Firm and Worker Dynamics in a Frictional Labor Market [*]

Adrien Bilal[†] *Niklas Engbom*[‡] *Simon Mongey*[§] *and* *Giovanni L. Violante*[¶]

January 21, 2022

Abstract

This paper integrates the classic theory of firm boundaries, through span of control or taste for variety, into a model of the labor market with random matching and on-the-job search. Firms choose when to enter and exit, whether to create vacancies or destroy jobs in response to shocks, and Bertrand-compete to hire and retain workers. Tractability is obtained by proving that, under a parsimonious set of assumptions, all worker and firm decisions are characterized by their joint surplus, which in turn only depends on firm productivity and size. The job ladder in marginal surplus that emerges in equilibrium determines net poaching patterns by firm characteristics that are in line with the data. As frictions vanish, the model converges to a standard competitive model of firm dynamics. The combination of firm dynamics and search frictions allows the model to: (i) quantify the misallocation cost of frictions; (ii) replicate elusive life-cycle growth profiles of superstar firms; and (iii) make sense of the failure of the job ladder around the Great Recession as a result of the collapse of firm entry.

Keywords: Diminishing Returns to Scale, Firm Dynamics, Frictional Misallocation, Great Recession, Job Turnover, Marginal Surplus, Net Poaching, On the Job Search, Unemployment, Vacancies, Worker Flows.

JEL Classification: D22, E23, E24, E32, J23, J63, J64, J69.

[*] The view included are those of the authors and do not represent those of the Federal Reserve System. We thank our discussants Björn Brügemann, Nir Jaimovich and Thibault Lamadon. We also thank Pierre Cahuc, V.V. Chari, Steven Davis, Nobu Kiyotaki, Per Krusell, John Moore, Giuseppe Moscarini, and Jean-Marc Robin, and many other seminar participants for their useful feedback. We thank Sebastian Guarda, Hugo Lhuillier and Alex Weinberg for excellent research assistance. The views expressed in this study are those of the authors and do not necessarily reflect the position of the Federal Reserve Bank of Minneapolis or the Federal Reserve System.

[†] Harvard University and NBER

[‡] New York University, CEPR and NBER

[§] Federal Reserve Bank of Minneapolis, University of Chicago and NBER

[¶] Princeton University, CEBI, CEPR, IFS, IZA, and NBER

1 Introduction

Aggregate production in the economy is divided into millions of firms, each facing idiosyncratic fluctuations in its productivity and demand. Understanding the process of labor reallocation across these production units is important for several reasons. In the long run, reallocating labor away from unproductive firms toward more productive firms enhances aggregate productivity and growth. In the short run, the propagation of sectoral and aggregate shocks depends on how quickly labor flows across firms and between unemployment and employment. From a normative perspective, understanding the potential welfare losses or gains due to reallocation is necessary for assessing the efficacy of policies that subsidize jobless workers, protect employment, or advantage particular sectors or firms.

The labor reallocation process has three key properties. First, it has distinct layers: the entry and exit of firms, the creation and destruction of jobs at existing firms, and the turnover of workers across existing jobs. Second, it is intermediated by labor markets that are frictional, as revealed by the coexistence of vacancies and job seekers. Third, around half of worker turnover occurs through direct job-to-job transitions: most new hires come from another firm rather than from unemployment.

Therefore, addressing labor reallocation requires a framework with two central elements. First, a theory of the firm (i.e., its boundaries) and of firm dynamics (entry, growth, separations, exit). Second, a theory of worker flows intermediated by frictional labor markets that allows for on-the-job search and job-to-job mobility (i.e., poaching). Quantitatively, such a framework should account for a new body of time series and cross-sectional evidence—emerging from matched employer-employee data—that describes the relationship between firm characteristics and the direction and composition of worker flows.[1]

This paper presents a new model with these features. A firm is a profit maximizing owner of a technology with decreasing returns to scale and stochastic productivity that chooses optimally whether to enter and when to exit the market. Equivalently, the firm could be a monopolistic producer facing a downward sloping demand curve with a stochastic shifter, with an isomorphic interpretation in our model. Firms grow by posting costly vacancies that are randomly matched to either unemployed or employed workers. Worker flows occur when matched workers determine that the value of working at the newly matched firm exceeds their value of unemployment or employment in their current firm. In general, with decreasing returns to scale in production, these values are a complicated function of a high dimensional state vector that includes distributions of wages or worker values inside the firm. This

[1] If we consider hires for a particular firm type (e.g., young, small and fast-growing), by *composition* we mean the split between hires from unemployment and those from employment. Within hires from employment, *direction* refers to the characteristics of the employers between which workers are reallocated.

makes the problem seem intractable.

Our first contribution is to set out a parsimonious set of assumptions that are *sufficient* for tractability. Our assumptions place a minimal structure on the contractual environment such that the state vector becomes manageable. Three assumptions on bargaining and surplus sharing are common to many single-worker firm environments: (i) lack of commitment; (ii) wage contract renegotiation by mutual consent; (iii) Bertrand competition among employers for employed jobseekers. Two further assumptions are required in our new multi-worker firm environment: (iv) no value is lost in internal wage renegotiations between a firm and its incumbent workers; and (v) vacancy policies maximize combined firm and worker value—for which we offer an explicit microfoundation. Under these assumptions, firm and workers' decisions are privately efficient, as if the firm and incumbent workers maximize their *total value*. The state variables of the total value function are only two: firm size (n) and productivity (z).

Two other ingredients are vital to achieve tractability. First, we work in continuous time. In a small interval of time only one random event may occur. For example, a firm only needs to deal with one of its employee meeting another firm, not all combinations of its employees meeting other firms. Second, we take the continuous limit of a discrete workforce. Worker flows are determined by comparing the change in total value that would arise if a worker joins or leaves a firm. With a continuous measure of workers, this *marginal value* can be conveniently expressed as a partial derivative of total value.

We show that total and marginal value are sufficient for characterizing firm and worker dynamics. Marginal value pins down hiring: facing a convex vacancy cost, firms post vacancies until the marginal cost of a vacancy is equal to the expected marginal value of hiring.[2] Marginal value also pins down separations: facing a decreasing marginal product of labor, firms fire workers until the marginal value of a worker equals the value of unemployment. When total value is less (more) than the firm owner's outside option, the firm exits (enters). Finally, in equilibrium, marginal values determine the direction of worker flows. Workers climb a *job ladder* in marginal value, quitting when on-the-job search delivers a match with a higher marginal value firm. An intuitive Bellman equation accounts for the evolution of the total value, while a law of motion reflecting frictional labor reallocation accounts for the evolution of the firm size and productivity distribution.[3]

Our second contribution exploits the mathematical tractability of our framework to analytically characterize equilibrium firm and worker reallocation. First, we analyze firm dynamics and job turnover

[2] Convex adjustment costs are among the solutions proposed by Elsby, Michaels, and Ratner (2019) to obtain empirically plausible sluggish adjustment of labor market aggregates, which is difficult to generate in models with fixed or linear costs.

[3] This representation uniquely pins down firm and worker dynamics, the subject of this paper, but is consistent with multiple wage determination mechanisms that determine how this joint value is split. Wages, therefore, are not allocative in that the distribution of firms and flows of workers across firms is independent of wage dynamics. In order to study the model's implication for wage dynamics, one has to make additional assumptions. We return on this point in Section 2.

graphically in (n, z)-space by describing the regions in which a firm exits, fires and hires. Firms that exit and fire always destroy jobs. Hiring firms may either grow on net (creating jobs) or shrink on net (destroying jobs) because some of their workers quit to firms with a higher rank on the marginal value ladder. Second, we decompose sources of net employment growth for hiring firms into the different types of gross flows: hires and separation from/to unemployment and from/to employment via poaching. This decomposition varies systematically with the firm states (n, z) that determine marginal surplus. Third, we establish that our framework generalizes existing work by studying the limiting behaviors of our economy. As *decreasing returns to scale* vanish, the economy converges to one in which single-worker firms operate in a frictional labor market à la Postel-Vinay and Robin (2002). As *frictions* vanish, the economy converges to one in which multi-worker firms operate in a competitive labor market à la Hopenhayn (1992). Surprisingly, on the job search is necessary for this result, as it provides the mechanism that equates the marginal products of labor across firms in the limit. As in Hopenhayn (1992), the limit features a non-degenerate firm size distribution. This is in sharp contrast to the frictionless limit of an economy with constant returns which would see one firm hire all workers in the economy.

Our third contribution exploits the computational tractability of our framework to quantitatively analyze equilibrium firm and worker reallocation. We estimate the model by Simulated Method of Moments, targeting cross-sectional moments of the size distribution of firms, firm dynamics, job flows and worker flows for the U.S. economy. We argue that parameters are well-identified.

As a test of the model, we show that our theory is quantitatively consistent with new facts from US employer-employee match data (Haltiwanger, Hyatt, Kahn, and McEntarfer, 2018). In the data, job-to-job flows vary systematically across firms: young firms poach workers from older firms, but firm size is only weakly correlated with net poaching. In our model, with decreasing returns, a small, young, high productivity firm that is yet to grow, has a high marginal value of a worker which places it near the top of the job ladder. Meanwhile, older firms that are small have reached that size because of low productivity, which places them at the opposite end of the ladder. Both are small, but the young firms' vacancies are more likely to attract workers from competitors. To guide future measurement we show that average labor productivity and firm growth are observables that are strongly positively correlated with marginal surplus, so predictive of net poaching and job ladder rank.

We then consider three applications of our model in which we highlight the misallocation effects of labor market frictions along three dimensions of the data: cross-section, firm life-cycle, and time-series.

First, we show that an increase in match efficiency that, by alleviating search frictions, drives unemployment close to zero also accelerates worker reallocation to more productive firms, and in doing so reduces the cross-sectional misallocation of labor across firms and raises TFP by nearly 5 percent.

In our second application, we argue that allowing for misallocation via labor market frictions overcomes a shortcoming of competitive firm dynamics models first identified by Luttmer (2011). In these environments, jointly matching the volatility of firm growth and the size of young firms requires, respectively, small shocks and very low productivity of entrants. Consequently, firms take several hundreds of years to reach the tail of the size distribution, in contrast to the data where this transition is much faster. In our environment, labor market frictions impede to reach the optimal size instantaneously and allow a distribution of young firms that are all small in size, but in which some have very high productivity. With decreasing returns and on the job search these firms are at the very top of the job ladder, and move quickly toward the tail of the size distribution.

Third, our model offers an intuitive interpretation for firm and worker dynamics around the Great Recession, and links them to the observed decline in TFP through a worsening in the degree of frictional misallocation of labor. The recession featured a sharp drop in firm entry and a decline in job-to-job reallocation of workers, which has been characterized as a 'failure of the job ladder' (Siemer, 2014; Moscarini and Postel-Vinay, 2016). Theoretically, our model offers a unifying explanation. A transitory shock to the discount rate, which is a commonly used stand-in for worsening financial frictions (Hall, 2017), lowers the value of entry and shrinks the population of young, high marginal surplus firms with high equilibrium net poaching rates. Vacancy posting collapses among these firms and labor reallocation up the ladder breaks down. Quantitatively, the shock generates the empirical contractions in aggregate employment, job-to-job mobility, firm entry, vacancies and output. In the cross-section, the model matches the decline in net poaching at high productivity firms, and increase in net poaching at low productivity firms (Haltiwanger, McEntarfer, and Staiger, 2021). The resulting rise in frictional misallocation of labor causes a slump in total factor productivity that accounts for a quarter of the large decline in output.

Collectively, these applications demonstrate that our new theoretical framework is a useful platform to jointly analyze the microeconomic dynamics of firms and workers in a frictional labor market, and how these shape macroeconomic outcomes.

Literature

Our paper connects two literatures that share an idea going back to Lucas (1978): the dominant force that delivers a non-degenerate firm-size distribution is the combination of diminishing returns in production and heterogeneity in productivity.

The first literature studies equilibrium models of single-product firm dynamics with competitive labor markets. Classic examples are Hopenhayn (1992), Hopenhayn and Rogerson (1993), and Luttmer

(2011).[4] Recent examples, with applications to the Great Recession, are Arellano, Bai, and Kehoe (2019), Clementi and Palazzo (2010) and Sedláček (2020). Like these models, our framework features entry, exit, and non degenerate distributions of firm size and age. Unlike these models, the employment adjustment costs that firms face are endogenous. They depend on the firm's likelihood of poaching workers from competitors and the expected transfers this requires. Both are a function of the firm rank on the marginal surplus ladder, which is an equilibrium object.

The second literature comprises a number of papers that model multi-worker firms in frictional labor markets. Here, two approaches have been taken: directed search and random search.

Under the directed search approach, Kaas and Kircher (2015) and Schaal (2017) generate firm employment dynamics resembling those in the micro data.[5] Building on Menzio and Shi (2011), Schaal (2017) allows for on the job search, and thus is the closest counterpart to our framework. A drawback of directed search is that the probability that a firm hires from a competitor versus from unemployment is not determined.[6] As a result, this class of models cannot speak to the systematic variation across firm types in net poaching rates or the composition of hires. A model consistent with these facts is one of the objectives of our analysis.

Under the random search approach, Elsby and Michaels (2013) and Acemoglu and Hawkins (2014) solve models where firms face decreasing returns in production, stochastic productivity, linear vacancy costs, and wages determined by Nash bargaining.[7] Both generate employment relationships with a large average surplus and small marginal surplus. Elsby and Michaels (2013) demonstrate that the latter yields a volatile job-finding rate over the cycle, while the former avoids a high separation rate. This resolves the tension identified by Shimer (2005) in the Diamond-Mortensen-Pissarides framework. Gavazza, Mongey, and Violante (2018) introduce recruiting intensity and financial constraints to account for the sharp drop in aggregate match efficiency around the Great Recession. All of these models abstract from search on the job.[8]

Random search models with wage posting feature both on-the-job search and a firm-size distribution.

[4] For a review of the literature see also Luttmer (2010).

[5] It is worth remarking that these two papers had very different objectives to ours. Kaas and Kircher (2015) illustrate that a key advantage of directed search, the efficiency and block-recursivity properties of equilibrium, extends to models with 'large' firms. Schaal (2017) proves this property is also robust to the addition of on-the-job-search and studies aggregate uncertainty shocks in the context of the Great Recession.

[6] In the equilibrium of directed search models, net hiring costs are equated across firms through free entry, which implies that firms are indifferent across the markets in which they search for workers. The probability that a separation from a firm is to employment or unemployment, however, is determined.

[7] Bertola and Caballero (1994) derive closed form results under a linear approximation to both marginal product and convex vacancy costs, and a two state Markov process for productivity.

