economy where only firm productivity determines mobility. Recall that heterogeneity in amenities was necessary to ensure consistency with the empirical frequency of down-the-ladder job moves in the Norwegian data. In the economy in column (D), there are zero down-the-ladder job moves. Hence, when we remove amenities, workers are sorted across firms only on firm productivity. This causes aggregate productivity to increase substantially, by more than 7 percent.

C. Search frictions. We next explore the role of search frictions by steeply increasing match efficiency A[27]. In practice, our counterfactual implies that all employed and unemployed workers meet with a firm every period with certainty. Since our model is in discrete time, it remains the case that workers can at most meet with one firm per period. Hence, search frictions still exist, but they are strongly mitigated.

In an economy with on-the-job search, mitigating search frictions leads to a large increase in productivity (12.74 percent), as workers rapidly ascend the job ladder (E-to-E rates triple). Table 6 E shows that this is accompanied by a large increase in concentration as workers agglomerate at the most productive firms. The Herfindahl index increases by a factor of 8, from 0.09 to 0.75. When at the highest-ranked firm, workers now quickly bump into the next-highest-ranked firm, which bids up their surplus share. As a result, reducing search frictions leads to a substantial narrowing of the markdown,

[27] Numerically, we implement this by setting as A to 100, which is more than 500 times the level in our baseline (Table 4). We also set ξ_e to 1, so that unemployed and employed workers have the same search efficiency. Since A is so large, this is almost irrelevant, but we do it nonetheless for completeness.

Variable	A. Baseline $\{M_j, \overline{N}_j\}$	B. Duplicate $\boldsymbol{z}_j$ $2 \times \{M_j, \overline{N}_j\}$	C. Meet yourself	D. One ϵ $\frac{1}{N}\int \epsilon_i di$ fixed	E. No Search $\xi_e = 1$, $A = \infty$
HHI_n	0.09	0.05	0.12	0.28	0.75
Markdown (w/z), employment weighted	0.79	0.80	0.80	0.86	0.93
E-to-E worker flow rate (%)	0.55	0.56	0.62	0.63	1.17
Labor productivity vs. Baseline (%)	-	0.42	0.92	7.59	12.74
Welfare vs. Baseline (%)	-	0.25	0.57	6.60	14.75
Average wage vs. Baseline (%)	-	0.68	1.43	16.82	30.88
Standard deviation of log wage vs. Baseline (%)	-	0.81	2.05	-8.98	-91.66

Table 6: Counterfactuals in Norway

with wages now only 7 percent below marginal product. Markdowns remain positive for two reasons: i) the recuperation of vacancy costs for firms that still operate and ii) the fact that, with granular firms, some workers at the top of the productivity ladder are employed at an amenity discount and run out of potential outside offers to increase their wage. Both the sorting to higher z firms and greater surplus share contribute to a 30 percent increase in wages; meanwhile, inequality compresses dramatically, as workers end up in approximately the same boat relatively quickly.

D. Bargaining. Our final set of counterfactuals addresses a source of market power that is implicit in the assumptions of this framework: duopsony in the bargaining protocol. What if a worker could instead have three firms at the negotiating table? Or four? It is beyond the scope of this paper to solve a model with these features; however, we think that future progress in this area is important. The aim of our final set of counterfactuals is to show that work in this area may be of future interest.

As a first step, we consider a counterfactual bargaining protocol that involves three firms at the table. First, we abstract from amenities, which gives an unambiguous ranking of firms. Second, we assume that whenever a worker meets with a firm of rank K, it simultaneously meets with the firm at rank $K+1$ (i.e., the next-best firm). We refer to this as the *next-highest-ranked* bargaining protocol. We assume that the arrival of firm $K+1$ at the table is frictionless and does not come at the cost of any vacancies. Since firm K presents the highest surplus value, the protocol affects only how surplus is split, but in making the bargaining table more competitive, it reduces the firm's value of vacancy posting, which is taken into account. Indeed, in partial equilibrium, with contact rates held fixed, there would be no change at all to allocations.

Table 7 D reports the results. The markdown narrows by 11 percentage points—more than half of the baseline value of 21 percent—relative to that in the single ε case. While the *next-highest-ranked* bargaining protocol does not alter the partial equilibrium allocations, it certainly alters the general equilibrium allocations. A lower surplus share due to more direct competition reduces the return to vacancy posting. Less productive firms cut vacancies disproportionately. As more productive firms post a greater share of vacancies in equilibrium, labor productivity increases by an additional 5 percent relative to that in the single ε economy (column (A)). This increases concentration, seen in the tripling of the Herfindahl index from 0.28 to 0.84. Welfare increases by 7.2 percent relative to that in the

Variable	A. Single ε $\frac{1}{N}\int \varepsilon_i di$ fixed	B. ... and $\theta = 0.5$	C. ... and $\theta = 0.9$	D. ... and meet next lowest rank
HHI_n	0.28	0.54	0.72	0.84
Markdown (w/z), employment weighted	0.86	0.97	1.00	0.97
E-to-E worker flow rate (%)	0.63	0.64	0.67	0.48
Labor productivity vs. Single ε (%)	-	3.75	5.08	4.90
Welfare vs. Single ε (%)	-	5.48	7.34	7.20
Average wage vs. Single ε (%)	-	15.45	20.3	18.3
Standard deviation of log wage vs. Single ε (%)	-	-87.21	-93.8	-95.2

Table 7: Bargaining counterfactuals in Norway

single ε economy. The wage variance falls, driven by compression in wages within firms and more sorting of workers to high-productivity firms. In summary, more competition at the bargaining table is disproportionately bad for low-productivity firms, leading to a positive reallocation of employment.

Counterfactual levels of θ provide context for the large effects of the next-highest-ranked protocol. A more straightforward way to understand the effect of bargaining power could be to study comparative statics with respect to the worker bargaining weight $\theta \in [0, 1]$. However, this is difficult to interpret by itself. Column (B) shows that compared to the next-highest-ranked protocol, an increase in θ from 0.18 (baseline) to 0.50 has an equivalent effect on the markdown, and Column (C) shows that an increase in θ to 0.90 has a similar effect on welfare. Expanding the bargaining table has implications similar to those of massive increases in the bargaining power of workers[28].

6.3 Summary and policy implications

Our experiments have explored how the three paradigms of monopsony—concentration, amenities, and search frictions—interact. Concentration is multifaceted, and we find a significant role for the hard-wired duopsonistic wage setting built into models descended from Cahuc, Postel-Vinay, and Robin (2006). Giving a seat at the bargaining table to the *next-highest-ranked* firm lowers markdowns by 50 percent. On the other hand, we find little effect of firms excluding themselves from future contacts, and leaving the bargaining protocol untouched while doubling the number of firms in Norway does very little to markdowns. These results do not necessarily imply futility of stricter antitrust enforcement. Welfare significantly increases when the number of firms in the market doubles (see Table 6, row 5, column (B)) and even more so when firms are prevented from removing themselves from future contracts.

While outside the scope of the current formulation of our model, antitrust law may be an effective way to alter the bargaining protocol of firms. For instance, theoretical work by Shi (2023) argues that a near-ban on noncompetes is an efficient policy. Moreover, such a ban may be within the scope of the Federal Trade Commission's mandate. As this article is written, the Federal Trade Commission has proposed a rule to ban noncompetes[29]. Our framework can be easily modified to include noncompete restrictions and thus be used to assess the effect of such policies in a setting with firm concentration.

[28] Anecdotally, this is well understood in the post-PhD economics job market and provides incentives for firms to make exploding offers, limiting workers' length of search and hence the number of competing firms at the negotiation stage.

[29] See https://www.ftc.gov/legal-library/browse/federal-register-notices/non-compete-clause-rulemaking

We find that amenity dispersion explains one-third of observed markdowns in Norway. Since our model features first-degree price discrimination, however, the presence of amenity-driven markdowns does not imply inefficiency *per se*. Nonetheless, any policy that tilts the bargaining protocol in favor of the worker (raising σ and θ) increases pay by an erosion of the compensating differential. Put simply, a firm obtains output from the worker, and when pushed—via outside offers—they will pay up to that level of output. This raises consumer surplus but may lower total surplus.

Finally, search frictions in our discrete-time setting account for two-thirds of the observed markdowns. We view search frictions as immutable, with no practical policy to alleviate them (see also Naidu and Posner, 2022). Despite search frictions alone accounting for the majority of observed markdowns, the nonlinear interaction of the sources of monopsony in our model imply that addressing concentration and neoclassical sources of monopsony can still yield sizable improvements in worker welfare and wages.

One direction that we hope that future work will explore further is the *distributional* consequences of monopsony and market power. As Schmitz (2016) and Herkenhoff and Raveendranathan (2019) emphasize, the deadweight costs of monopoly are sizable and borne primarily by low-income households. This might also be true for any deadweight costs due to monopsony. Our framework is well suited to addressing these questions, as it allows for wage dispersion both within and across firms.

Two additional factors that can be accommodated relatively well in the model are human capital and risk aversion. Adding human capital to the model would enable researchers to explore the impact of monopsony and market power on investment in education and skill development. This would help us understand how these market structures affect workers' long-term career prospects and their ability to adapt to changes in the labor market. Work by Jungerman (2023) makes progress along these lines by incorporating human capital accumulation into a framework with dynamic oligopsony. Incorporating risk aversion would allow a more comprehensive analysis of the welfare costs of market power. This would help researchers better understand the costs associated with job loss and the ways in which individuals' decisions and well-being are affected by market power dynamics. Last, similarly to Berger, Herkenhoff, and Mongey (2022a), we observe that worker mobility across markets may be important for quantifying the aggregate consequences of policies. By exploring these additional factors, future research could provide a more complete picture of the distributional consequences of monopsony and market power and guide policy interventions aimed at mitigating their negative effects on workers and society at large.

7 Conclusion

In this paper, we develop a general equilibrium theory of monopsony that features (1) search frictions, (2) amenities, and (3) firm granularity. We estimate the strength of each source of monopsony using administrative data from Norway. Our approach introduces a novel method for defining markets and offers an extensive empirical overview of the conditional covariances of concentration, job flows, wages, and wage dispersion. Our model successfully replicates these relationships.

We use our model as a testing ground to investigate the sources of labor market power, focusing on wage markdowns—a measure defined as the ratio of a worker's current wage to her productivity at the firm. Our findings indicate that over 50 percent of observed wage markdowns can be attributed to firm

granularity and strategic wage setting during bargaining. Amenities account for 33 percent of observed markdowns, while eliminating search frictions (by setting contact rates to one in a discrete-time setting) reveals that they account for 66 percent of markdowns. Due to strong nonlinear interactions between the three monopsony channels, the sum of this decomposition does not equal 100 percent.

These results can inform policy discussions. Markdowns related to concentration could potentially be influenced by antitrust laws or restrictions on noncompete agreements, which may help reduce amenity wage penalties. In contrast, markdowns driven by search and matching processes are likely immutable and more influenced by technology than policy.

Our paper points to a number of fruitful avenues for future research. Allowing worker mobility across markets, incorporating human capital, and allowing for risk aversion are necessary to account for these first-order factors that may have an important bearing on policy recommendations. Likewise, structurally modeling natural experiments that affect the composition and productivity of firms within a market may yield novel insights into how labor markets respond to market structure and provide strong tests of our own and other existing theories of oligopsony.

References

Azar, J., I. Marinescu, and M. Steinbaum (2022): "Labor market concentration," *Journal of Human Resources*, 57(S), S167–S199.

Azar, J. A., S. T. Berry, and I. Marinescu (2022): "Estimating labor market power," Discussion paper.

Azar, J. A., I. Marinescu, M. I. Steinbaum, and B. Taska (2018): "Concentration in US labor markets: Evidence from online vacancy data," Discussion paper, National Bureau of Economic Research.

Azkarate-Askasua, M., and M. Zerecero (2022): "The aggregate effects of labor market concentration," *Available at SSRN 4323492*.

Bagga, S. (2022): "Firm Market Power, Worker Mobility, and Wages in the US Labor Market," *Manuscript*.

Bagger, J., and R. Lentz (2019): "An empirical model of wage dispersion with sorting," *The Review of Economic Studies*, 86(1), 153–190.

Benmelech, E., N. K. Bergman, and H. Kim (2022): "Strong Employers and Weak Employees How Does Employer Concentration Affect Wages?," *Journal of Human Resources*, 57(S), S200–S250.

Berger, D., T. Hasenzagl, K. Herkenhoff, S. Mongey, and E. Posner (2022): "Merger Guidelines for Labor Markets," *Working Paper*.

Berger, D., K. Herkenhoff, and S. Mongey (2022a): "Labor market power," *American Economic Review*, 112(4), 1147–93.

Berger, D. W., K. F. Herkenhoff, and S. Mongey (2022b): "Minimum wages, efficiency and welfare," Discussion paper, National Bureau of Economic Research.

Bhaskar, V., A. Manning, and T. To (2002): "Oligopsony and monopsonistic competition in labor markets," *Journal of Economic Perspectives*, 16(2), 155–174.

Bhaskar, V., and T. To (1999): "Minimum wages for Ronald McDonald monopsonies: A theory of monopsonistic competition," *The Economic Journal*, 109(455), 190–203.

Bhuller, M. (2009): "Classification of Norwegian Labor Market Regions," *SSB Notater*, 24.

Bhuller, M., K. O. Moene, M. Mogstad, and O. L. Vestad (2022): "Facts and Fantasies about Wage Setting and Collective Bargaining," *Journal of Economic Perspectives*, 36(4), 29–52.

Bloesch, J., and B. Larsen (2023): "When do Firms Profit from Wage Setting Power? New vs. Classical Monopsony," Manuscript.

Brooks, W. J., J. P. Kaboski, Y. A. Li, and W. Qian (2019): "Exploitation of Labor? Classical Monopsony Power and Labor's Share," Discussion paper, National Bureau of Economic Research.

Burdett, K., and D. Mortensen (1998): "Wage differentials, employer size, and unemployment," *International Economic Review*, pp. 257–273.

Burdett, K., S. Shi, and R. Wright (2001): "Pricing and matching with frictions," *Journal of Political Economy*, 109(5), 1060–1085.

Cahuc, P., F. Postel-Vinay, and J.-M. Robin (2006): "Wage bargaining with on-the-job search: Theory and evidence," *Econometrica*, 74(2), 323–364.

Card, D., A. R. Cardoso, J. Heining, and P. Kline (2018): "Firms and labor market inequality: Evidence and some theory," *Journal of Labor Economics*, 36(S1), S13–S70.

Dube, A., J. Jacobs, S. Naidu, and S. Suri (2019): "Monopsony in Online Labor Markets," *American Economic Review: Insights*, Forthcoming.

Engbom, N., and C. Moser (2022): "Earnings inequality and the minimum wage: Evidence from Brazil," *American Economic Review*, 112(12), 3803–47.

Felix, M. (2022): "Trade, Labor Market Concentration, and Wages," *Job Market Paper*.

Flinn, C., and J. Mullins (2021): "Firms' Choices of Wage-Setting Protocols," Discussion paper, Discussion paper, New York University.

Hazell, J., C. Patterson, H. Sarsons, and B. Taska (2022): "National Wage Setting," .

Herkenhoff, K. F., and G. Raveendranathan (2019): "Who bears the welfare costs of monopoly? The case of the credit card industry," .

Hurst, E., P. Kehoe, E. Pastorino, and T. Winberry (2022): "The Distributional Impact of the Minimum Wage," Discussion paper.

Jarosch, G. (2014): "Searching for Job Security and the Consequences of Job Loss," .

Jarosch, G., J. S. Nimcsik, and I. Sorkin (2019): "Granular Search, Market Structure, and Wages," *Manuscript*.

Jungerman, W. (2023): "Monopsony and human capital," Manuscript.

Kostol, A. R. (2017): "Mismatch and the Consequence of Job Loss," Discussion paper, Preliminary Working Paper.

Kroft, K., Y. Luo, M. Mogstad, and B. Setzler (2020): "Imperfect competition and rents in labor and product markets: The case of the construction industry," Discussion paper, National Bureau of Economic Research.