[8] Fujita and Nakajima (2016) introduce on-the-job search and study the dynamics of job-job flows over the business cycle. However, solving their equilibrium requires worker's outside option to always equal value of unemployment. Hence workers are always indifferent between searching/working and staying/moving.

Bilal and Lhuillier (2021) introduce decreasing returns in the steady-state Burdett and Mortensen (1998) environment, but handling productivity shocks remains out of reach. With constant returns to scale, out of steady-state dynamics in the Burdett and Mortensen (1998) model are studied by Moscarini and Postel-Vinay (2013, 2016), Coles and Mortensen (2016), Engbom (2017), Gouin-Bonenfant (2018) and Audoly (2019). In these models the size distribution is non degenerate only because of the existence of search frictions: as frictions disappear, all workers become employed at the most productive firm. Instead, as explained, the frictionless limit of our model is a version of Hopenhayn (1992). Another implication of such environments is that large firms, which pay higher wages in the model, should systematically poach from small firms, while the data suggest otherwise.

Within the random search literature, we build on the set-up developed by Postel-Vinay and Robin (2002): Bertrand competition between employers for workers and wage renegotiation under mutual consent. This environment has become another workhorse of the literature due to its tractability and empirically plausible wage dynamics.[9] As opposed to the Postel-Vinay and Robin (2002) framework, the probability of filling a vacancy in our model is not a function of the exogenous distribution of firms' productivity, but a function of the *endogenous* distribution of firms' marginal surpluses, which itself depends on how the equilibrium of the frictional labor market allocates workers across heterogeneous firms. Kiyotaki and Lagos (2007) develop a version that is a step closer to us. Their firms have a capacity of one position—an extreme version of decreasing returns— and when an occupied firm meets another worker, it engages in renegotiation with its incumbent worker. Our contribution is to generalize this sequential auction protocol to multi-worker firms, show how one can still solve the model's equilibrium through the notion of joint surplus, and do so in a tractable way.

The final expression for joint surplus that features among our equilibrium conditions is reminiscent of that in Lentz and Mortensen (2012): a version of Klette and Kortum (2004) with on-the-job search in which a firm's demand for labor is limited by demand for its portfolio of products. While they assume that all decisions are based on joint firm-workers values, we derive this result from primitives, provide a characterization of the equilibrium and illustrate how to use the model for a quantitative analysis of newly documented empirical patterns. Our central finding that a job ladder in marginal surplus arises in equilibrium is closely related to contemporaneous work by Elsby and Gottfries (2021) who elegantly characterize a special case of our environment with linear vacancy costs and no endogenous entry and exit. In their setting, firm value and policies are a function of a single state variable, the marginal product of labor. In our theory, marginal surplus is related to the current and future marginal products of labor,

[9] Recent examples are Postel-Vinay and Turon (2010); Jarosch (2021); Lindenlaub and Postel-Vinay (2016); Borovicková (2016); Lise and Robin (2017).

and also depends on average surplus through the exit decision. Nevertheless, in our calibrated model, the correlation between marginal surplus and the current marginal product of labor is high. Finally, the endogenous entry decision, not in Elsby and Gottfries (2021), is at the heart of three main applications of our model.

Outline. Section 2 establishes the physical and contractual environment. Section 3 states our joint value representation and its key properties. Section 4 defines an equilibrium and characterizes firm dynamics and worker flows. Section 5 estimates the model on US data and discusses the model's fit. Section 6 uses the estimated model to examine, under different angles, how search frictions impede reallocation of labor across firms. Section 7 concludes. The Appendix contains all proofs and a discussion of identification.

2 Model

2.1 Physical environment

Time is continuous and there is no aggregate uncertainty. There are two types of agents. An exogenous mass $\bar{\mathtt{n}}$ of ex-ante identical, infinitely-lived *workers* that are risk neutral, discount the future at rate ρ and are endowed with one unit of time each period which is inelastically supplied to production. An infinite mass of homogeneous *potential firms*, of which an endogenous mass become *operating firms*.

Production technology. There is a single homogeneous good. Workers may either be employed or unemployed. Unemployed workers produce b units of the final good. A firm with productivity $z \in Z$ employing n workers produces $y(z, n)$ units of the final good, where $y(z, n)$ is strictly increasing in z and n and concave in n, i.e. $y_{nn}(z, n) \leq 0$.[10],

Firm demographics. A potential firm becomes an operating firm by paying a fixed cost c_0. This cost entitles the firm to a draw of productivity z from the distribution $\Pi_0(z)$ and to n_0 workers, taken from unemployment. After entry, z evolves stochastically. At any point in time a firm may exit, at which point all of its workers become unemployed and the firm produces $\vartheta > 0$ units of the final good which we refer to as its *scrap value*. Denote the mass of entrants $\mathtt{m}_0$ and the mass of operating firms $\mathtt{m}$.

Matching technology. Hiring firms and job-seekers meet in a frictional labor market. The total number of meetings is given by the CRS aggregate matching technology $m(\mathtt{s}, \mathtt{v})$. Inputs to this function are total

[10] In addition, we assume that for any z the Inada conditions hold with respect to n: (i) $y(z, 0) = 0$, (ii) $\lim_{n\to 0} y_n(z, n) = +\infty$, and (iii) $\lim_{n\to +\infty} y_n(z, n) = 0$.

vacancies $\mathtt{v}$ and total units of search efficiency $\mathtt{s} = \mathtt{u} + \xi(\bar{\mathtt{n}} - \mathtt{u})$, where the parameter ξ determines the relative search efficiency of employed workers (labor force minus the unemployed). Search is random in the following sense. A firm pays a cost $c(v; z, n)$ to post v vacancies, where c is increasing and convex in v. Each vacancy is matched with a worker at rate $q(\mathtt{s}, \mathtt{v}) = m(\mathtt{s}, \mathtt{v})/\mathtt{v}$. The worker is unemployed with probability $\phi = (\mathtt{u}/\mathtt{s})$, and employed with probability $(1 - \phi)$. A worker faces no cost of search. An unemployed worker meets a firm at rate $\lambda^U(\mathtt{s}, \mathtt{v}) = m(\mathtt{s}, \mathtt{v})/\mathtt{s}$. An employed worker meets a firm at rate $\lambda^E(\mathtt{s}, \mathtt{v}) = \xi\lambda^U(\mathtt{s}, \mathtt{v})$. The rates q and λ^U can be expressed in terms of *market tightness* $\theta = (\mathtt{v}/\mathtt{s})$. If constituted, the match of a worker to a firm exogenously expires at rate δ, and the worker becomes unemployed.

States. Let x be the vector of state-variables for the firm. This vector includes all individual state variables of all workers at the firm. For now, we do not specify exactly what is in x and, along the way, define a number of functions that map x at instant t into a new state vector at $t + dt$. Let $H(x)$ be the measure of x across firms in the economy, $v(x)$ the number of vacancies created by a firm with state x, and $n(x)$ employment at firm x. The total mass of vacancies and employed workers in the economy are

$$\mathtt{v} = \int v(x)\, dH(x) \quad , \quad \mathtt{n} = \bar{\mathtt{n}} - \mathtt{u} = \int n(x)\, dH(x) .$$

Densities that appear in agents' problems describe vacancy- and employment-weighted distributions:

$$h_v(x) = \frac{v(x)h(x)}{\mathtt{v}} \quad , \quad h_n(x) = \frac{n(x)h(x)}{\mathtt{n}} .$$

Timing. We separate the within-dt timing of events in the model into two parts.

First, events up to the opening of the labor market are described in Figure 1. A firm's productivity z is first realized. Next, incumbent workers are fired, choose whether to quit the firm, or their employment contracts are renegotiated. Next, the firm decides whether to stay in operation or exit. An operating firm produces $y(z, n)$, pays wages according to contracts with its workers, and posts vacancies.

Second, the mutually exclusive events that may occur to a worker or firm are described in Figure 2.[11] The first branch in Figure 2 describes events that may occur to an unemployed worker. The second and third branch distinguish between direct and indirect events that may affect the value of incumbent worker i. *Direct* events involve worker i meeting with another firm, or the destruction of the worker's job. *Indirect* events involve worker i's co-worker j meeting with another firm, or the destruction of a co-worker's job. The final branch describes events that directly impact the firm. The firm may meet an

[11] The mutual exclusivity property is a consequence of continuous time.

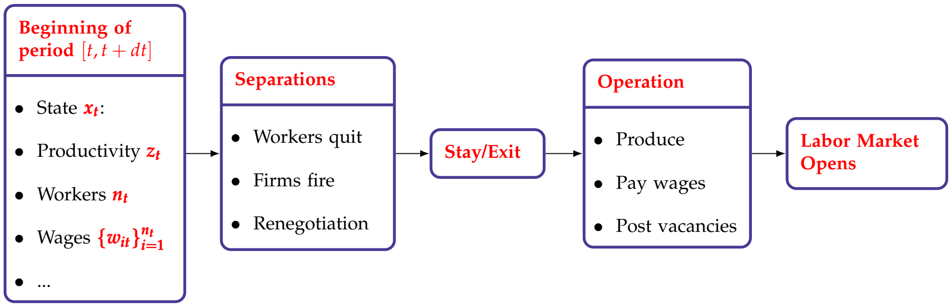


Figure 1: Timing of events prior to the opening of the labor market

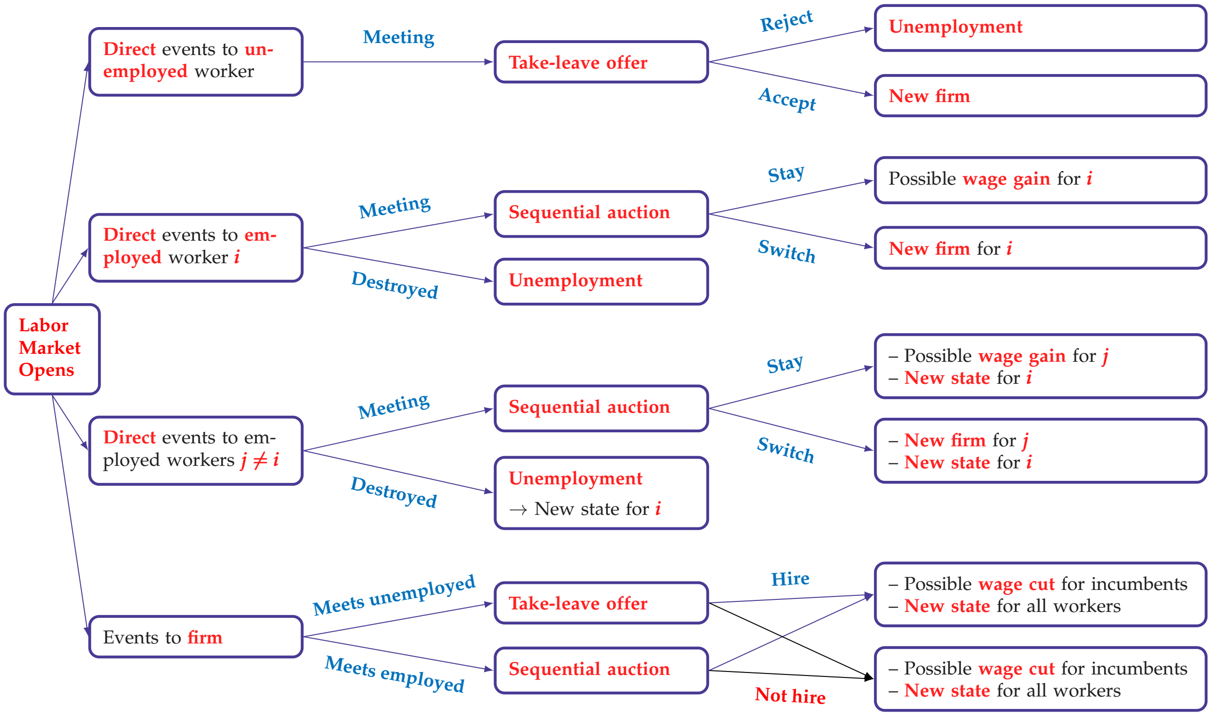


Figure 2: Labor market: Set of mutually exclusive possible labor market events

employed or unemployed worker, emerge either with a new hire or not and new allocation of values to its workers, reflected in updates to the state x. Following any of these events, the state vector x changes, potentially affecting the value of the match to worker i. Through the following assumptions, we put structure on the states in which these events occur and how values evolve in each case.

2.2 Information and contractual environment

The information structure is such that everything that is payoff relevant is observable by both firms and workers, and thus we rule out private information by assumption.[12] The contractual environment is rooted in incomplete contract theory, where there is a key distinction between the information available to firm and workers, and what is verifiable by a third party, e.g. a court, and hence contractible. The only verifiable and hence contractible objects are the wage, whether the firm made the wage payment, and whether the worker provided labor services. Therefore, a contract between the firm and one of its workers is a binding agreement that specifies a constant wage, i.e. a fixed payment from the firm to the worker, in exchange for labor services. This contract satisfies five assumptions:

(A-LC) **Limited commitment.** All parties are subject to limited commitment. In particular,

(a) **Layoffs -** Firms can fire workers at will.

(b) **Quits -** Workers can always quit into unemployment or to another firm when they meet one.

(c) **Collective agreements -** Workers cannot commit to any other worker inside the firm. *De facto* this assumption rules out transfers among workers.

(A-MC) **Mutual consent.** The wage (contract) can be renegotiated only by mutual consent, i.e. only if one party can credibly threaten to dissolve the match (the firm by firing, the worker by quitting). A threat is credible when one of the two parties has an outside option that provides her with a value that is higher than the value under the current contract.

(A-EN) **External negotiation.** An *external negotiation* is a situation where, through search, the firm comes into contact with an external job seeker or an incumbent worker comes into contact with another firm. In external negotiations, all offers are *take-it-or-leave-it*.

- In a meeting with an unemployed worker, the firm makes a take-leave offer to the worker.
- In a meeting with an employed worker, the two firms Bertrand compete through a *sequential auction*. First, the poaching firm makes the take-leave wage offer. Second, the target firm makes a take-leave counteroffer to the worker. Third, the worker decides.

(A-IN) **Internal negotiation.** An *internal negotiation* is any other situation where contracts between firm and any incumbent workers are modified (following **(A-MC)**, an internal negotiation takes place

[12] For example, the number of vacancies posted by the firm is observable to workers, and whether a particular incumbent worker has an outside offer (as well as the identity of the competing firm) is observable to the current firm and other incumbent workers. See Lentz (2015) for an environment where on-the-job search behavior, including the identity of firms in outside meetings, is unobservable to the firm and cannot be directly contracted upon.

when any party has a credible threat). The only parties involved in an internal negotiation are those that have a threat and those that are under that threat. We assume that—with respect to worker and firm values—the internal negotiation is a *zero-sum game* and that participation is individually rational for all parties.[13] Apart from these assumptions we leave internal negotiation unrestricted.