Lamadon, T., M. Mogstad, and B. Setzler (2022): "Imperfect competition, compensating differentials, and rent sharing in the US labor market," *American Economic Review*, 112(1), 169–212.

Lindenlaub, I., and F. Postel-Vinay (2021): "The Worker-Job Surplus," *National Bureau of Economic Research No. 28402*.

Lise, J., and F. Postel-Vinay (2020): "Multidimensional skills, sorting, and human capital accumulation," *American Economic Review*, 110(8), 2328–76.

Lise, J., and J.-M. Robin (2017): "The macrodynamics of sorting between workers and firms," *American Economic Review*, 107(4), 1104–1135.

Manning, A. (2003): *Monopsony in motion: Imperfect competition in labor markets*. Princeton University Press.

Naidu, S., and E. A. Posner (2022): "Labor Monopsony and the Limits of the Law," *Journal of Human Resources*, 57(S), S284–S323.

Postel-Vinay, F., and J.-M. Robin (2002): "Equilibrium wage dispersion with worker and employer heterogeneity," *Econometrica*, 70(6), 2295–2350.

Rinz, K. (2022): "Labor market concentration, earnings, and inequality," *Journal of Human Resources*, 57(S), S251–S283.

Robinson, J. (1933): *The Economics of Imperfect Competition*. Palgrave Macmillan.

Røed, K., and T. Zhang (2003): "Does Unemployment Compensation Affect Unemployment Duration?," *Economic Journal*, 113(484), 190–206.

Roussille, N., and B. Scuderi (2022): "Bidding for talent: Equilibrium wage dispersion on a high-wage online job board," Discussion paper, Working Paper.

Rubens, M. (2023): "Market structure, oligopsony power, and productivity," *Forthcoming, American Economic Review*.

Schmitz, J. A. (2016): "The cost of monopoly: A new view," *Federal Reserve Bank of Minneapolis Region*.

Schmutte, I. (2014): "Free to Move? A Network Analytic Approach for Learning the Limits to Job Mobility," *Labour Economics*, 29(C), 49–61.

Schubert, G., A. Stansbury, and B. Taska (2022): "Employer Concentration and Outside Options," *Mimeo*.

Shi, L. (2023): "Optimal Regulation of Noncompete Contracts," *Econometrica*, forthcoming.

Taber, C., and R. Vejlin (2020): "Estimation of a roy/search/compensating differential model of the labor market," *Econometrica*, 88(3), 1031–1069.

Trottner, F. (2022): "Misallocations in Monopsonistic Labor Markets," Discussion paper, Working Paper.

Yeh, C., C. Macaluso, and B. Hershbein (2022): "Monopsony in the US Labor Market," *Available at SSRN 4049993*.

Zhu, H. (2012): "Finding a good price in opaque over-the-counter markets," *The Review of Financial Studies*, 25(4), 1255–1285.

Appendix For Online Publication

A Data appendix

This section describes the main data sources and key variables and offers some suggestive evidence of the quality of our measures of vacancies and local labor markets.

A.1 Data sources

Linked employer–employee register. Statistics Norway and NAV jointly maintain the Norwegian Matched Employer–Employee Register, a linked employer–employee database (LEED) covering workers' earnings and transitions between employers. The employer reports the data to tax authorities at the end of the year and includes separate identifiers for the firm and its establishments. The data serve multiple purposes, including third-party tax reporting, as the basis for pension contributions and eligibility for safety net programs.

Wages and hours. There were some noteworthy changes to the data structure from 2014 to 2015. The reporting system was automated and based on monthly payments after 2015 and covers a slightly broader set of low-paying jobs and more detailed information on hours, bonuses, overtime, fixed pay, and variable pay. Before 2015, contracts with fewer than 4 hours per week or below an annual NOK10,000 were not reported. Importantly, for the vast majority of jobs, we observe the dates of alterations to the contract and the corresponding wage, industry and occupational codes, geographic location, and tenure at the establishment in both data sources. Hours reported are reasonably well measured in brackets from 4 to 19, 19 to 30, and above 30 hours per week, as pension contributions depend on them. We classify workers as full-time employees if their weekly hours are at least 30.

Occupations and workplace locations. The employment registers report 5-digit occupations. There are about 6,000 different occupations, where some job descriptions have been adjusted from the EU version to meet Norwegian standards and occupational licensing rules. For some positions, the descriptions include information about the rank of the occupation in the hierarchy, e.g., assistants, mid-level managers, top-level management, or members of the executive board[30].

We use the four-digit version. This version combines industry variation in certain occupations, such as code 3114 (machine engineers), which combines machine engineers in shipbuilding and construction. This version is also a natural definition of the career ladder for several skilled occupations, such as the code 2224 (pharmacists), which includes over-the-counter shop assistants, the licensed occupation for handling prescriptions, and senior positions in private pharmacies and clinical and hospital pharmacists.

[30]The first digit gives the skill level: 1: CEO/manager/politician, 2: master's degree, 3: bachelor's degree, 4: customer relations, 5: sales, health care, and service, 6: agriculture, 7: manual vocational occupations, 8: routine vocational, 9: no skill requirement, and 10: military. Occupations 4–8 require 10–12 years of schooling. 1 and 9 and 10 have no formal educational requirements. The second digit gives the field.

The tax registers also include the workplace location of the establishment—the municipality in which the employee must be present to perform her tasks. There are about 400 municipalities and 19 counties in Norway, none of which represent the natural boundaries of a local labor market. We instead use the information about the residence and workplace to define a commuting zone, which is our preferred definition of a local labor market. We follow Bhuller (2009), who aggregate municipalities into 46 regions allowed to cross counties and combine commuting statistics with natural boundaries, such as mountains and fjords.

Population registers. Demographic information on workers comes from longitudinal administrative registers provided by Statistics Norway. These data cover every Norwegian resident from 2006 to 2018 and contain the individual residential location, educational background, and demographic information (including on the worker's sex, age, residential location, spouse, and children).

The national education database (NUDB) includes the highest obtained degree from 1983 to 2018. These files provide information on the field (e.g., plumbing, mechanical engineering, nursing), level (e.g., vocational track in high school, bachelor's, master's, PhD), and years of schooling. There are approximately 5,000 educational codes.

A.2 Combining occupational codes

The International Standard Classification of Occupations (ISCO) underwent a major revision due to a resolution at the Meeting of Experts on Labor Statistics in 2007[31]. Statistics Norway and the Norwegian Labour and Welfare Administration (NAV) jointly adopted the international versions, with 354 and 403 unique four-digit occupations from ISCO88 and ISCO08. The Norwegian versions are named STYRK98 and STYRK08. STYRK98 is currently used in the employer–employee register, available from 2003, and STYRK08 is used in the vacancy data from the employment agency/NAV and after 2011 in the unemployment data. Several occupations, e.g., computer systems designers and computer programmers, were split into smaller groups during the revision. Other codes were collapsed into a larger group; for example, many types of machine operators were made obsolete by technological change. An official version of a two-way correspondence table from STYRK98 to STYRK08 is unavailable.

For our analysis, we need a crosswalk in both directions. Employees look for jobs classified in the new version. Similarly, employers must be able to see potential candidates whose skills are classified by the old version. To create a consistent measure of occupations, we proceed in three steps. First, we identify occupations with an exact match in the two versions. This gives a match of 50 occupations. Second, we identify revised occupations using unemployment periods that overlap with both versions in 2011/2012. To reduce noise from case-worker reporting, we keep 1:1 and 1:many mappings but keep only those with at least 30 percent of each unique STYRK98 code's total. Third, we keep all occupations in the official ISCO correspondence table with 1:1, 1:2, and 2:1 mappings. The remaining occupations are manually identified based on the available descriptions of STYRK codes and the ISCO crosswalk table[32].

[31] See ILO, "ISCO–08 Volume I: International Standard Classification of Occupations, Structure, Group Definitions and Correspondence Tables."

[32] Statistics Norway 1998 (NOS C521); Statistics Norway 2011 (Notater 17/2011).

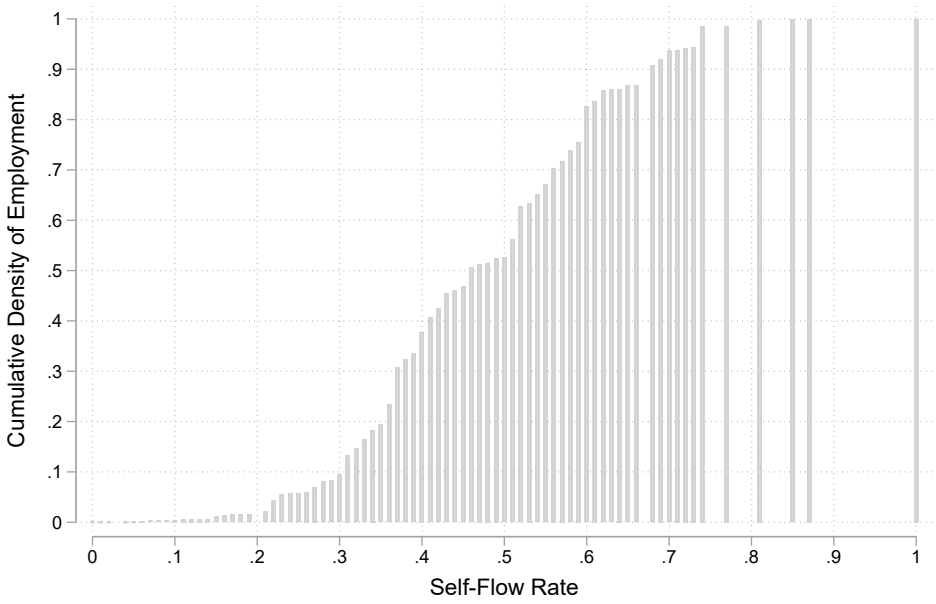


Figure 6: CDF of labor force by self-flow rates

Notes: The self-flow rates are calculated from all EE rates over the period 2006–2016, aggregated by 4-digit occupation.

A final set of consistent occupations depends on the linked sets of occupations with 1:many and many:1 mappings. The resulting number of occupations is reduced from 403 unique versions to 259 consistent occupations. Our crosswalked occupations can be mapped from either STYRK08 or STYRK98 and are available from this link.

A.3 Calculating occupational self-flow rates

The self-flow rates are calculated from the entire dataset of employer-to-employer (EE) transitions as the fraction of EE moves that move from and to the same occupation. We calculate the average self-flow rate of 4-digit occupations to be 43 percent and that of 3-digit occupations to be 45 percent. Next, we calculate the cumulative distribution of employment by the self-flow rate in an occupation. Figure 6 shows that a 50 percent cutoff captures approximately 50 percent of all employment.

B Additional empirical results

Quintiles of firms per market. Table 8 reports our main specifications when we use quintiles of the number of firms in the market as an alternate measure of concentration. Columns (1) through (3) illustrate that greater numbers of firms are associated with greater employer-to-employer transitions, greater wages, and a greater variance of wages, respectively.

Table 8: Alternate measures of concentration in a market: Number of firms

	(1) EE Rate	(2) Log Wage	(3) SD Log Wage
Number of firms, quintile 2	0.000598***	0.0150***	0.0564***
	(3.70e-05)	(0.00344)	(0.00394)
Number of firms, quintile 3	0.00115***	0.0366***	0.0849***
	(4.34e-05)	(0.00386)	(0.00415)
Number of firms, quintile 4	0.00177***	0.0690***	0.0991***
	(4.74e-05)	(0.00410)	(0.00425)
Number of firms, quintile 5	0.00260***	0.114***	0.114***
	(5.52e-05)	(0.00444)	(0.00451)
FE	O-Y	O-Y	O-Y
Controls	Y	Y	Y
Obs.	892,774	892,927	892,176
R^2	0.185	0.821	0.409

Unemployment vs. firms per worker. Table 9 illustrates the relation between the unemployment rate and the number of firms per worker. Column (1) of Table 9 regresses the unemployment rate (at the 4-digit occupation by region by time level) on firms per worker (at the market level). While the bottom quintile of firms per worker has marginally lower unemployment, the two specifications disagree on the relationship. Occupation–year fixed effects suggest higher unemployment in the highest quintile of firms per worker. Occupation–region fixed effects imply a nonmonotonic relationship. Column (2) yields an insignificant difference between the unemployment rate in quintiles 1 and 5 of the firms per worker distribution. Given this nonrobust relationship between unemployment and firms per worker, our vacancy posting cost is designed to deliver a flat relationship between the two variables.

Table 9: Relationship between unemployment and firms per worker in a market

	(1)	(2)
	——Unemployment Rate——	
Firms per worker, quintile 2	0.00330***	0.00323***
	(0.00118)	(0.00122)
Firms per worker, quintile 3	0.00323***	0.00514***
	(0.00125)	(0.00175)
Firms per worker, quintile 4	0.00518***	0.00659***
	(0.00122)	(0.00222)
Firms per worker, quintile 5	0.00640***	0.00452
	(0.00131)	(0.00275)
FE	O-Y	O-R
Controls	N	N
Obs.	964,179	964,132
R^2	0.417	0.514

C Robustness to an alternative definition of labor markets

In this section, we consider an alternative definition of local labor market markets. Rather than clustering occupations by K-means, we instead build on the work of Schmutte (2014) and extract the relevant clusters of occupations within a CZ using modularity maximization.

C.1 The modularity maximization approach

Modularity maximization (MM) is the workhorse model for community detection in network analysis. It identifies groups of nodes (also known as communities) that are densely interconnected within themselves but sparsely connected to nodes outside their group (like a labor market). This is achieved by maximizing a metric called modularity, with a high modularity score indicating that the observed network has more within-group connections than would be expected by chance.

Similarly to Schmutte (2014), who uses individual-level mobility data across occupations, we start from the aggregated occupational flows, grouped by 4-digit occupation separately for each commuting

zone (CZ). Our algorithm proceeds as follows:

1. First, we isolate single-occupation markets with high self-flow rates (e.g., 50 percent of EE moves).

2. To detect the occupational network among the remaining occupations, we perform modularity maximization, which searches for the grouping of occupations that maximizes the total sum of worker moves within the network relative to a randomly assigned group of occupations. For example, "electrical engineer" is considered a potential network if its inclusion yields a higher modularity score than its combination with a random occupation. We implement the Louvain algorithm, which assigns each occupation i to a network by calculating the change in modularity by moving i into the network of each occupation j to which i is connected (by at least one worker). In each iteration, it picks the j that increases the modularity score the most.

3. This process continues, where the network grows (or stays the same) until there is no change in the modularity score. We use the final network of occupations as the relevant clusters.

C.2 Results

We repeat the main graphical analysis using the MM market definition. Figure 7 provides a graphical presentation of our regression evidence using across-region, within-occupation–year variation, but now using the markets defined based on the MM algorithm. Reassuringly, the pattern remains the same as that for the markets defined based on the K-means clustering algorithm. We repeat this robustness check using the MM market definition with occupation–region fixed effects in Figure 8. Again, the correlations between the Herfindahl index and labor market outcomes remain quantitatively similar and qualitatively the same as in our main empirical analysis.