(A-VP) **Vacancy posting.** The firm posts the privately efficient amount of vacancies, which is the one that maximizes the sum of the values of the firm and its workers. Below we propose one possible micro-foundation for **(A-VP)**.

Discussion of assumptions. Assumption **(A-LC)** implies an environment with at-will employment. **(A-MC)** is common under incomplete contracts and in the terminology of MacLeod and Malcomson (1989) yields *self-enforcing contracts*, a feature consistent with most legal frameworks. **(A-EN)** is a particular protocol to resolve the game between two firms competing for a worker. Combined, these three assumptions amount to the contractual environment of Postel-Vinay and Robin (2002). The authors show that they lead to a joint value representation in the one-worker-one-firm model. We now discuss how **(A-IN)** and **(A-VP)** are sufficient to extend this convenient representation to an environment with multi-worker firms and diminishing marginal product of labor.

(A-IN) is a standard assumption in virtually all bargaining protocols. As such, it allows for a large class of possible micro-foundations for the internal renegotiation game. Each would imply different wage dynamics. The central takeaway is that, no matter the details of such a game and the ensuing wages, if **(A-IN)** is satisfied then *allocations* are uniquely determined by joint value dynamics. This paper focuses on *allocations*, i.e. firm and worker dynamics. We leave for future research an investigation of what different internal renegotiation games imply for wage dynamics at the firm and worker level, and which is most consistent with data on wages.

(A-VP) is admittedly a strong assumption, but necessary to simplify the environment for analytical characterization and quantitative analysis. Absent **(A-VP)**, the firm would over-post vacancies relative to the privately efficient amount. The incentive to over-post comes from credible threats to layoff incumbents and hence lower their wages. First, *over-hiring* threatens layoffs by lowering the marginal product of labor, as extensively discussed by Stole and Zwiebel (1996) and Brügemann, Gautier, and Menzio (2018). Second, if a posted vacancy matches with a job seeker with a low outside option, the firm may have no intention of hiring but the match nonetheless generates a threat to *swap* the incumbent

[13] We adopt the standard definition of a zero-sum game: each individual's gain or loss is exactly offset by losses and gains of other participants. We also adopt the standard definition of individual rationality: after internal negotiation each player who remains employed at the firm receives at least the outside option that was present before internal negotiation.

worker. Proceeding under either would require the full distribution of wages as a state variable, ruling out tractability. Assumption **(A-VP)** resolves these issues.[14]

The presence of these inefficiencies and the need for an assumption like **(A-VP)** is unique to an environment with DRS, on-the-job search and endogenous vacancy posting. With DRS and on-the-job-search, but exogenous contact rates (as in Kiyotaki and Lagos, 2007), there is no endogenous vacancy choice and these inefficiencies do not arise. With on-the-job search and constant returns, over-hiring does not occur due to a constant marginal product of labor (Postel-Vinay and Robin, 2002), and hiring a worker that matches with a vacancy is always profitable leaving the swap threat hollow. Without on-the-job search but with decreasing returns, incumbents are all hired from unemployment and with the same outside option are paid the same wage (Elsby and Michaels, 2013; Acemoglu and Hawkins, 2014). Over-hiring occurs, but with a degenerate distribution of wages within the firm this does not impede tractability, and swapping is not a threat because the job seeker and incumbent are paid the same wage. If not addressed, both inefficiencies would render the model intractable.

We propose one possible micro-foundation for assumption **(A-VP)**. The idea is to ex-ante remove any gains to the firm from expected future wage cuts that would otherwise encourage excess vacancy posting. We assume that workers anticipate that firm's behavior and offer a preemptive wage cut that leaves the firm indifferent between the efficient vacancy policy and the firm's privately optimal policy.[15]

(A-VPI) After the firm announces its proposed vacancies for dt, a randomly selected incumbent worker has the opportunity to make a take-leave counter-offer to the firm. The counter-offer specifies acceptable wages for incumbents in exchange for an alternative spot vacancy policy.[16]

We conclude by noting that that in directed search environments, full state-contingent contracts and one-sided commitment by firms deliver bilateral efficiency between a one-worker-one-firm pair (Menzio and Shi, 2011), and private efficiency between a firm and its many workers (Schaal, 2017). We extend this literature by showing that a similar joint value representation can also be achieved in an environment with random search, incomplete contracts, and no commitment.

[14] In a different environment Hawkins (2015) allows full commitment to a fixed wage. This assumes away wage cuts, and hence delivers efficient vacancy posting.

[15] Alternative implementations could be based on the introduction of 'social norms' that prevent firms from cutting the wage of a worker and swapping an incumbent worker with a new worker. Because they would involve deviations from lack of commitment **(A-LC)**, we do not emphasize these alternative implementations in this paper.

[16] This assumption does not require commitment because it is not state-contingent. It is a 'spot contract' between the parties involved: a transfer in exchange for an immediate action.

3 Joint value representation

Having described the economy's environment and the contract space, we now describe the main theoretical result of the paper. For presentation purposes, the environment is specialized in two ways. First, each firm employs a continuum of workers n. Second, productivity follows a diffusion $dz_t = \mu(z_t)dt + \sigma(z_t)dW_t$.[17]

Result. All *allocative decisions* in the economy—entry, exit, vacancy posting and mobility of workers between firms—are determined by the *joint value*, $\Omega(z,n)$. The joint value equals the present discounted value of an operating firm's profits plus the present discounted value of lifetime utility of its incumbent workers, and satisfies the following, where U is lifetime utility of an unemployed worker:

$$\begin{aligned}
\rho\Omega\left(z,n\right) = & \quad \max_{v\geq 0}\, y\left(z,n\right) - c\left(v;z,n\right) && (1)\\
\textit{EU job destruction}: & \quad + \quad \delta n\Big[U - \Omega_n\left(z,n\right)\Big] \\
\textit{UE unemployed hire}: & \quad + \phi q(\theta)v\Big[\Omega_n\left(z,n\right) - U\Big] \\
\textit{EE poaching hire}: & \quad + (1-\phi)q(\theta)v\int \max\Big\{\Omega_n\left(z,n\right) - \Omega_n\left(z',n'\right)\,,\,0\Big\}dH_n\left(z',n'\right) \\
\textit{Shock}: & \quad + \mu(z)\Omega_z\left(z,n\right) + \frac{\sigma(z)^2}{2}\Omega_{zz}\left(z,n\right).
\end{aligned}$$

Firms' operation requires (z,n) to be interior to an *exit boundary*, and an additional *layoff boundary* determines when separations occur:

$$\textit{Exit boundary}: \quad \Omega(z,n) \geq \vartheta + nU, \quad , \quad \textit{Layoff boundary}: \quad \Omega_n(z,n) \geq U. \tag{2}$$

Conditions (1) and (2) represent the solution of the Hamilton-Jacobi-Bellman variational inequality, which we include for completeness in Appendix B, equation (22), along with a discussion on the derivation of the boundary conditions. The entry decision can be written in terms of joint value as:

$$\textit{Entry condition}: \quad \int \Omega(z,n_0)d\Pi_0(z) \geq c_0 + n_0U \tag{3}$$

The first term in (1) is simply output net of vacancy costs. Next, the firm exogenously loses one of its n workers to unemployment at rate δ. The separated worker receives the value of unemployment U, and the remaining workers and firm see their joint value decline by the marginal value of the lost worker.

[17] As we show in the Appendix of Bilal, Engbom, Mongey, and Violante (2019), our results also hold with an integer-valued workforce and when the productivity process is a jump-diffusion.

The firm hires by posting vacancies which are matched to a worker at rate $q(\theta)$. The probability that this worker is unemployed is ϕ, and the firm always hires unemployed workers. This investment increases the value of the firm and incumbents by Ω_n but dilutes their equity, as U is pledged to the new worker. The firm also hires from other firms by poaching. Workers at other firms are met according to the employment-weighted distribution of productivity and size, H_n. Upon meeting, the total value increases by $\Omega_n(z,n) - \Omega_n(z',n')$. The first term is the gain in value to the firm and incumbent workers due to the new hire. The second term is the value that the firm and incumbent workers pledge to the new worker, which is equal to the highest value its former employer would pay to retain them. Hence poaching is successful if this difference is positive and workers flow to the highest marginal value firm.[18]

Conversely, an incumbent worker may quit to a higher marginal value firm. The firm and remaining workers will lose $\Omega_n(z,n)$ and so are prepared to increase the worker's value by $\Omega_n(z,n)$ to retain them. Knowing this, the external firm hires the worker by offering the worker exactly $\Omega_n(z,n)$. The joint value of the firm, remaining workers and poached worker are therefore unchanged and, as in Postel-Vinay and Robin (2002), no *'EE Quit'* term appears in (1).[19]

Boundary conditions (2) describe firm exit and layoffs. First, firms operate if the value of doing so exceeds the joint value of exit: the scrap value ϑ plus unemployment for its n workers. Second, if productivity falls, the marginal value of a worker will fall, but must remain above the opportunity cost of employment. To ensure this, firms layoff workers to sustain $\Omega_n(z,n) \geq U$. Finally, (3) states that firms enter if the joint value of operating $\Omega(z,n_0)$ net of the entry cost c_0 exceeds the joint outside value for the n_0 initial employees.

The joint value representation has three appealing properties.

3.1 Properties of the joint value representation

(1) Parsimony. Firm and worker policies are characterized by a low-dimensional state vector: productivity and size. Given decreasing returns to scale in production and on-the-job search, this simplification is a contribution. First, with decreasing returns spillovers exist as bargaining moves from one worker to the next. This problem has been addressed in environments where workers have homogeneous outside options, which restricts attention to labor market transitions between employment and unemploy-

[18] This term reads as if the poaching coalition induces a breach of contract between worker and former employer, and compensates the latter exactly for the associated loss of value. This scheme is reminiscent of the result in Diamond and Maskin (1979) and Kiyotaki and Lagos (2007) that compensatory damages in breach of contracts restore efficiency.

[19] This result implies that if workers' search effort was costly and endogenous, its privately efficient level would be zero, and thus workers' job to job transitions would only occur through exogenous contacts. To make search effort salient, one needs to modify (A-EN) and introduce positive bargaining power for workers in the contractual environment.

ment.[20] With on the job search, however, past offers create heterogeneous outside options within the firm, precluding these approaches. Second, in models with on-the-job search these bargaining spillovers are assumed away either by (i) constant returns to scale, which reduces decision making units to one-worker-one-firm pairs and impedes a proper study of firm dynamics; or (ii) the combination of full commitment to complex state-contingent contracts and directed search. Our contribution is to prove that a plausible set of minimal assumptions on the contractual environment is sufficient to micro-found a parsimonious representation of allocations.

(2) Private efficiency. All agents' decisions (entry, exit, separations, vacancies, and hires) maximize their joint value. Put differently, in external and internal negotiations all privately attainable gains from trade are exploited such that no transfer could yield a Pareto improvement. Crucially, our assumptions on bargaining do not directly imply this result. Instead, these assumptions provide the basis upon which decentralized negotiation decisions lead to a privately efficient outcome in all scenarios that matter for allocations: when the firm negotiates internally with its incumbent workers, when the firm negotiates with a potential new hire, and when the firm decides how many vacancies to open. We have therefore shown how the Coase theorem arises in our context without the need to assume full commitment and complex state contingency in contracting.[21]

(3) Endogenous job ladder. In one-worker-one-firm models, it is the firm's *exogenous* productivity that fully determines its position on the job ladder. Here the ladder is in *endogenous marginal values* of labor $\Omega_n(z, n)$. These equilibrium objects are determined by the current marginal product of labor together with expectations of future productivity, worker mobility, exit, market tightness and composition of vacancies and workers across firms and unemployment.

Proof. To convey the economics of how our assumptions lead to this result, we use a static model in Appendix A. One by one, we cover the construction of each term in (1). The approach and arguments for the proof in the case of the dynamic model are extensions of the proof of the static model. While the proof for the static model is compact, the complete proof of the joint value representation for the dynamic model requires much additional notation and is contained in the Appendix of the working paper version (Bilal, Engbom, Mongey, and Violante, 2019).

[20] See Stole and Zwiebel (1996), recently revisited by Brügemann, Gautier, and Menzio (2018).

[21] We leave the characterization of the socially efficient allocations to future work, but note that the decentralized and planner's allocations will not coincide. Besides the standard congestion externality à la Hosios, an additional composition externality arises. As in Acemoglu (2001), low-productivity firms do not internalize that their vacancies divert workers away from high-productivity firms. These distorted vacancy decisions affect the equilibrium distribution of workers across firms $H_n(n, z)$ which, in turn, influences the hiring opportunities of all other firms and distorts output.

3.2 Surplus formulation

A convenient formulation of (1) is in terms of *joint surplus*, defined $S(z,n) := \Omega(z,n) - nU$, such that

$$S_n(z,n) = \Omega_n(z,n) - U \quad , \quad S_z(z,n) = \Omega_z(z,n) \quad , \quad S_{zz}(z,n) = \Omega_{zz}(z,n).$$

Hence the marginal (joint) surplus $S'_n = S_n(z',n')$ at a competitor is sufficient to characterize how surplus changes over an EE hire. With these definitions, along with the value of unemployment $\rho U = b$, the joint value (1) becomes joint surplus:

$$\begin{aligned} \rho S\left(z,n\right) \;\; = \;\; \max_{v\geq 0} \quad & y\left(z,n\right) \; - \; nb \; - \; \delta n S_n\left(z,n\right) + \mu(z) S_z\left(z,n\right) \; + \; \frac{\sigma^2(z)}{2} S_{zz}\left(z,n\right) \\ & + q(\theta) v \left[\phi S_n\left(z,n\right) \; + \; (1-\phi) \int_0^{S_n(z,n)} \Big(S_n\left(z,n\right) - S'_n \Big) \; dH_n\left(S'_n\right) \right] - c\Big(v; z, n\Big) \end{aligned} \tag{4}$$

subject to the same two boundary conditions now expressed in terms of surplus:

$$S(z,n) \geq \vartheta \quad \text{for exit, and} \quad S_n(z,n) \geq 0 \quad \text{for layoffs.} \tag{5}$$

Properties of $S(z,n)$. To analyze worker and firm dynamics we first establish some properties of the joint surplus function under standard assumptions on technology. Suppose (i) productivity follows a geometric Brownian motion with $\mu(z) = \mu \cdot z$, $\sigma(z) = \sigma \cdot z$, (ii) the vacancy cost function is isoelastic in vacancies only $c(v) = \bar{c} v^{1+\gamma}$, and (iii) the production function satisfies $y_z > 0, y_n > 0, y_{nn} < 0, y_{zn} > 0$.[22] In Appendix B we show that under these assumptions the following Properties hold inside the boundaries: **(P1)** S is increasing and concave in employment: $S_n > 0$, $S_{nn} < 0$; **(P2)** S is increasing in productivity: $S_z > 0$; **(P3)** S is supermodular in productivity and labor: $S_{zn} > 0$. We now combine these with the surplus formulation to characterize firm optimal polices.

3.3 Vacancy policy

From (4), the first order condition for the firm's vacancy decision gives

$$c_v\big(v; z, n\big) = q(\theta) R\big(S_n(z,n)\big) \quad , \quad \text{where} \quad R(S_n) = \phi S_n + (1-\phi) \int_0^{S_n} \Big(S_n - S'_n \Big) \, dH_n\left(S'_n\right) \tag{6}$$

The return on a vacancy R is independent of v, and is a strictly increasing and strictly convex function of only marginal surplus:

$$R'(S_n) = \underbrace{[\phi + (1-\phi) H_n(S_n)] \cdot 1}_{\text{Higher surplus on each hire}} \; + \; \underbrace{(1-\phi) h_n(S_n) \cdot 0}_{\text{Surplus on additional hires} = 0} \qquad , \qquad R''(S_n) = (1-\phi) h_n(S_n)$$

[22] The functional form assumed in our quantitative analysis satisfies these assumptions: $y(z,n) = zn^{\alpha}$ with $\alpha \in (0,1)$.

On the *intensive margin*, an increase in S_n increases the return to hiring an unemployed or employed worker one-for-one. On the *extensive margin*, an increase in S_n widens the set of firms from which the firm will poach, increasing the probability of a hire by $(1-\phi)h_n(S_n)$, but hiring from these additional firms yields zero additional value as the target firm's marginal surplus associated with the worker is close to that of the poaching firm.

Endogenous hiring cost. The literature on firm dynamics models exogenous employment adjustment costs. Instead, search frictions and the job ladder induce an endogenous firm-specific *hiring cost function* that depends on both equilibrium market tightness and on the firm rank on the job ladder.

The vacancy yield of a firm with marginal surplus $S_n(z,n)$ is $q(\theta)[\phi + (1-\phi)H_n(S_n)]$. Attaining h hires costs $\mathcal{C}(h,n,z,S_n)$, due to the $v(h,S_n)$ vacancies required:

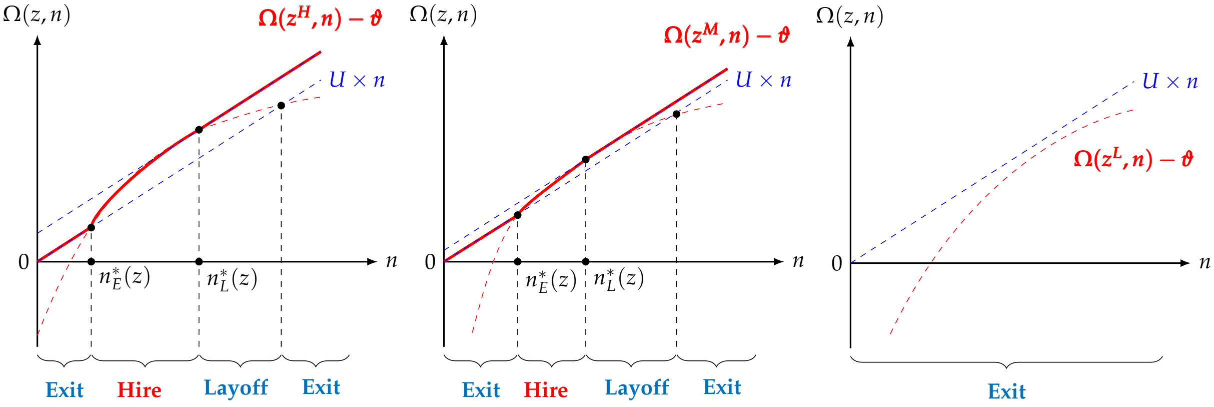
$$\mathcal{C}\Big(h,z,n,S_n\Big) = c\Big(v\Big(h,S_n\Big);\, z,\, n\Big) = c\left(\frac{h}{q(\theta)\left[\phi + (1-\phi)H_n(S_n)\right]};z,n\right). \tag{7}$$

This reduced form hiring cost function is increasing and convex in h and decreasing in marginal surplus, and is also determined by two equilibrium objects: the aggregate distribution of marginal surplus $H_n(S_n)$ and overall market tightness via $q(\theta)$. The cost function (7) makes clear the role of frictions and on-the-job search as endogenous sources of adjustment cost: the cost is low for firms at the top of the job ladder, and for all firms under a slack labor market.[23]

3.4 Hire and separation policies

Figure 3 characterizes the firm's hiring and separation choices for alternative pairs (z,n). Consider panel (a). The red dashed line is the value of hiring net of the scrap value: $\Omega(z,n) - \vartheta$. The lower blue dashed line extending from the origin gives the total value of unemployment to the firms' employees: $U \times n$. The exit threshold $n_E^*(z)$ is determined by their intersection, at which point the per-worker value net of ϑ is equal to the value unemployment: $\left(\Omega(z,n_E^*(z)) - \vartheta\right)/n = U$. If $n < n_E^*(z)$, the firm fires its n workers. As opposed to this condition on *average values*, the layoff threshold $n_L^*(z)$ equates the *marginal value*, i.e. the slope of $\Omega(z,n)$, to U. If $n > n_L^*(z)$, the firm fires $(n - n_L^*(z))$ incumbents who each receive U, and the joint value is $\Omega(z,n_L^*(z)) + (n - n_L^*(z))\,U$. The upper envelope of these choices is given by the solid red line.

[23] To draw a comparison, the standard convex adjustment cost in firm dynamics models is independent of equilibrium objects and depends only on firm employment growth. In the directed search model of Kaas and Kircher (2015) or random search model of Gavazza, Mongey, and Violante (2018) the equilibrium meeting rate $q(\theta)$ enters, but without on the job search there is no additional role for the distribution of firms.



(a) High productivity (b) Medium productivity (c) Low productivity

Figure 3: Values of exit, hiring and layoff for fixed levels of productivity z

Panel (b) and (c) describe these policy regions for lower productivity firms. Under a lower productivity, the exit and layoff regions extend, while the hiring region shrinks (Panel b). At an even lower productivity it is optimal for the firm to exit for all n (Panel c).

3.5 The gross worker flow composition of net employment growth

The model decomposes firms' net job growth into the four worker flows discussed in the introduction: hires from unemployment (UE), poaching inflows (EE^+), separations into unemployment (EU), and poaching outflows (EE^-). Firm's net job growth in the hiring region is given by

$$\frac{dn}{n} = \underbrace{q(\theta)\frac{v(z,n)}{n}\Big[\phi+(1-\phi)H_n(S_n(z,n))\Big]}_{\text{Inflows: }(UE)\text{ and }(EE^+)} - \underbrace{\Big[\delta+\lambda^E(\theta)\overline{H}_v(S_n(z,n))\Big]}_{\text{Outflows: }(EU)\text{ and }(EE^-)}. \tag{8}$$

Under assumptions (i)-(iii) stated above, we can also prove an additional property: **(P4)** Net employment growth dn/n is increasing with productivity z and decreasing with size n. See Appendix B.

Figure 4 illustrates how the four worker flows vary with n for a given level of z. Consider a firm that is at the layoff frontier: $n = n^*_L(z)$. Marginal surplus is zero so the firm posts zero vacancies and shrinks due to exogenous separations and poaching. Conditional on a meeting, any worker employed in that firm leaves $(\overline{H}_v(0) = 1)$, so separations occur at rate $\delta + \lambda^E(\theta)$. As the firm shrinks, decreasing returns in production cause the firm's marginal surplus to increase **(P1)**. In terms of outflows, the firm loses fewer workers to competitors. In terms of inflows, the firm posts vacancies which always generate hires from unemployment and, as marginal surplus increases further, hires from employment too. Firms

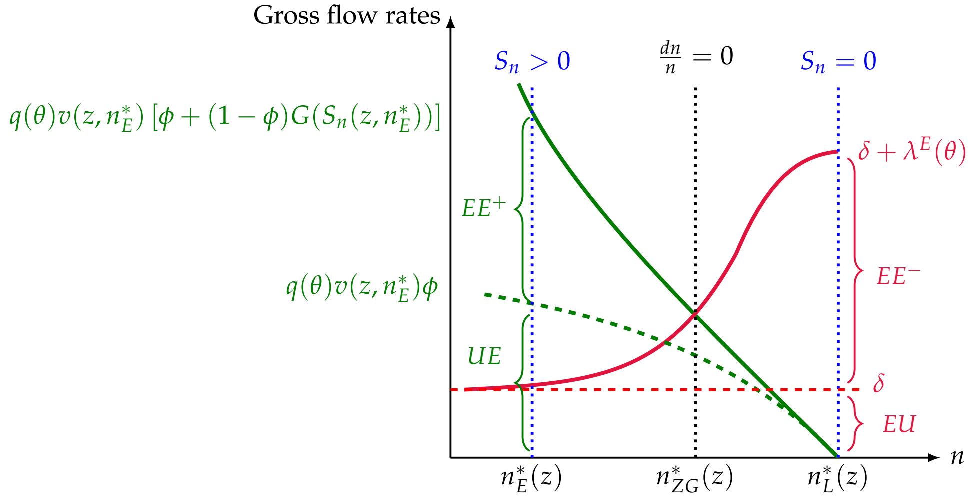


Figure 4: Gross worker flows by employment level, for given productivity

Notes: The solid red curve represents total separations $(EU + EE^-)$ and the dashed red horizontal line exogenous quits EU. The green curve represents total hires $(UE + EE^+)$ and the dashed green curve hires from unemployment (UE).

shrink towards $n_{ZG}^*(z)$ where there is zero growth but gross flows in both directions are still positive. For any given productivity z, the firm with the highest marginal surplus has the smallest size compatible with operating, i.e. size $n_E^*(z)$, and grows quickly away from $n_E^*(z)$ with high vacancy posting and net poaching.

Moreover, if $c(v; z, n) = c(v, S_n)$, then faster growing firms have:

(1) Higher rates of EE^+, lower rates of EE^- and higher rates of net-poaching: $(EE^+ - EE^-)$

(2) Higher shares of hires from employment EE^+ and lower shares from unemployment UE

(3) Higher shares of separations to unemployment EU and lower shares to employment EE^-.

The intuition is simply that fast growing firms have high marginal surplus. For example, the pattern in (2) can be observed from Figure 4. As one moves leftward along the x-axis, S_n increases, the firm's growth rate increases, and EE^+ as a share of total hires increases as well.

We conclude by noting that this type of analysis on the composition of hires by firm size and productivity cannot be performed in current directed search models. As explained in the Introduction, in that class of models, the composition of hires at the firm level is indeterminate.

4 Equilibrium

We formally define an equilibrium, and employ a phase diagram to characterize firm and worker dynamics in (n, z)-space. We also study two limiting equilibria, one where decreasing returns in production vanish, and another where matching frictions vanish.

4.1 Equilibrium

A stationary equilibrium with positive entry is: (i) a joint surplus function $S(z,n)$; (ii) a vacancy policy $v(z,n)$; (iii) a law of motion for firm level employment $\frac{dn}{dt}(z,n)$; (iv) a stationary distribution of firms $H(z,n)$; (v) vacancy- and employment-weighted distributions of marginal surplus $H_v(S_n)$ and $H_n(S_n)$; (vi) a positive mass of entrants $\mathtt{m}_0$, (vii) a vacancy meeting rate $q(\theta)$ and conditional probability of meeting an unemployed worker ϕ, such that:

(i) Total surplus $S(z,n)$ satisfies the HJB equation (4) and associated boundary conditions.

(ii) The vacancy policy $v(z,n)$ satisfies the first order condition:

$$c_v(v(z,n);z,n) = q(\theta)\left[\phi S_n(z,n) + (1-\phi)\int_0^{S_n(z,n)} \left(S_n(z,n) - S_n'\right)\, dH_n(S_n')\right].$$

(iii) The law of motion for firm level employment is

$$\frac{dn}{dt}(z,n) = \begin{cases} -\frac{n}{dt} & n < n_E^*(z) \\ q(\theta)v(z,n)\left[\phi + (1-\phi)H_n(S_n(z,n))\right] - n\left[\delta + \lambda^E(\theta)(1 - H_v(S_n(z,n)))\right] & n \in \left[n_E^*(z), n_L^*(z)\right) \\ \frac{n_L^*(z)-n}{dt} & n \geq n_L^*(z), \end{cases}$$

where the notation $\frac{n}{dt}$ denotes a jump of size n, and where the exit threshold satisfies value-matching consistent with (4), and the exit and layoff boundaries satisfy smooth-pasting conditions in productivity and employment:[24]

$$\underbrace{S\left(z, n_E^*(z)\right) = \vartheta}_{\text{Value-matching from (4)}}\ ,\ \underbrace{S_z\left(z, n_E^*(z)\right) = 0\,,\ S_n\left(z, n_E^*(z)\right) = 0 \quad \text{if } \frac{dn}{dt}\left(z, n_E^*(z)\right) < 0\,,\ S_n\left(z, n_L^*(z)\right) = 0}_{\text{Smooth-pasting conditions from (4)}}$$

(iv) Vacancy- and employment-weighted distributions of marginal surplus are consistent:

$$H_v(S_n) = \int \mathbb{1}_{[S_n(z,n)\leq S_n]}\frac{v(z,n)}{\mathtt{v}}dH(z,n) \quad , \quad \mathtt{v} = \int v(z,n)dH(z,n)$$
$$H_n(S_n) = \int \mathbb{1}_{[S_n(z,n)\leq S_n]}\frac{n}{\mathtt{n}}dH(z,n) \quad , \quad \mathtt{n} = \int n\, dH(z,n)$$

(v) The measure of firms $H(z,n)$ is stationary, and admits a density function $h(z,n)$ that satisfies:

$$0 = -\frac{\partial}{\partial n}\left(\frac{dn}{dt}(z,n)\, h(z,n)\right) - \frac{\partial}{\partial z}\left(\mu(z)\, h(z,n)\right) + \frac{\partial^2}{\partial z^2}\left(\frac{\sigma(z)^2}{2}h(z,n)\right) + \mathtt{m}_0\, \pi_0(z)\, \mathbf{\Delta}(n)$$

where $\mathbf{\Delta}$ is the Dirac delta "function" which is zero everywhere except $n = n_0$ where it is infinite.

[24] Smooth pasting conditions obtain only when firms are actually crossing the exit or layoff boundaries. Firms may then choose the exit or layoff boundaries by taking an interior first-order optimality condition. For additional details and discussion see Appendix B.1.