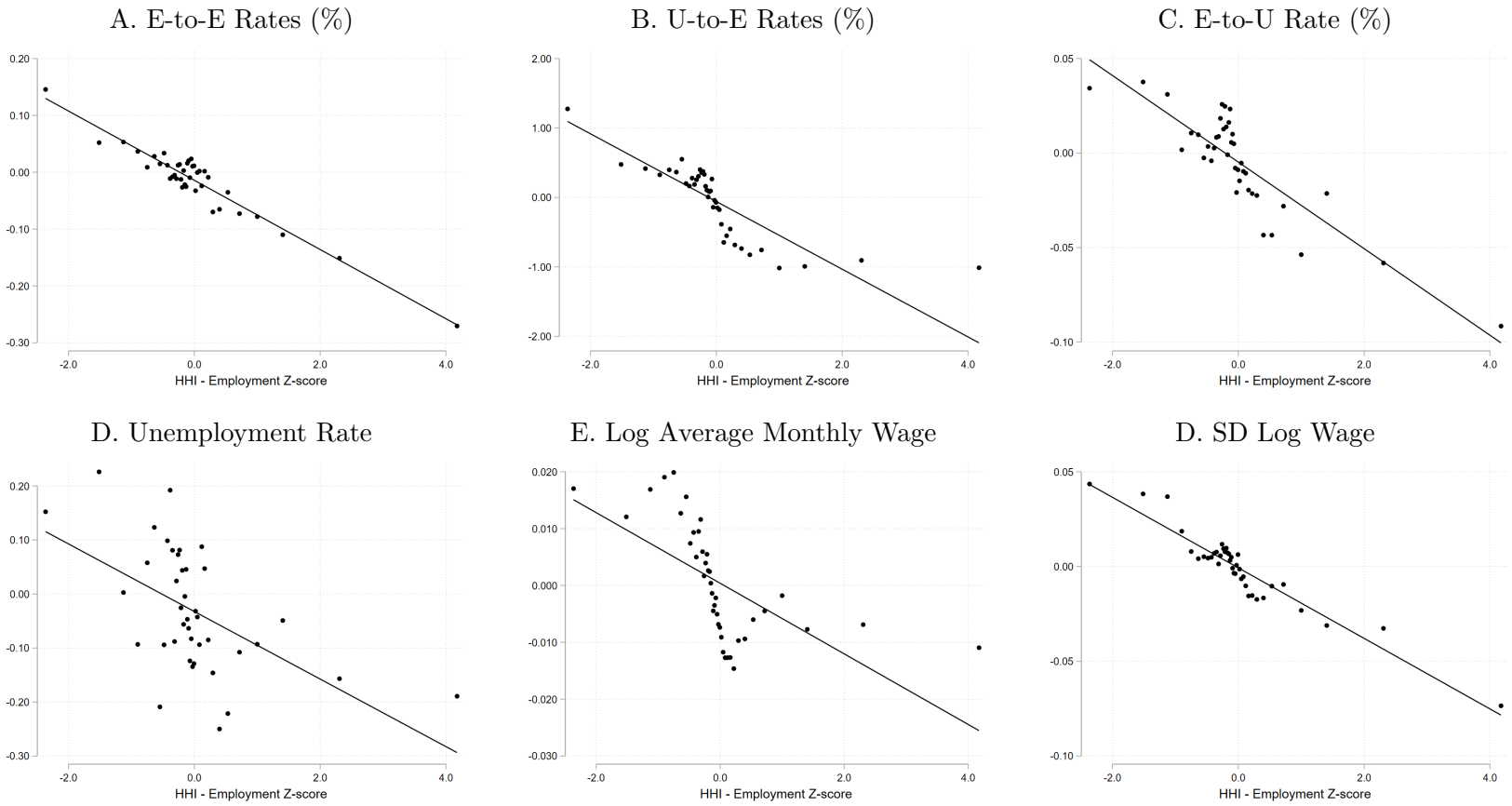


Figure 7: Concentration and labor market outcomes residualized on **occupation–year FEs** – Modularity maximization

Note: For each market (where a market is defined as a cluster of occupations within a commuting zone that maximizes the modularity score), we compute the employment Herfindahl index (HHI). For each 4-digit occupation–commuting zone–year, we compute the average of the dependent variable within 40 centiles of the market HHI, unweighted. We then residualize all x and y variables on occupation–year fixed effects, age composition, gender composition, education composition, lagged firms-per-worker ventiles, lagged labor force growth, and month-of-year dummies. The average NOK/USD in 2021 was 9.

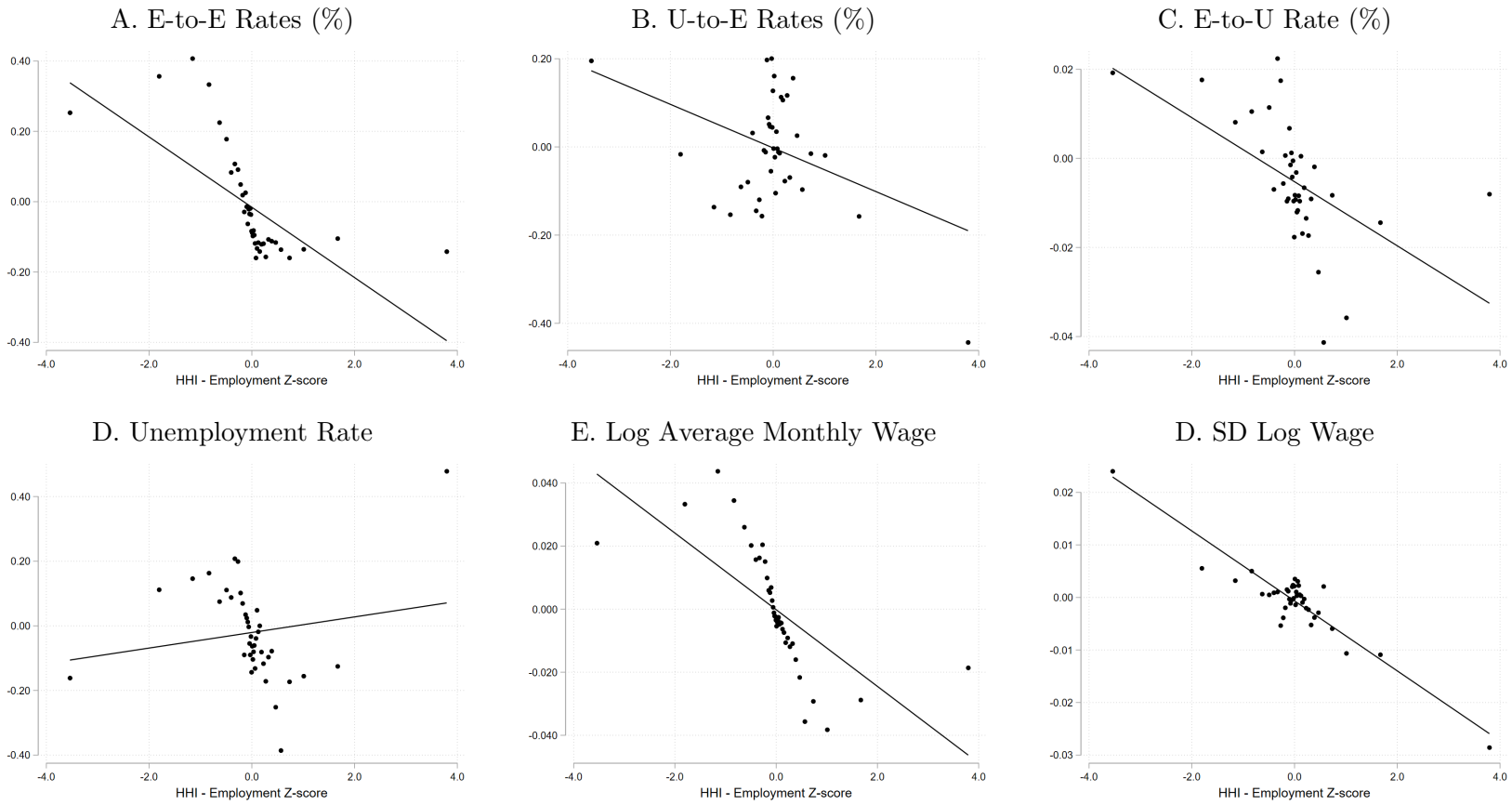


Figure 8: Concentration and labor market outcomes residualized on **occupation–region FEs** – modularity maximization

Note: For each market (where a market is defined as a cluster of occupations within a commuting zone that maximizes the modularity score), we compute the employment Herfindahl index (HHI). For each 4-digit occupation–commuting zone–year, we compute the average of the dependent variable within 40 centiles of the market HHI, unweighted. We then residualize all x and y variables on occupation–region fixed effects, age composition, gender composition, education composition, lagged firms-per-worker ventiles, lagged labor force growth, and month-of-year dummies. The average NOK/USD in 2021 was 9.

C.3 Industry and occupations as markets

Table 10 computes self-flow rates for raw industry and occupation codes in the Austrian and Norwegian data. The information for the Austrian markets is taken from Jarosch, Nimcsik, and Sorkin (2019). We find that self-flow rates are very low with industry definitions. Flows based on occupations are higher and the raw 4-digit occupation definition of a market yields a self flow rate of 43 percent while the raw 3-digit occupation definition of a market yields a self flow rate of 45 percent.