(vi) Entry $\mathtt{m}_0$ is such that the expected value of a new entrant is zero:

$$c_0 = \int S(z, n_0) d\Pi_0(z),$$

(vii) Vacancy meeting rate $q(\theta)$ and conditional probability of meeting an unemployed worker ϕ are consistent with the aggregate matching function given employment $\mathtt{n}$, unemployment $(\mathtt{u} = \bar{\mathtt{n}} - \mathtt{n})$, and vacancies $\mathtt{v}$.

The numerical procedure to compute the equilibrium of the model is described in the Appendix of Bilal, Engbom, Mongey, and Violante (2019).

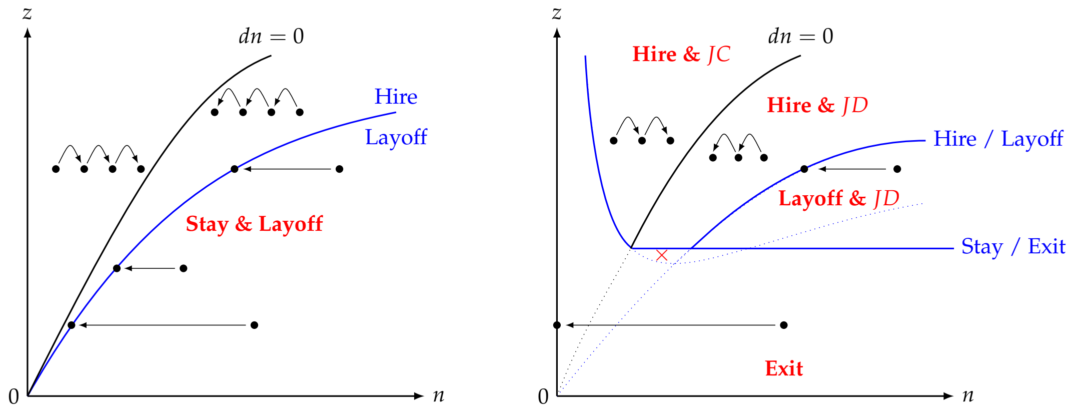
4.2 Firm dynamics, job reallocation and worker turnover: a phase diagram

We now can represent firm dynamics (entry and exit), job reallocation (net growth), and worker reallocation (hires and separations) in the (n, z)-space. Figure 5 describes the functions that determine the stay/exit frontier $n_E^*(z)$, hire/layoff frontier $n_L^*(z)$, and the zero growth locus $n_{ZG}^*(z)$.

First, Panel (a) considers the model without a scrap value such that there is no endogenous exit. From (5) the layoff frontier has slope $dz/dn = -S_{nn}/S_{zn}$. Therefore properties **(P1)** $(S_{nn} < 0)$ and **(P3)** $(S_{zn} > 0)$, imply the layoff frontier is positively sloped. Recall from Figure 3 that, fractionally to the left of the layoff frontier $n_L^*(z)$, $S_n \approx 0$, so vacancy posting is low and the firm shrinks due to EE^- and EU flows. Therefore the zero growth locus along which $dn = 0$ must be located strictly to the left of the layoff frontier. Between the zero-growth locus and the layoff frontier, firms hire but lose even more workers, and so experience job destruction (JD). To the left of the zero-growth locus, marginal surplus is sufficiently large that firms are successful in hiring and retaining workers, and so experience job creation (JC). In all cases some endogenous separations through quits also occur, thus the model generates both hires for shrinking firms and endogenous separations for growing firms. To the right of the layoff frontier, firms fire workers, destroying jobs en masse, and in doing so jump back to the frontier.

Panel (b) introduces a positive scrap value which induces endogenous exit. First, let us ignore the smooth-pasting conditions. From (5) the exit frontier would have slope $dz/dn = -S_n/S_z$. Therefore properties **(P1)** $(S_n > 0, S_{nn} < 0)$ and **(P2)** $(S_z > 0)$, imply the exit frontier would have a minimum at $S_n = 0$, where it crosses the layoff frontier, increasing on either side.

The smooth pasting conditions modify this frontier. A necessary condition for optimal exit is that $S_n = 0$: if marginal surplus was positive on the exit boundary $(S = \vartheta)$, then the firm could post vacancies and increase $S > \vartheta$, and hence would not want to exit. Now recall that by Property **(P1)**, S is strictly concave in n, but optimal layoffs imply $S_n = 0$ on the layoff boundary, implying that S_n is not zero again away from the layoff frontier. Combined, these observations have two implications. First, firms do not



(a) No endogenous exit ($\vartheta = 0$)

(b) Endogenous exit ($\vartheta > 0$)

Figure 5: Exit, layoff and zero-growth frontiers in the (n, z)-space

Notes: This figure plots exit, layoff and zero-growth frontiers for two cases: without and with positive scrap value. It also includes examples of hypothetical firm paths, in each case keeping productivity fixed. A firm (black dot) that begins in the layoff region jumps to the layoff frontier, firing $n - n_S^*(z)$ workers. Subsequent declines in productivity smoothly move the firm along the layoff frontier until, possibly, exit. A firm that is located in the hiring region smoothly converges toward the $dn = 0$ line by growing or shrinking.

exit along the downward sloping section of the exit frontier in the *Hire & JC* region. Indeed, firms in this region have very high marginal surplus and drift to the right. Second, firms cannot be located in the part of the *Hire & JD* region with a red ×-mark. A firm located here would be headed toward exit with $S_n > 0$, which violates optimality. As a result, to the right of the intersection of the zero-growth locus and the $S(z, n) = \vartheta$ locus, the exit frontier is flat.

The stationary distribution of firms in the economy has support along the layoff frontier, and to its left, with zero mass along the left exit frontier. Growing firms do not exit, but shrinking firms may experience productivity shocks that force them over the horizontal section of the exit frontier. All firms—except those on the layoff frontier—post vacancies, and hire workers from employment and unemployment, and lose workers to employment and unemployment.[25] Finally, not depicted in the figure, the mass $\mathtt{m}_0$ of entrants is distributed according to $\Pi_0(z)$, along a vertical line at $n = n_0$. Only those with a high enough value of z keep operating.

4.3 Limiting economies

Our economy includes as special cases two well known frameworks for worker flows on the one hand, and firm dynamics on the other.

[25] The results derived in Section 3.5 regarding gross flows fully describe employment dynamics of the firm when interior to these boundaries. In particular, one can think of Figure 4 as describing gross firm hiring along a straight horizontal line drawn in the (n, z) space of Figure 5 and running from the exit to the layoff frontier.

4.3.1 Vanishing decreasing returns in production

In traditional search and matching models with random, on the job search, the constant returns to scale assumption implies an indeterminate firm size distribution. These are models of *jobs* rather than *firms*, with each job having match output $y(z)$. Here we show that the Bellman equation (4) is a natural generalization of expressions of firm value found in these settings.

Consider the limiting case of (4) when the function y is linear in n, $y(z,n) = zn$. Depending on the form of the recruiting cost function, we obtain either the surplus representation in Lise and Robin (2017) or a slight variant thereof. The formulation in Lise and Robin (2017) arises when vacancy costs are independent of n, $c(v,n) = c(v)$, and the scrap value is zero (no endogenous exit). Under these assumptions (4) becomes affine in size, where the slope gives the value of an existing match and the intercept gives the value of a vacancy:

$$\rho S(z,n) = S_0(z) + n \times \rho \widehat{S}(z) \quad \text{where} \quad \begin{cases} \rho \widehat{S}(z) = y(z) - b - \delta \widehat{S}(z) + \mu(z) \widehat{S}_z(z) + \frac{\sigma^2(z)}{2} \widehat{S}_{zz}(z) \\ S_0(z) = \max_v q(\theta) v \left[\phi \widehat{S}(z) + (1-\phi) \int_0^{\widehat{S}(z)} \left(\widehat{S}(z) - S' \right) dH_n(S') \right] - c(v) \end{cases}$$

These correspond to equations (3), (6) and (7) in Lise and Robin (2017). Since $\widehat{S}$ is independent of n, $H_n(S') = H(z)$. The rank of a firm on the job ladder is determined only by its exogenous productivity z. The value of new jobs therefore depends block-recursively on the value of existing ones. In this limit, firm size is irrelevant for joint surplus and the distribution of marginal surplus.

A similar limiting economy arises when the production function is linear, but vacancy costs are homogeneous of degree one in (v,n) and, again, the scrap value is zero. See Appendix C for details.

4.3.2 Vanishing frictions

We first consider the limit as matching efficiency goes to infinity in the absence of on the job search. We then add on-the-job search and show that this is key for obtaining a version of Hopenhayn (1992) in the limit. Formal proofs are in Appendix D.

Frictionless limit without on-the-job-search. Let A be a scalar in front of the matching function, and take this parameter to infinity in an economy without on-the-job search ($\xi = 0$). Now consider the free-entry condition $\int S(n_0, z) d\Pi_0(z) = c_0$ and the Bellman equation for joint surplus,

$$\rho S(n,z) = \max_v y(n,z) - bn - c(v) - \delta n S(n,z) + qv S_n(n,z) + \mu(z) S_z(n,z) + \frac{\sigma(z)^2}{2} S_{zz}(n,z) \tag{9}$$

The only general equilibrium object in total surplus is q. Therefore there is a unique value of q, independent of A, such that the free-entry condition holds. With q unaffected by the increase in A, firm vacancy policies $v(n, z)$ and firm dynamics conditional on entry remain the same, while the measure of firms adjusts such that q remains constant. Since there exists dispersion in the marginal product of labor for any arbitrary A, then this dispersion continues to exist even in the limit. This limiting result differs from the competitive equilibrium of a frictionless model in which all firms' marginal products of labor are equal and equated to the wage.[26]

Frictionless limit with on-the-job search. On-the-job search breaks this result, allowing equilibrium q to increase. As A increases unemployment $\mathtt{u}$ still goes to zero but, with on-the-job search total units of search efficiency remain positive $\mathtt{s}_\infty = \xi\bar{\mathtt{n}}$. Meanwhile aggregate feasibility ensures that vacancies $\mathtt{v}_\infty$ remain finite. Hence in the limit $\theta_\infty = \mathtt{s}_\infty/\mathtt{v}_\infty$ is constant, and $q = A\theta_\infty^{-(1-\beta)}$ increases in A. As q increases, reallocation of labor via job-to-job quits accelerates. These quits cause marginal surplus to increase at low-S_n origin firms and decline at high-S_n poaching firms, compressing the distribution of S_n to a point.

Our main result, shown in Appendix D, is that in the limit as q goes to infinity, firm behavior is described by the following Bellman equation, employment policy function and boundary condition:

$$\rho S(z) = \max_n y(z, n) - nb + \mu(z)S_z(z) + \frac{\sigma^2(z)}{2}S_{zz}(z) \quad , \quad y_n(z, n^*(z)) = b \quad , \quad S(z) \geq \vartheta \quad , \tag{10}$$

This characterization contains three results. First, with infinitely fast reallocation, there is no dispersion in marginal surplus in equilibrium and hence no surplus gained from on-the-job search. Second, behavior is *as if* firms choose their optimal size each instant, with all hires realized through immediate job-to-job reallocation. Third, this implies the only state variable is z and the productivity-size distribution is degenerate on $(z, n^*(z))$, along which marginal products are equalized. Thus the limit of our model is isomorphic to Hopenhayn (1992) with respect to job reallocation, firm exit, and the expected profits of entrants, which are zero net of entry costs.

We conclude by noting that this natural limiting behavior of our economy stands in stark contrast to traditional search models with constant returns to scale. In bargaining and wage posting models, all employment would go to the most productive firm.

[26] To draw an analogy, consider a comparative statics in the Hopenhayn (1992) model with respect to a parameter that does not directly enter the free entry condition. The effect is a change in the aggregate supply of labor and, thus, the mass of firms, but not their distribution. See also Kaas (2020).

5 Estimation

Having provided a qualitative discussion of the model, we now turn to its quantitative implications. To that end, we estimate the model on U.S. data. Because the model is set and solved in continuous time, we can construct correctly time-aggregated measures at any desired frequency.

5.1 Methodology

We make the following functional form assumptions. The production function is $y(z,n) = zn^{\alpha}$. The vacancy cost function is $c(v,n) = \frac{\bar{c}}{1+\gamma}\left(\frac{v}{n}\right)^{\gamma} v$, as in Kaas and Kircher (2015), such that the per-vacancy cost is increasing in the vacancy rate. The matching function is Cobb-Douglas with vacancy elasticity β: a worker meets a vacancy at rate $f(\theta) = A\theta^{\beta}$ and a vacancy meets a worker at rate $q(\theta) = A\theta^{-(1-\beta)}$. The distribution of entrant productivity draws is Pareto with a minimum of one and shape parameter ζ. Log productivity follows a random walk, $d\log z(t) = \mu dt + \sigma dW(t)$. We add exogenous firm exit at rate d. These assumptions leave 16 parameters to determine. We proceed in three steps.

Externally set or normalized. We normalize or set to standard values five parameters, as summarized in Table 1A. The discount rate ρ implies an annual real interest rate of five percent. The elasticity of the matching function $\beta = 0.50$ is based on standard values in the literature (Petrongolo and Pissarides, 2001). When solving the model we add a fixed cost of operation c_f and set the scrap value ϑ to zero; the two are isomorphic. Without loss of generality we are then able to normalize c_f.[27] From the first order condition for vacancies (6) it is clear that we cannot identify $\bar{c}$ and A separately, so we normalize $\bar{c}$. Finally, we set employment of entering firms, n_0, to one which we interpret as the labor input of the entrepreneur or founder.

Estimated offline. We set three parameters to target directly three moments in the data, as summarized in Table 1B. First, the entry cost c_0 is pinned down by an average firm size of 23. The measure of active firms $\mathtt{m}$ that delivers an average firm size of 23 when there is a unit measure of workers and a non-employment rate of 10 percent is 0.90/23. While $\mathtt{m}$ is an equilibrium outcome, the fact that a higher $\mathtt{m}$ decreases the value of entry through a tighter labor market implies a unique c_0 that satisfies the free-entry condition under a given $\mathtt{m}$. Here, and throughout, we use a broader definition of the pool of job-seekers than in the standard unemployment definition in the CPS. This accounts for the fact that a significant

[27] The argument that we can normalize c_f to one has two pieces. First, for a given fixed cost c_f, we can always choose an entry cost c_0 that generates the desired mass of firms $\mathtt{m}$, see below. Second, with Pareto initial draws of productivity followed by a random walk, a higher fixed cost simply scales the economy up.

Parameter		Value	Moment	Data	Model
			A. Externally set/normalized parameters		
ρ	Discount rate	0.004	5% annual real interest rate		
β	Elasticity of matches w.r.t. vacancies	0.5	Petrongolo and Pissarides (2001)		
c_f	Fixed cost of operation	1	Normalization		
$\bar{c}/(1+\gamma)$	Scalar in the cost of vacancies	100	Normalization		
n_0	Size of entrants	1	Normalization		
			B. Estimated offline		
$\mathtt{m}$	Number of active firms	0.043	Average firm size (BDS)	23.340	20.851
d	Exogenous exit rate	0.000	Exit rate, 1000–2499 empl. firms	0.002	0.002
γ	Curvature of vacancy cost function	3.450	Vacancy filling rate vs. hiring rate	3.450	3.450
			C. Internally Estimated		
μ	Drift of productivity	-0.001	Exit rate (annual)	0.076	0.076
σ	St.d of productivity shocks	0.016	St.d. of log empl. growth (annual)	0.420	0.354
α	Curvature of production	0.817	Empl. share of 500+ firms	0.518	0.527
ζ	Shape of entry distribution	11.844	JC rate, age 1 firms (annual)	0.247	0.255
A	Matching efficiency	0.195	Nonemployment rate	0.100	0.100
ξ	Relative search efficiency of employed	0.151	EE rate (quarterly)	0.048	0.041
δ	Exogenous separation rate	0.017	EN rate (quarterly)	0.056	0.055
b	Flow value of leisure	1.029	JD rate of incumbents (annual)	0.092	0.093

Table 1: Estimated parameters and targeted moments

Notes: Annual firm dynamics moments are from HP-filtered Census BDS data between 2011–2016, with the exception of the standard deviation of annual growth rates, which is from Elsby and Michaels (2013). Quarterly worker flows are from HP-filtered Census J2J data between 2011–2016.

number of hires come directly from out of the labor force and some of our data sources (JOLTS and Census J2J) do not identify whether the origin of hires or destination of separations is unemployment or non-participation.[28]

Second, the exogenous exit rate d is pinned down by an annual exit rate of firms with 1000–2499 employees of 0.2 percent. Such firms may layoff workers, but never exit endogenously in the model, as total surplus is far from the exit frontier.

Third, the vacancy cost elasticity γ is pinned down by the cross-sectional relationship between vacancy- and vacancy-filling rates. In JOLTS microdata, Davis, Faberman, and Haltiwanger (2013) document a nearly log-linear relationship between each of these and the hiring rate, which implies a log-linear relation between them. In the model, these relationships are driven by marginal surplus. The following cross-sectional relationship can be derived from the firm's optimality condition, using a log linear approximation that is valid for firms with small growth rates:

$$\log \frac{v^*(z,n)}{n} \approx \kappa_0 + \kappa_1 \log\left(\frac{h^*(z,n)}{v} - \kappa_2\right) \quad \text{, where} \quad \gamma = \frac{1}{\kappa_1} \tag{11}$$

Firms with high marginal surplus post more vacancies per worker, and fill them more quickly as they

[28] Our definition of the non-employment rate is constructed as follows. The numerator equals the sum of the unemployed (FRED series UNEMPLOY) plus those out-of-the-labor-force who answer that they 'currently want a job' in the CPS (NILFWJN). The denominator equals the sum of the civilian labor force (CLF16OV) plus the same subgroup of those out-of-the-labor-force (NILFWJN). From 2011-2016 this ratio is, on average, just above 10 percent.

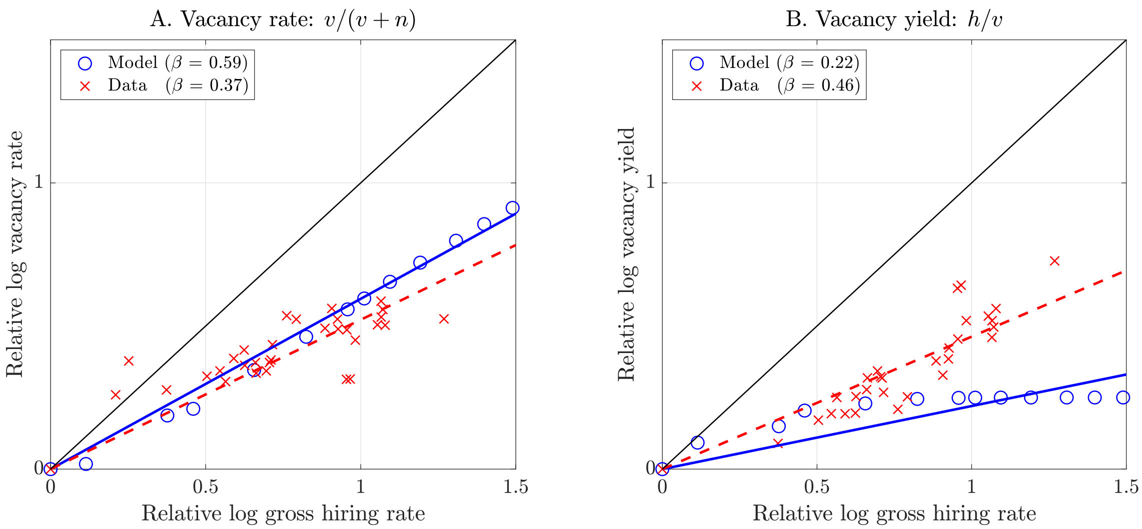


Figure 6: Vacancy rates by firm gross hiring rate

Notes **Data**: Establishment-month observations in JOLTS microdata 2002-2018 are pooled in bins, where bins are determined by net monthly growth rate, and have a width of 1 percent. Growth rates computed as in Davis, Faberman, and Haltiwanger (2013). Within bin b, total hires h_b, total vacancies v_b, total employment n_b are computed. From these, the gross hiring rate h_b/n_b, and implied daily vacancy posting rate $vr_b = v_b/(n_b)$ are computed using the daily recruiting model of Davis, Faberman, and Haltiwanger (2013). **Model**: The variables are constructed in the same way as in the data. Points plotted are logs of these variables, differenced about the bin representing a one percent net growth rate.

can poach labor from more firms. We compute these objects in JOLTS microdata in narrow monthly growth rate bins, then estimate (11) by non-linear least squares.[29] Our estimates imply $\gamma = 3.45$. Figure 6 shows this value for the vacancy cost elasticity provides a good fit to the microdata.

Internally estimated. The remaining parameters are estimated by minimizing the objective function

$$\mathcal{G}(\boldsymbol{\psi}) = \Big(\widehat{\boldsymbol{m}} - \boldsymbol{m}(\boldsymbol{\psi})\Big)' \mathbf{W}^{-1} \Big(\widehat{\boldsymbol{m}} - \boldsymbol{m}(\boldsymbol{\psi})\Big) \quad , \quad \boldsymbol{\psi} = \Big\{\mu, \sigma, \alpha, \zeta, A, \xi, \delta, b\Big\} ,$$

where $\widehat{\boldsymbol{m}}$ is a vector of empirical moments and $\boldsymbol{m}(\boldsymbol{\psi})$ are their model counterpart. The matrix $\mathbf{W}$ contains squares of the data moments on the main diagonal and zeros elsewhere.[30] We target eight moments that are relatively standard to firm dynamics and frictional labor market literatures. While this remaining subset of parameters is jointly estimated, some moments are particularly informative about some parameters. Next, we briefly outline our logic.

The drift of productivity, μ, is informed by the (unweighted) firm exit rate. The more negative the

[29] We use establishment-month observations in JOLTS microdata 2002–2018 (see Mongey and Violante, 2019). As in Davis, Faberman, and Haltiwanger (2013), we pool all vacancies, hires and employment in net monthly growth rate bins of one percent width. We then use these to compute the vacancy rate and vacancy yield at the bin level. To be consistent with our approximation we use growth rate bins between 2.5 and 9.5 percent when estimating (11).

[30] Our moments are taken from various data sources and in most instances we cannot compute variances of the moments, let alone covariances with other moments.

drift, the faster firms exit. The standard deviation of productivity shocks, σ, is informed by the standard deviation of annual log employment growth. If shocks are larger, employment is more volatile. Decreasing returns, α, is informed by the employment share of firms with more than 500 employees. A smaller span of control allows for fewer large firms. The thickness of the tail of the productivity distribution of entrants, ζ, is informed by job creation among young firms (Decker, Haltiwanger, Jarmin, and Miranda, 2020).[31] Matching efficiency, A, is informed by the nonemployment rate, as a more efficient labor market reduces nonemployment. Relative search efficiency of employed workers, ξ, and the exogenous separation rate, δ, are informed by quarterly EE and EN rates. Finally, the flow value of leisure, b, is informed by the job destruction rate of incumbent firms. The direct effect of b on marginal surplus $S_n(n,z)$ is one-for-one, so under a higher b, a productivity shocks is more likely to lead a firm to hit the layoff frontier $S_n(n,z) = 0$, and destroy jobs. Table 1C summarizes the parameters estimated by minimum distance. In Appendix E we discuss identification more formally, and plot the marginal effect of each parameter on its associated target moment and on the objective function.

5.2 Model fit

The estimated model is consistent with micro data that was not directly targeted by the estimation. This data sits at the intersection of firm and worker dynamics: (i) the distribution of firms and employment, (ii) job and worker flows across the distribution, and (iii) patterns of net poaching across the distribution.

1. Distribution of firms and employment. Figure 7 shows that the model reproduces the skewed firm size and age distributions. In both data and model, around 90 percent of firms are small (less than 20 employees), but these firms account for only around 20 percent of employment. Symmetrically, firms with more than 500 employees represent less than 1 percent of firms, but more than 50 percent of employment. By age, around half of firms are older than 10 years, but these account for 80 percent of employment in the data, and somewhat more in the model.

2. Firm, job and worker reallocation. Figure 8 examines turnover at the level of firms, jobs and workers. As in the data the rate of job creation peaks for young firms and then declines with age, while job destruction rates are relatively flat. The model is also consistent with the mild decline in job turnover by size. Not shown in the figure, in the data (model) 16 percent (15 percent) of all jobs are created by new

[31] A natural alternative would have been to target the productivity gap between entrants (younger than 1 year old) and incumbents. The model does well in this respect. At the estimated parameter vector, this gap is 27 (35) percent in the model (data) (Gavazza, Mongey, and Violante, 2018).

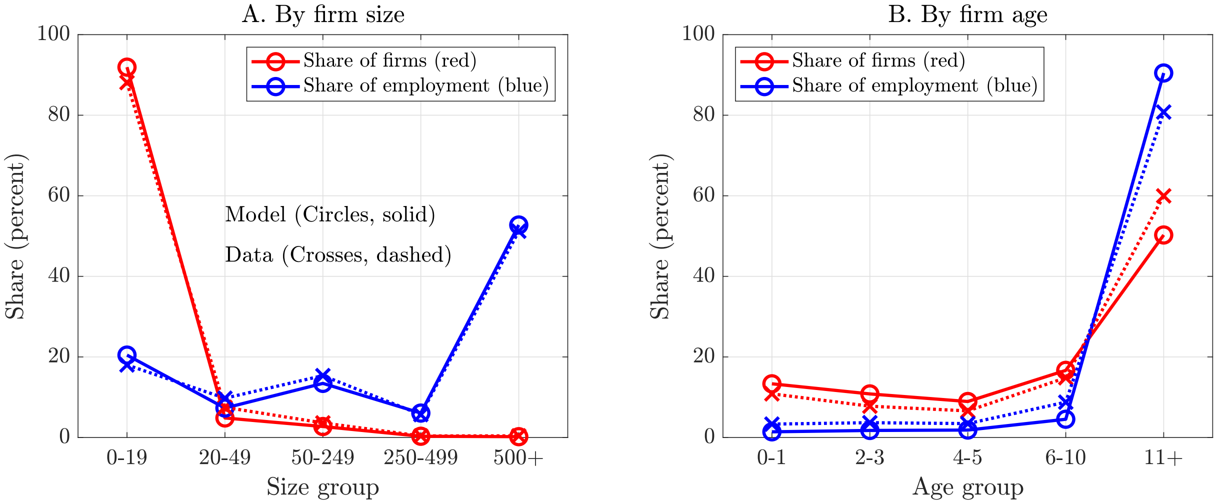


Figure 7: Distribution of firms and employment by firm age and size in data and model

Source: Census BDS

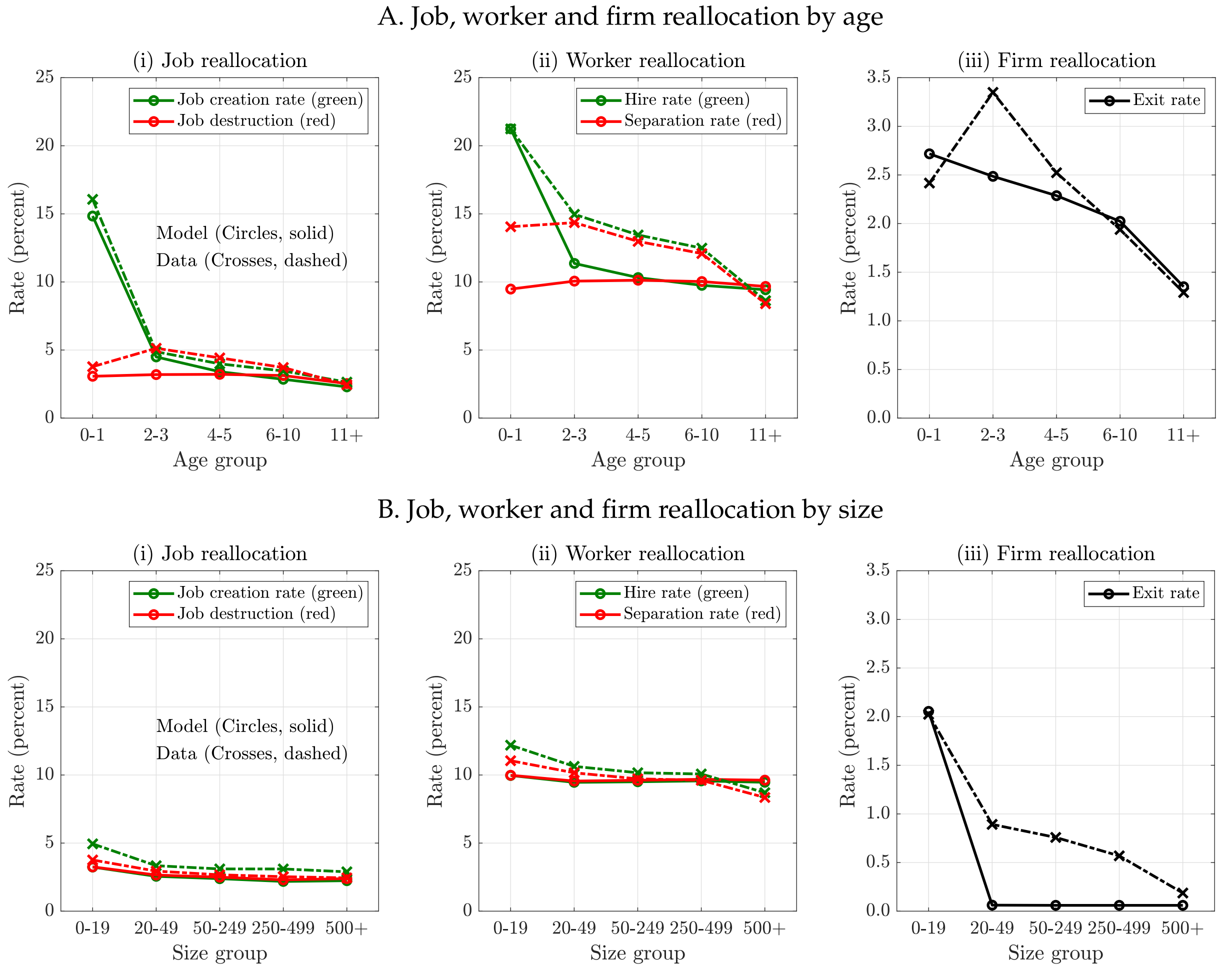
firm births and 28 percent (26 percent) by firms less than 10 years old. In terms of worker flows, EE transitions allow the model to account for the key fact that worker reallocation rates are around three times as large as job reallocation rates, thus generating the right amount of churning. In the cross-section, the model can reproduce the stark negative empirical relationship between hiring rates and firm age, but struggles to match the gradient at which the separation rate declines with age in the data after age 3. The model is also in line with the mild decline in worker turnover by size. The model reproduces the negative gradient for firm exit with respect to age observed in the data. Absent jumps in the productivity process, firms slowly shrink before exiting, and so every firm above medium size only exits exogenously: as a result, the model matches small and very large firm exit rates, but not those of intermediate size firms.[32]