Table 10: Summary Statistics: Within Market EE Shares

	Industry		Occupation			Flow-based	
Market definition: CZ×	4	2	7	4	3	Firm	Occu
Austria	0.18	0.25				0.41	
Norway	0.32	0.39	0.33	0.43	0.45		0.52

Notes: Within-market transitions among all employer-employer (EE) transitions defined as a change of the firm with at most 30 days of non-employment between the two spells and at least one year of tenure in the old and the new job. Market is industry/occupation×commuting zone. Statistics from Austria is from JNS22 and Norway are own calculations. Outflow from CZ in Austria is 40% and Norway is 15% (see Bhuller 2009).

D Within-market and all flows

In the main text, we compute all flows using within-market stocks and flows. Table 11 computes flows using all workers regardless of whether they stay or leave the market. The E-to-E rate is higher (mechanically), but the separation and job-finding rates remain very similar.

Table 11: Within-market flows and all flows

	Within market flows	All flows
E-to-E rate (monthly %)	0.65%	1.19%
U-to-E rate (monthly %)	8.08%	8.82%
E-to-U rate (monthly %)	0.35%	0.35%

Figure 9 reports the covariances between the Herfindahl valuse and flows using all workers regardless of whether they stay or leave the market. Panels A through C include occupation–year FEs, and Panels D through F include occupation–region FEs. We find patterns very similar to those in the main text.

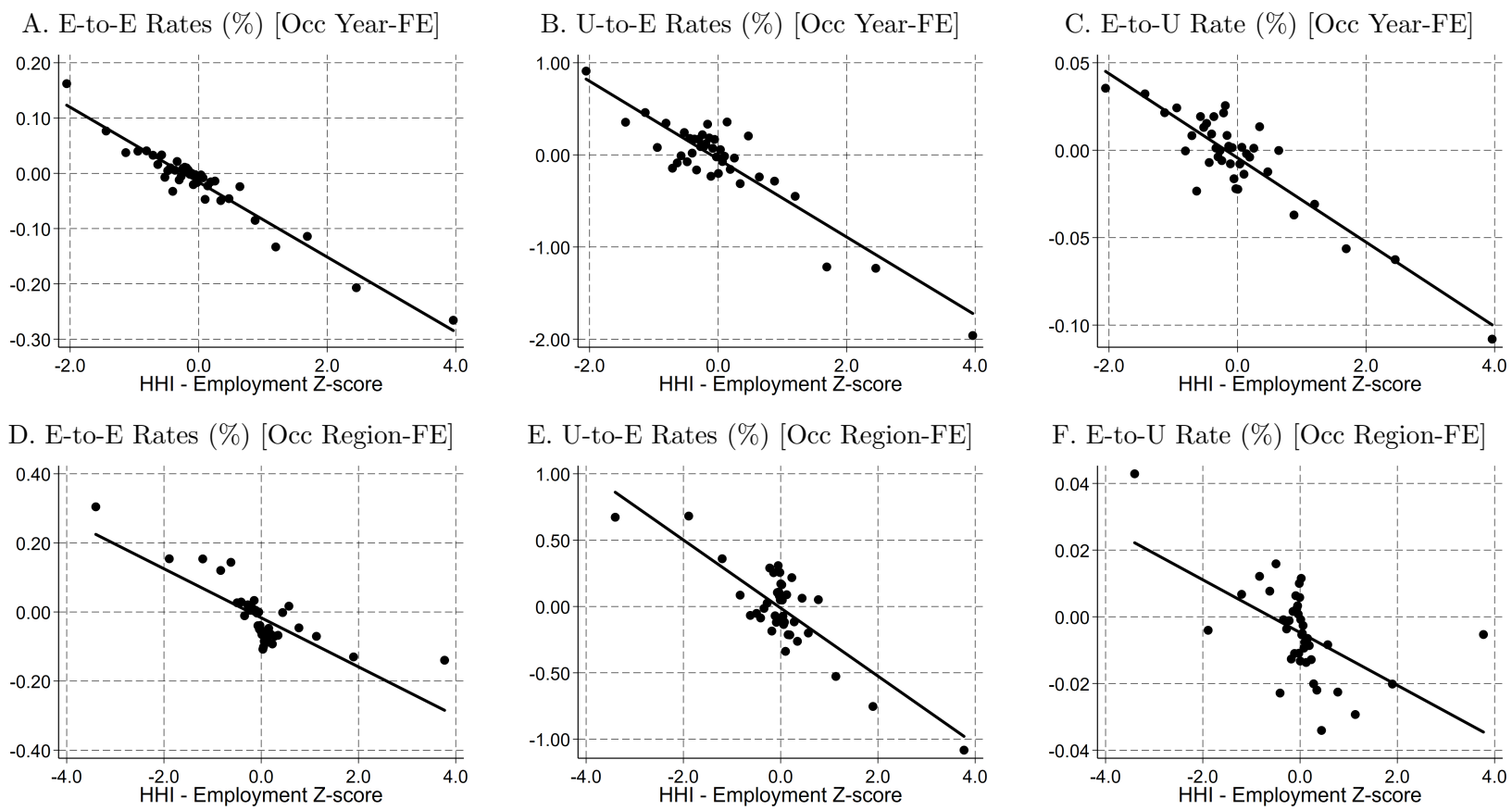


Figure 9: Concentration and job flows using main-text, K-means definitions of markets

Note: For each market (where a market is defined, as in Section 2, as a cluster of occupations within a commuting zone), we compute the employment Herfindahl index (HHI). For each 4-digit occupation–commuting zone–year, we compute the average of the dependent variable within 40 centiles of the market HHI, unweighted. We then residualize all x and y variables on occupation–year fixed effects, age composition, gender composition, education composition, lagged firms-per-worker ventiles, lagged labor force growth, and month-of-year dummies. The average NOK/USD in 2021 was 9.

E Model derivations

In this appendix, we derive the main equations in the text. To keep the algebra simple, we derive the equations without amenities. Adding amenities involves a straightforward modification to these equations. As defined in the text, $P_i := W_i + J_i$ is the joint value of a match, where W_i is the worker continuation value and J_i is the firm continuation value. The surplus is defined as $S_i := W_i + J_i - U_i - V_i = P_i - U_i - V_i$, where we assume that the outside option of the firm is zero, $V_i = 0$ (i.e., the job position is destroyed if not filled by the current worker).

Bargaining and surplus shares. We begin with the Nash bargaining problem of a worker moving from firm i to firm k. This job flow happens only when $P_k > P_i$, and thus, the worker's outside option is $W_i = P_i$ (i.e., to extract the full joint value of the match or, equivalently, to extract the full surplus of the match $W_i - U_i = P_i - U_i = S_i$):

$$\begin{aligned} &\max_{W_k} \left(P_k - W_k\right)^{1-\theta} \left(W_k - P_i\right)^{\theta}, \\ &\quad \Rightarrow W_k = P_i + \theta \left[P_k - P_i\right]. \end{aligned} \tag{10}$$

As a convenient accounting device, we write the worker value as the value of unemployment plus some share σ of the match surplus S (see Lise and Postel-Vinay (2020)). Thus, the definition of the worker value $W_k(\sigma)$ is given by a σ share of surplus.

$$W_k(\sigma) := U + \sigma S_k = U + \sigma \left[P_k - U\right]. \tag{11}$$

Equating the solution to the Nash bargaining problem in equation (10) to our accounting definition of the worker continuation value in equation (11) yields the following expression for the surplus share:

$$\begin{aligned} W_k(\sigma) &= P_i + \theta \left[P_k - P_i\right], \\ U + \sigma \left[P_k - U\right] &= P_i + \theta \left[P_k - P_i\right], \\ \sigma\left(P_i, P_k\right) &= \frac{\theta \left[P_k - U\right] + (1-\theta)\left[P_i - U\right]}{\left[P_k - U\right]}. \end{aligned} \tag{12}$$

We can write the surplus share in equation (12) equivalently in terms of surplus values:

$$\sigma\left(S_i, S_k\right) = \theta + (1-\theta)\frac{S_i}{S_k} \tag{13}$$

We use equation (13) in the law of motion for the surplus share, which we denote σ':

$$\sigma' = \begin{cases} \left(\frac{\theta S_k + (1-\theta) S_i}{S_k}\right) & \text{if } S_k > S_i \\ \max\left\{\sigma, \frac{S_k}{S_i}\right\} & \text{if } S_k \leq S_i. \end{cases}$$

In the event that the surplus at the new firm k is greater than that at firm i ($S_k > S_i$), then the worker

moves and Nash bargains; thus, $\sigma' = \left(\frac{\theta S_k + (1-\theta) S_i}{S_k}\right)$. In the event that the worker meets a firm that offers less surplus, the worker stays at the incumbent firm, and the offer is either matched (with no Nash bargaining) or ignored; thus, $\sigma' = \max\left\{\sigma, \frac{S_k}{S_i}\right\}$ $(S_k \leq S_i)$.