3. Net poaching by firm characteristics. Figure 9A plots the distribution of marginal surplus $H(S_n)$ together with the net poaching rate as a function of marginal surplus. The CDF reveals that the equilibrium density $h(S_n)$ is quite dispersed. Net poaching is flat and negative at the low end of the distribution of S_n, after which it starts increasing steadily. What explains this particular shape? Under our assumptions on vacancy costs, the vacancy rate of the firm ($\widetilde{v} = v/n$) depends only on marginal surplus.[33] The net poaching rate $p(S_n)$ is therefore:

$$p(S_n) = \widetilde{v}(S_n)q(\theta)(1-\phi)H_n(S_n) - \lambda^E(\theta)\overline{H}_v(S_n).$$

[32] Note that the fact that the model slightly overestimates both (i) the hiring and separation rates at old firms (Figure 8A), and (ii) the employment share at old firms (Figure 7B), implies that the total share of hires and separations at old firms is larger than in the data (e.g. 15 ppt higher for age 11+ firms).

[33] To see this note that the marginal cost of a vacancy is $c_v(v,n) \propto (v/n)^\gamma$ and, as characterized in Section 4, the marginal benefit of a vacancy depends only on S_n.



A. Job, worker and firm reallocation by age

B. Job, worker and firm reallocation by size

Figure 8: Job, worker and firm reallocation by size and age

Notes: **Data**: Census BDS firm data for annual job creation, job destruction and exit. Quarterly rates constructed by dividing by four. Census J2J firm data for quarterly hiring and separation rates. Authors aggregate data into bins given in table which reflect the granularity of J2J data. Census J2J separations (hires) include separations to (hires from) non-employment. **Model**: Time aggregated to a quarterly frequency.

Higher marginal surplus increases net poaching through three channels: (i) a higher return to vacancies leads to higher vacancy posting, increasing EE hires ($\uparrow \widetilde{v}(S_n)$); (ii) conditional on a vacancy, a greater fraction of meetings result in a hire ($\uparrow H_n(S_n)$); (iii) firm incumbents match with fewer competitors that result in an EE quit ($\downarrow \overline{H}_v(S_n)$). Figure 9B plots these three forces using the following decomposition:

$$p(S_n) = \underbrace{\int_0^{S_n} \frac{\partial \widetilde{v}(u)}{\partial u} q(\theta)(1-\phi) H_n(u) du}_{\text{Increasing} \uparrow \widetilde{v}(S_n)} + \underbrace{\int_0^{S_n} \widetilde{v}(u) q(\theta)(1-\phi) h_n(u) du}_{\text{Increasing} \uparrow H_n(S_n)} - \underbrace{\lambda^E(\theta)\overline{H}_v(S_n)}_{\text{Decreasing} \downarrow \overline{H}_v(S_n)}$$

Firms with very low marginal surplus do not hire and lose all their employees who meet other firms, so net poaching for them approaches $-\lambda^E(\theta)$. In the middle range of $\log S_n$ a rise in marginal surplus

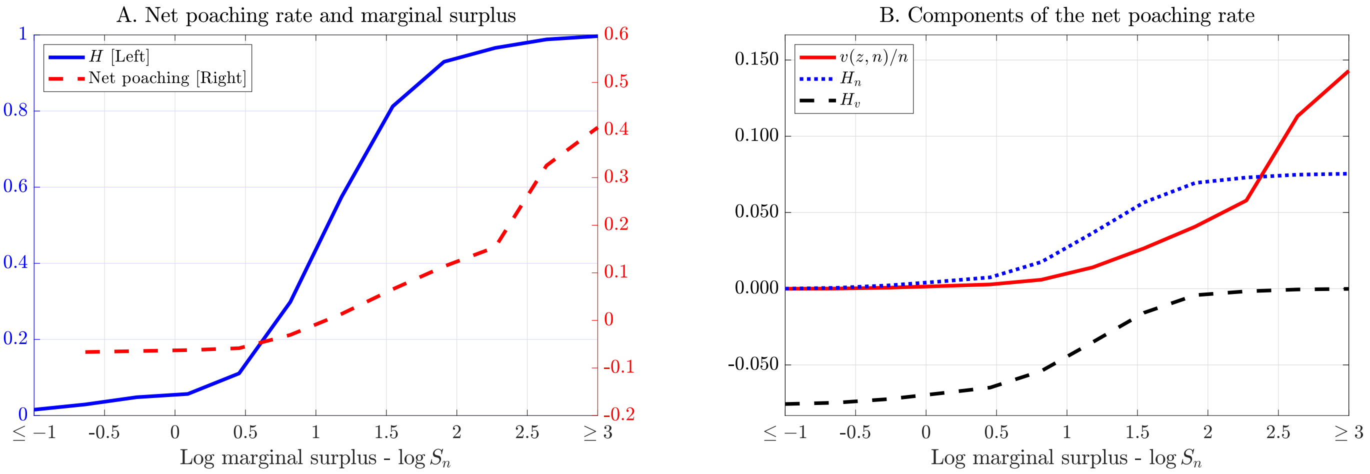


Figure 9: Net poaching and marginal surplus distribution

Notes: **Panel A.** Net poaching rate $p(S_n)$ by log marginal surplus S_n and the CDF of log marginal surplus. **Panel B.** Decomposition of the change in net poaching rate as S_n rises into three components: (i) higher vacancies (red line), (ii) more poaching hires due to higher rank on the job ladder (green line), and (iii) lower poaching separations due to higher rank on the job ladder (blue line).

increases firms' net poaching mostly through changes in its marginal surplus rank which, in turn, expands hires from other firms (EE^+) and reduces quits (EE^-). The vacancy rate initially rises slowly, but as firms get toward the top of the job ladder, vacancies are the only way to keep growing. This explains why net poaching keeps rising even in the region where the CDF is flattening out, over which poaching translates into negligible jumps up the ladder.

We now project this relationship between net poaching and marginal surplus onto observables in order to compare model and data. Haltiwanger, Hyatt, Kahn, and McEntarfer (2018) document two key empirical patterns: (i) a negligible gradient of net poaching by size, which is inconsistent with wage posting models, and (ii) a steeper negative gradient by age, as young poach from old. Panels A and B of Figure 10 show that the model matches these patterns quite well, albeit the slope by age is somewhat more pronounced in the model relative to the data. Size is not a particularly good predictor of where a firm sits on the marginal surplus job ladder. Consider a vertical slice of Figure 5. At a given size some firms are highly productive, have a high S_n, have positive net poaching and create jobs on net. Meanwhile, other firms are less productive, have a low S_n, negative net poaching and destroy jobs on net. In contrast, young firms are on average small and productive: they sit to the left of $dn = 0$ and, having not yet grown, are high on the marginal surplus job ladder. They therefore display large and positive net poaching rates.[34]

[34] Put differently, the reason why small firms do not have high net poaching rates on average is because some of them are young and highly productive, but did not have had time to grow yet, while others are small simply because they are

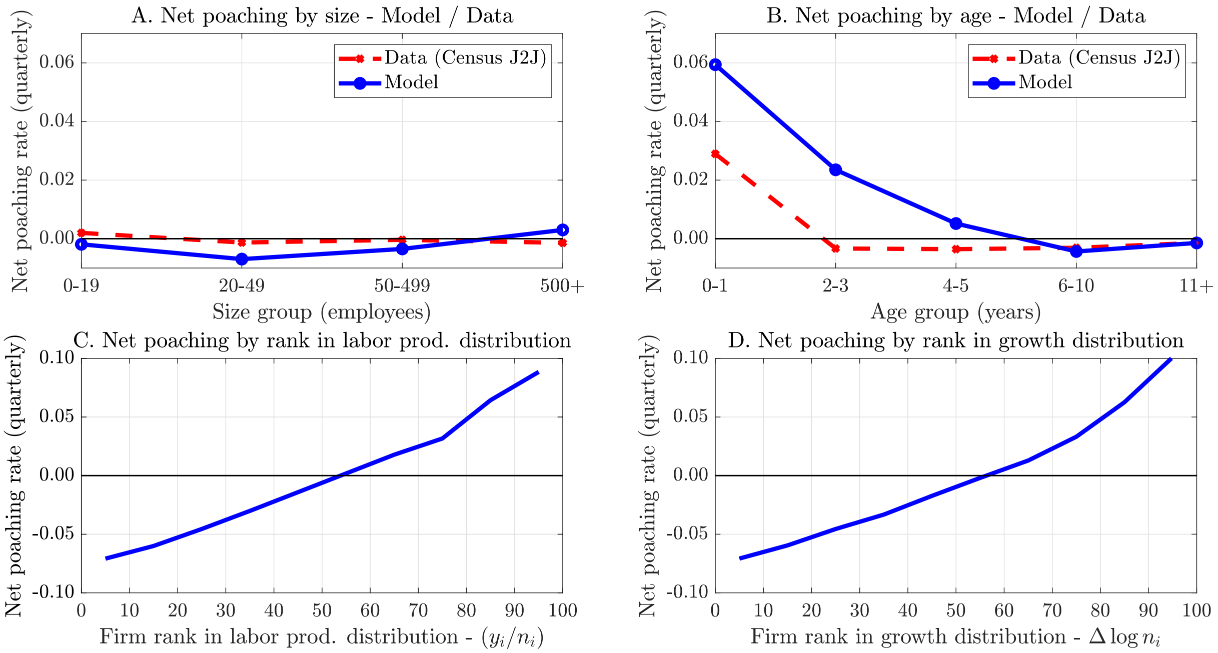


Figure 10: Net poaching rates by size, age, labor productivity and net employment growth rate

Notes: In panels C and D, firms are ranked according to the employment weighted distribution of labor productivity (C) and employment growth (D).

Panels C and D of Figure 10 plots net poaching rates as a function of two other, potentially observable firm characteristics, labor productivity and net employment growth rate. The model predicts a much higher gradient between these two variables and net poaching rates compared to size and age. First, marginal surplus is highly correlated with the static marginal product of labor, and the latter is proportional to the average product under our functional form for $y(z,n)$.[35] Second, both in the data and in the model, hires from employment account for much of firm employment growth. This implies a tight positive relation between net growth rate and net poaching rate.

6 Search frictions and labor misallocation

It is search and matching frictions, in our model, that impede the instantaneous reallocation of labor across firms. In this section we analyze and quantify the implications of this source of misallocation along three dimensions of the data: cross-section, firm life-cycle and the aggregate business cycle. Because our model is consistent with micro data on both the speed and the direction of the poaching flows across

unproductive.

[35] One generalization of the model that would weaken this relation is the addition of heterogeneity in the scale of production parameter α, as in Gavazza, Mongey, and Violante (2018). This would create an additional source of cross-sectional variation in marginal surplus that is orthogonal to z. Another one is the addition to the model of forced EE moves ('godfather shocks').

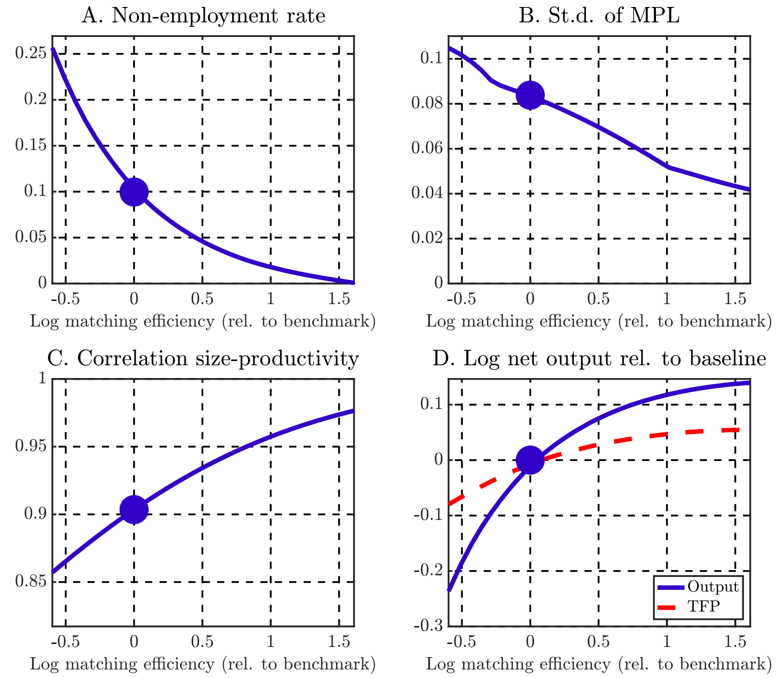


Figure 11: Frictionless limit: the effect of increasing match efficiency

firms that resolve such *frictional misallocation*, it offers a credible platform for these three exercises.

6.1 Misallocation cost of labor market frictions

Our model puts us in a unique position to quantify such misallocation by computing the limit as search frictions vanish. This exercise would not make sense in a model without decreasing returns since it would predict that the most productive firm would hire the entire labor force. Conversely, we have shown that our economy converges to a competitive equilibrium with a non-degenerate firm size distribution.