Surplus. To derive an equation for surplus, we begin by describing the continuation value for a worker who is unemployed, U. A worker who finds a job coming from unemployment has an outside option of U, so $W_i = U + \theta\left[P_i - U\right]$, which implies directly that $\sigma = \theta$. We can write the continuation value of unemployment as follows:

$$U = b + \beta\left[U + \sum_k \lambda_{uk} \mathbf{1}\left[W_k(\theta) > U\right]\left[W_k(\theta) - U\right]\right],$$
$$U = b + \beta\left[U + \theta \sum_k \lambda_{uk} \mathbf{1}\left[P_k > U\right]\left[P_k - U\right]\right],$$
$$U = b + \beta\left[U + \theta \sum_k \lambda_{uk} \max\left\{S_k, 0\right\}\right].$$

We can also derive an expression for the joint value of the worker and firm. This joint value takes into account the worker's value from a new match and the worker's value in unemployment:

$$P_i = z_i + \beta\left[P_i + \sum_{k \neq i} \lambda_{ik} \mathbf{1}\left[P_k > P_i\right]\left[W_k\left(\sigma\left(P_i, P_k\right)\right) - P_i\right] + \delta\left(U - P_i\right)\right]. \tag{14}$$

The term $\sum_{k \neq i} \lambda_{ik} \mathbf{1}\left[P_k > P_i\right]\left[W_k\left(\sigma\left(P_i, P_k\right)\right) - P_i\right]$ captures the value of the worker's meeting with a new firm $k \neq i$. The contact rate is λ_{ik}; the worker moves if the surplus in the new firm is greater $[P_k > P_i]$; and the worker bargains and obtains a surplus share given by equation (12). The term $\delta\left(U - P_i\right)$ captures the value of unemployment if the worker separates. Substituting the definition of the surplus into equation (14) yields

$$P_i = z_i + \beta\left[P_i + \theta \sum_{k \neq i} \lambda_{ik} \max\left\{S_k - S_i, 0\right\} - \delta S_i\right]. \tag{15}$$

By subtracting U from both sides of equation (15), we can derive an expression for the surplus:

$$P_i - U = \left\{z_i + \beta\left[P_i + \theta \sum_{k \neq i} \lambda_{ik} \max\left\{S_k - S_i, 0\right\} - \delta S_i\right]\right\} - \left\{b + \beta\left[U + \theta \sum_k \lambda_{uk} \max\left\{S_k, 0\right\}\right]\right\},$$
$$S_i = z_i - b + \beta\left[S_i + \theta \sum_{k \neq i} \lambda_{ik} \max\left\{S_k - S_i, 0\right\} - \delta S_i - \theta \sum_k \lambda_{uk} \max\left\{S_k, 0\right\}\right]. \tag{16}$$

An important point to note is that with a finite number of firms, the surplus equation can be solved

via matrix inversion (see Section E.1). Given contact rates λ_{uk}, the surplus equation can be solved independently of its components W_i, U, and J_i. This remains true with amenities. Moreover, the worker's share of surplus does not alter the size of the surplus—hence its independence of σ. This is a standard result in models with linear utility.

Wage equation. We derive the wage equation in three steps. First, we derive the worker's value from first principles. Second, we use our accounting device whereby worker values are expressed as a share of surplus. Third, we equate these two formulas for the worker's value to solve for the wage. The resulting wage depends only on the surplus.

Step I. The first step is to compute the worker value from first principles. The worker value under the Cahuc, Postel-Vinay, and Robin (2006) (CPVR) bargaining protocol can be written as

$$W_i(\sigma) = w_i(\sigma) + \beta\Big[W_i(\sigma) + \sum_{k\neq i}\lambda_{ik}\left\{\mathbf{1}\left[P_k > P_i\right]\left[\theta P_k + (1-\theta)P_i - W_i(\sigma)\right] + \mathbf{1}\left[W_i(\sigma) < P_k < P_i\right]\left[P_k - W_i(\sigma)\right]\right\} + \delta\left[U_i - W_i(\sigma)\right]\Big].$$

This equation can be written compactly using min and max notation:

$$W_i(\sigma) = w_i(\sigma) + \beta\left[W_i(\sigma) + \sum_{k\neq i}\lambda_{ik}\max\left\{0, \min\left\{\theta P_k + (1-\theta)P_i, P_k\right\} - W_i(\sigma)\right\} + \delta\left[U - W_i(\sigma)\right]\right].$$

Last, we use the accounting definition of the worker value $W_i(\sigma) = U + \sigma\left[P_i - U\right]$, and by the definition of surplus $W_i(\sigma) - U = \sigma\left[P_i - U\right]$, we arrive at the following expression for the worker value:

$$W_i(\sigma) = w_i(\sigma) + \beta\left[W_i(\sigma) + \sum_{k\neq i}\lambda_{ik}\max\left\{0, \min\left\{\theta\left[P_k - P_i\right], P_k - P_i\right\} + (1-\sigma)\left[P_i - U\right]\right\} - \delta\sigma S_i\right].$$

Continuing to simplify, we can write the worker continuation value in terms of surplus:

$$W_i(\sigma) = w_i(\sigma) + \beta\left[W_i(\sigma) + \sum_{k\neq i}\lambda_{ik}\max\left\{0, \min\left\{\theta\left[S_k - S_i\right], S_k - S_i\right\} + (1-\sigma)S_i\right\} - \delta\sigma S_i\right]. \quad (17)$$

Step II. Our second step is to derive the worker's wage value from our accounting definition of the worker value in equation 11. Starting from equation 11, we have

$$W_i(\sigma) = U + \sigma S_i,$$
$$(1-\beta)W_i(\sigma) = (1-\beta)U + \sigma(1-\beta)S_i.$$

We then substitute the expression for surplus $(1-\beta)S_i$ by subtracting βS_i from both side of the

surplus equation (16) to arrive at an alternate expression for the worker's continuation value:

$$(1-\beta)\,W_i(\sigma) = \sigma z_i + (1-\sigma)\,b + \beta\left[\sigma\theta\sum_{k\neq i}\lambda_{ik}\max\{S_k - S_i, 0\} - \sigma\delta S_i + (1-\sigma)\,\theta\sum_{k}\lambda_{uk}\max\{S_k, 0\}\right]. \tag{18}$$

Step III. Equating the first expression for the worker's value in equation (18) to the second expression for the worker's value in equation (18), we obtain an expression that can be solved to obtain the wages:

$$\begin{aligned} &w_i(\sigma) + \beta\left[\sum_{k\neq i}\lambda_{ik}\max\{0, \min\{\theta\,[S_k - S_i], S_k - S_i\} + (1-\sigma)\,S_i\} - \delta\sigma S_i\right] \\ &= \sigma z_i + (1-\sigma)\,b + \beta\left[\sigma\theta\sum_{k\neq i}\lambda_{ik}\max\{S_k - S_i, 0\} - \sigma\delta S_i + (1-\sigma)\,\theta\sum_{k}\lambda_{uk}\max\{S_k, 0\}\right]. \end{aligned}$$

Rearranging the above expression, we can derive an expression for the wages of a worker in terms of surplus:

$$\begin{aligned} w_i(\sigma) = \sigma z_i + (1-\sigma)\,b + \beta\Bigg[&(1-\sigma)\,\theta\sum_{k}\lambda_{uk}\max\{S_k, 0\} \\ &- \sum_{k\neq i}\lambda_{ik}\max\{0, \min\{(1-\sigma)\,\theta\,(S_k - S_i), (S_k - S_i)\} + (1-\sigma)\,S_i\}\Bigg]. \end{aligned} \tag{19}$$

Similarly to the baseline model with amenities, this simpler wage expression includes three terms: (i) workers obtain σ of production, (ii) workers obtain (1 - σ) of their outside option, and last, (iii) there is backloading since firms that offer greater future pay prospects can initially pay less.

E.1 Solving for surplus and wages via matrix inversion

Because there is a finite number of firms, the solution for surplus and wages (conditional on contact rates) can be reduced to a series of matrix inversions. We briefly describe how we vectorize the model to make it easier to solve computationally. We first define $\tilde{z}_i = z_i - b$ and rearrange (16) to obtain the following expression for surplus:

$$(1 - \beta\,(1-\delta))\,S_i = \tilde{z}_i + \beta\theta\sum_{k\neq i}\lambda_{ik}\max\{S_k - S_i, 0\} - \beta\theta\sum_{k}\lambda_{uk}\max\{S_k, 0\}. \tag{20}$$

For pedagogical purposes, we work through the vectorization for the case of $N = 3$ firms. We first define the matrix Ψ to be the lower triangular matrix of contact rates, where firms are ordered by their productivity. We let $i = 1$ correspond to the highest productivity value, and since there are no amenities, no workers ever leave the highest-productivity firm. Thus, Ψ describes the optimal policy of workers

and their flow rates across employers:

$$\Psi := \begin{bmatrix} 0 & 0 & 0 \\ \lambda_{21} & 0 & 0 \\ \lambda_{31} & \lambda_{32} & 0 \end{bmatrix}.$$

With amenities, the ranking of firms in terms of (z, ε)-tuples must be guessed and then iterated on until convergence. However, a similar vectorization is easily derived.

We must introduce some additional notation to vectorize the surplus equation. Let $\boldsymbol{I}_N$ denote the identity matrix, $\boldsymbol{i}_N$ be an $N \times 1$ column vector of ones, $\boldsymbol{\lambda}_u$ denote the column vector of unemployed worker contact rates, $\boldsymbol{S}$ denote the column vector of surpluses, and $\otimes$ denote the Kronecker product.

We can continue to work with (20) in the case of three firms to derive the factorized surplus formula:

$$\left(\boldsymbol{I}_N - \beta\left(1-\delta\right)\boldsymbol{I}_N\right)\begin{bmatrix} S_1 \\ S_2 \\ S_3 \end{bmatrix} = \begin{bmatrix} \tilde{z}_1 \\ \tilde{z}_2 \\ \tilde{z}_3 \end{bmatrix} + \beta\theta \begin{bmatrix} 0 & 0 & 0 \\ \lambda_{21}\left(S_1 - S_2\right) & 0 & 0 \\ \lambda_{31}\left(S_1 - S_3\right) & \lambda_{32}\left(S_2 - S_3\right) & 0 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} - \beta\theta \begin{bmatrix} \lambda_{u1} & \lambda_{u2} & \lambda_{u3} \\ \lambda_{u1} & \lambda_{u2} & \lambda_{u3} \\ \lambda_{u1} & \lambda_{u2} & \lambda_{u3} \\ \lambda_{u1} & \lambda_{u2} & \lambda_{u3} \end{bmatrix}\begin{bmatrix} S_1 \\ S_2 \\ S_3 \end{bmatrix},$$

$$\left(\boldsymbol{I}_N - \beta\left(1-\delta\right)\boldsymbol{I}_N\right)\boldsymbol{S} = \tilde{\boldsymbol{z}} + \beta\theta \begin{bmatrix} 0 \\ \lambda_{21}\left(S_1 - S_2\right) \\ \lambda_{31}\left(S_1 - S_3\right) + \lambda_{32}\left(S_2 - S_3\right) \end{bmatrix} - \beta\theta\left(\boldsymbol{i}_N \otimes \boldsymbol{\lambda}_u'\right)\boldsymbol{S},$$

$$\left(\boldsymbol{I}_N - \beta\left(1-\delta\right)\boldsymbol{I}_N + \beta\theta\left(\boldsymbol{i}_N \otimes \boldsymbol{\lambda}_u'\right)\right)\boldsymbol{S} = \tilde{\boldsymbol{z}} + \beta\theta\left(\begin{bmatrix} 0 & 0 & 0 \\ \lambda_{21} & 0 & 0 \\ \lambda_{31} & \lambda_{32} & 0 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ 0 & \lambda_{21} & 0 \\ 0 & 0 & \lambda_{31}+\lambda_{32} \end{bmatrix}\right)\begin{bmatrix} S_1 \\ S_2 \\ S_3 \end{bmatrix}. \tag{21}$$

The first matrix is Ψ, and the second matrix is $diag\left(\Psi \times \boldsymbol{i}_N\right)$:

$$\Psi \times \boldsymbol{i}_N = \begin{bmatrix} 0 & 0 & 0 \\ \lambda_{21} & 0 & 0 \\ \lambda_{31} & \lambda_{32} & 0 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ \lambda_{21} \\ \lambda_{31}+\lambda_{32} \end{bmatrix}$$

Inspecting (21), it becomes clear that we can generalize the expression to N firms as follows:

$$\left(\boldsymbol{I}_N - \beta\left(1-\delta\right)\boldsymbol{I}_N + \beta\theta\left(\boldsymbol{i}_N \otimes \boldsymbol{\lambda}_u'\right)\right)\boldsymbol{S} = \tilde{\boldsymbol{z}} + \beta\theta\left(\Psi - diag\left(\Psi \times \boldsymbol{i}_N\right)\right)\boldsymbol{S}$$

$$\left(\boldsymbol{I}_N - \beta\left(1-\delta\right)\boldsymbol{I}_N + \beta\theta\left(\boldsymbol{i}_N \otimes \boldsymbol{\lambda}_u' - \Psi + diag\left(\Psi \times \boldsymbol{i}_N\right)\right)\right)\boldsymbol{S} = \tilde{\boldsymbol{z}}$$

We can then invert the resulting matrix to solve for the vector of surpluses, $\boldsymbol{S}$:

$$\boldsymbol{S} = \left(\boldsymbol{I}_N - \beta\left(1-\delta\right)\boldsymbol{I}_N + \beta\theta\left(\boldsymbol{i}_N \otimes \boldsymbol{\lambda}_u' - \Psi + diag\left(\Psi \times \boldsymbol{i}_N\right)\right)\right)^{-1} \times \tilde{\boldsymbol{z}} \tag{22}$$

With the surpluses given by (22), we can then solve for wages using equation (19). Importantly, our accounting device in which we express worker continuation values as a share of surplus allows us to solve for wages in terms of surplus values *alone*. These features of our model make it fast to solve and readily tractable to the incorporation of amenities. The framework can be easily modified to include rich worker

heterogeneity, as well.

E.2 Solution method with amenities

We describe the solution method for the full model with amenities. Within each market j, perform the following steps to solve for the market equilibrium:

i. Guess a vector of vacancies $\{v_i\}$.

ii. Guess a vector of employment $\{n_i(\varepsilon)\}$.

iii. Combine vacancies and employment to compute contact rates for workers using equation (2) and for firms using equation (3) (note that this computation of contact rates can be easily vectorized).

iv. Guess a ranking of firms in the (z, ε) space.

v. Solve for the vector of surplus values $\boldsymbol{S}$ via matrix inversion.

vi. Iterate on the ranking of firms until it agrees with $\boldsymbol{S}$.

vii. Solve for the stationary distribution of workers across firms using (8) (note that this law of motion can be easily vectorized).

viii. Solve the firm's objective value for a new vector of vacancies using equation (7) (note that the update for vacancies can be easily vectorized).

ix. Iterate until the vacancies converge.

x. Recover wages using (19), and simulate individual worker/wage histories as required to compute moments.