Our counterfactual experiment shifts the value of matching efficiency A holding all other parameters fixed at our baseline calibration. Figure 11 plots model outcomes for a wide range of values for A. As frictions vanish, non-employment falls, the dispersion of marginal products across firms shrinks toward zero, the correlation between size and productivity rises toward one, and aggregate TFP and output grow.[36]

To isolate the role of frictional misallocation, we decompose the change in output into a component due to the allocation of workers across firms, and a component due to higher employment in the economy as a whole, the scale effect. Imposing an aggregate production function $Y = Z\mathrm{n}^{\alpha}$, then across steady

[36] Output is net of vacancy costs. Note that the relationship between frictions and output is concave because the non-employment rate is convex in match efficiency.

states

$$\Delta \log Y = \Delta \log Z + \alpha \Delta \log \mathtt{n} \quad , \quad Z := \int_{\mathcal{N}\times\mathcal{Z}} z \left(\frac{n}{\mathtt{n}}\right)^{\alpha} dH(n,z).$$

The *TFP* term Z captures misallocation and is constant if the distribution of employment across firms is unchanged.[37] The misallocation from labor market frictions is sizable in our model. For example, reducing frictions to an extent that cuts the non-employment rate by half (from 10 to 5 percent) raises aggregate TFP by 3 percent and output by 7 percent. Thus reallocation accounts for nearly half the gains in output. To put this finding in context, the NE rate increases from 0.49 to 0.94 and the EE-rate increases from 0.04 to 0.07 (all at quarterly frequency). A five-fold rise in match efficiency relative to the benchmark would virtually eliminate frictional non-employment and boost TFP permanently by 5 percent.

6.2 Frictional misallocation and the life cycle of superstar firms

The addition of labor market frictions to a firm dynamics model overcomes a notable shortcoming of competitive environments first identified by Luttmer (2011). When these models are calibrated to generate the correct cross-sectional variation in firm employment growth rates and the empirical size distribution of firms, they imply that the median age of 'superstar' firms ($n > 10,000$ workers) is 750 years. In the data, the median age of such firms is only about 75 years.[38]

The root of the problem is that in the data young firms are almost uniformly small and firm employment volatility is not that large. Viewed through the lens of a frictionless model where size and productivity are perfectly correlated, young firms must be low productivity. This moment is therefore matched by young firms being way out in the left tail of the productivity distribution. If shocks are driven by a geometric Brownian motion and chosen to match the empirical volatility of firm growth, then only small shocks are required. Initially low productivity and small shocks means that it takes a very long time for any firm to get from the left to the very right tail of the productivity distribution. Since productivity is the only determinant of firm size, then it also takes a very long time for a firm to become a superstar firm.

In our environment, instead, labor market frictions imply that some young firms can be extremely productive but still small as frictions have prevented them from immediately growing large. Because their initial productivity is already near the upper tail, they remain at the top of the job ladder even

[37] In a constant returns to scale economy, the frictionless limit would be $Z = \max z$, since all employment would be at the firm with the highest productivity. Gains from eliminating search frictions would be implausibly large.

[38] For description of data see Luttmer (2011), Appendix A. The data in Luttmer (2011) was collected for 2008, on 813 firms with 10,000 or more workers. To compute firm age, Luttmer (2011) collected data on incorporation taken from a variety of historical sources, for example *Mergent Online*.

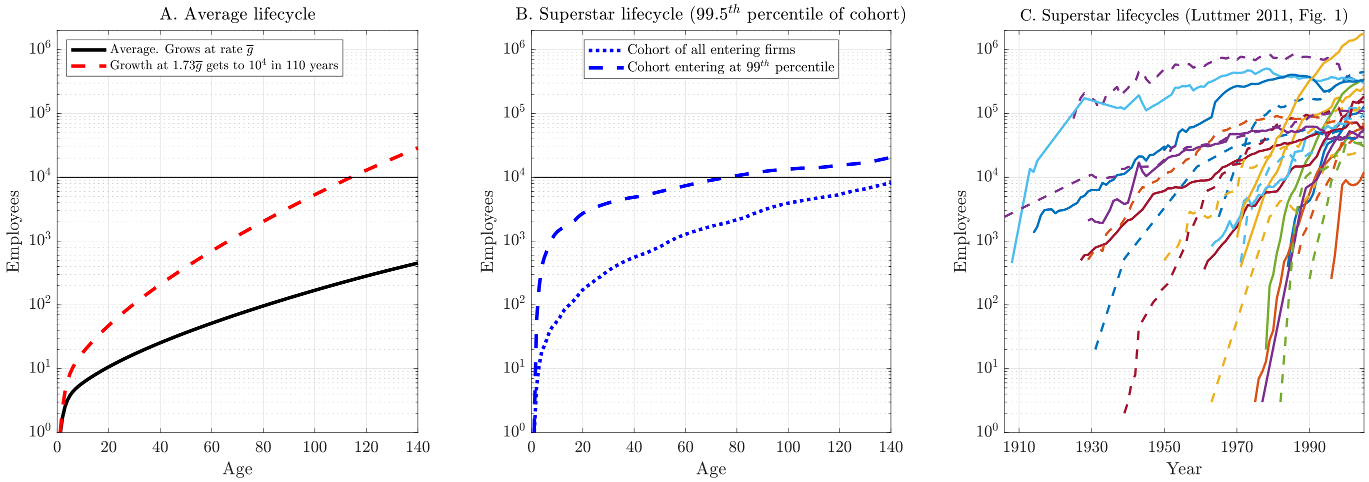


Figure 12: Comparing the average firm lifecycle to the lifecycle of superstars

Notes: **Panel A.** plots the average firm size by age, $\overline{n}_a$, (black) which grows at rate $g_a = \overline{n}_a/\overline{n}_{a-1} - 1$. The red dashed line plots a counterfactual under which n_0 grows at rates $\lambda \times g_a$, where $\lambda = 2.05$. This delivers a firm size of 10,000 at 110 years, which is the median age of a 10,000 employee firm in the model. In **Panel B.** we take all entering firms (blue dotted), and plot the 99.8th percentile of the firm size distribution at each age. The blue dashed line takes a mass of firms that enter at the 99th percentile of the entrant productivity distribution, and plots the 99.8th percentile of the firm size distribution at each age for this mass of firms. **Panel C.** uses the data made publicly available by Luttmer (2011) to re-create Figure 1 of that paper. It plots the employment histories of 25 of the 1,000 firms that had more than 10,000 employees in 2008.

after hiring many workers. This makes expansion cheap, and accommodates rapid growth. They can therefore move relatively quickly to the tail of the size distribution. When we simulate firm life cycles in the model, we find that the median age of firms with more than 10,000 workers is 110 years, and hence much closer to the data. It is thus the existence of frictional labor misallocation—high productivity, but small size firms—that allows the model to be consistent with the cross-sectional size distribution, volatility of firm growth rates, and life-cycle growth trajectories.[39]

Figure 12 sheds further light on the mechanism behind this finding. Panel A shows than in the model, on average, firms achieve much smaller sizes conditional on surviving for 110 years (black line). One would need a growth rate permanently twice as large in order for the expected firm size at age 110 to exceed 10,000 employees (red line). Instead, from the perspective of the model, being a superstar firm after a century of activity is an extreme tail event both in terms of initial productivity at entry and in terms of sequence of realized productivity shocks along the life cycle (Panel B). Finally, note that the employment trajectory of firms destined to stardom is highly concave both in the data (panel C) and in the model (panel B), and especially so for those firms that already start in the tail (dashed line in panel B).

[39] Recall that we target the dispersion of firm employment growth in our estimation exercise (Table 1). The model reproduces quite well the distribution of firms and employment by firm age (Figure 7), and average firm size (Table 1). Together these imply that we match average size by age and hence average lifecycle growth of firms.

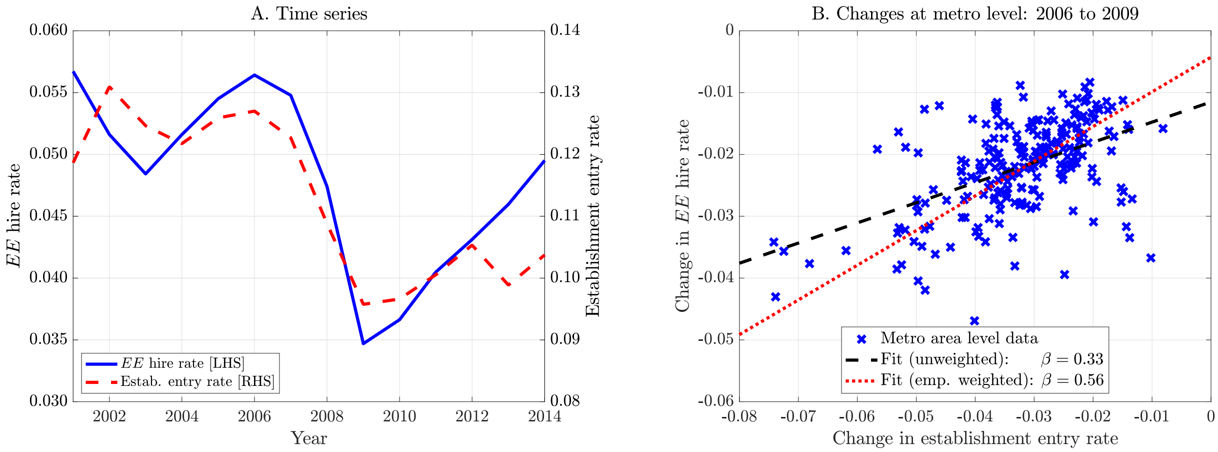


Figure 13: Entry and job-to-job hiring rates over the Great Recession: Aggregate and Cross-section

Notes: Both panels are constructed from the same data at the metro level. Establishment entry and number of establishments are from the Census BDS data, and used to construct *establishment entry rate*. Job-to-job hires and employment are from the Census J2J data, and used to construct *EE hire rate*. The data cover the subset of states that participate in these Census data release programs. These cover more than half of the US population.

6.3 Frictional misallocation in the Great Recession

The Great Recession provided a raft of new facts regarding the cyclical reallocation of workers across firms. In particular the two key mechanisms through which labor gets reallocated across productive units are (i) the entry of new firms which replace unproductive exiting firms and (ii) the upward movement of workers up the job ladder toward more productive firms. During the Great Recession both mechanisms slowed considerably. Firm entry (measured as the number of firms less than 1 year old in the BDS) dropped by almost 30 percent between 2007 and 2009 and has since recovered very slowly. Even allowing for the secular decline in firm entry documented by Pugsley and Sahin (2019), the drop around the Great Recession would be at least 20 percent.[40] The EE rate also fell markedly over the same period (Figure 13A).[41]

The decline in job-to-job transitions implied a tapering in the process of upgrading from low- to high-rank firms. Haltiwanger, Hyatt, Kahn, and McEntarfer (2018) document a fall in net poaching of high wage firms, those who are presumably at the top of the job ladder. Similarly, Moscarini and Postel-

[40] The entry rate in Figure 13A is for establishments for consistency with panel B which is constructed at the city level for which there are no firm-level data. The percentage fall in entry rates of firms and establishments in 2008-2009 is very similar and over the last 40 years the correlation between the two series is 0.98.

[41] The exact size of this decline is still debated. Haltiwanger, Hyatt, Kahn, and McEntarfer (2018) use Census $J2J$ data and report a decline around 30 percent, which Figure 13 replicates. Fujita, Moscarini, and Postel-Vinay (2019) argue that Census data overestimate this drop because employment status is only measured quarterly and spurious poaching transitions (EE from quarter to quarter, but with an 'invisible' non-employment spell in between) were much less likely during the recession, when unemployment spells were quite long. When they reassess measurement error in CPS data, these authors estimate a drop in the EE rate around 15 percent.

Vinay (2016) use a structural model to rank firms on the job ladder and estimate that high-rank firms disproportionately curtailed their demand for new labor in the recession. In short, as they put it: *the job ladder failed, starting from the upper rungs.*

To date, these two facts have not been connected in the literature, since there is a paucity of quantitative structural models that can integrate firm dynamics with on-the-job search. Theoretically, our environment suggests a natural link between the two, and since our model matches key data on firm and worker dynamics we can also test this link quantitatively.

Theoretically, the mechanism is as follows. New entrants and young firms account for a sizable share of vacancies and have higher marginal surplus than other firms in the economy. Following a shock that leads to a drop in the number of entrants, poaching would fall at the top of the ladder which reduces worker reallocation through the middle of the ladder, and so on down to unemployment.

Empirically, this idea is consistent with the correlation observed across cities. Figure 13B combines newly released Census *J2J* data with Census *BDS* data at the metro level. The time-series decline in entry and job-to-job mobility is mirrored in the cross-section of labor markets: cities with larger declines in establishment entry also display larger declines in job-to-job mobility.

We now simulate the Great Recession in our model. The aggregate shock that best describes the Great Recession is one that worsens financial conditions. To proxy for a financial shock in our framework, we solve the model under an unexpected temporary increase in the discount rate ρ (as in Hall, 2017). We calibrate the initial jump and the rate of convergence of ρ to match the 5 ppt increase in the unemployment rate and the seven years it took to return to pre-recession levels.[42]

Because the focus is on short-run dynamics, we replace the long-run free-entry condition with a simple imperfectly elastic entry rule. We posit that every instant, a measure 1 (a normalization) of entrepreneurs contemplate entering. They draw preference shocks for opening a new firm (ϵ_1) and for taking an outside option (ϵ_2) normalized to deliver a payoff of 1. Thus, entrepreneurs solve

$$\max \left\{ \epsilon_1 \times \frac{\overline{S}_t}{c_0} \, , \, \epsilon_2 \times 1 \right\}$$

where $\overline{S}_t = \int S_t(z, n_0) d\Pi_0(z)$ denotes the expected surplus at firm entry. We assume that ϵ_i are Frechet

[42] We calibrate the discount rate shock as follows. We feed in an AR(1) path for ρ_t with a half-life of two years, starting above its steady-state value. To generate a 5 ppt increase in unemployment, the model requires a twenty-fold initial decline in the discount factor. While this value may seem large at first sight, it actually represents a plausible reduction in the stochastic discount factor of firms' shareholders. Indeed, in a 'large family' model with curvature in the utility function and consumption equal to output every period, the initial 6% monthly decline in output that follows from our shock translates into a twenty-fold decline in the stochastic discount factor provided risk aversion is around 10. Values in this range are not uncommon in macro-finance, especially in in the long-run risk (Bansal and Yaron, 2004) and the countercyclical idiosyncratic risk (Constantinides and Ghosh, 2017) literatures.

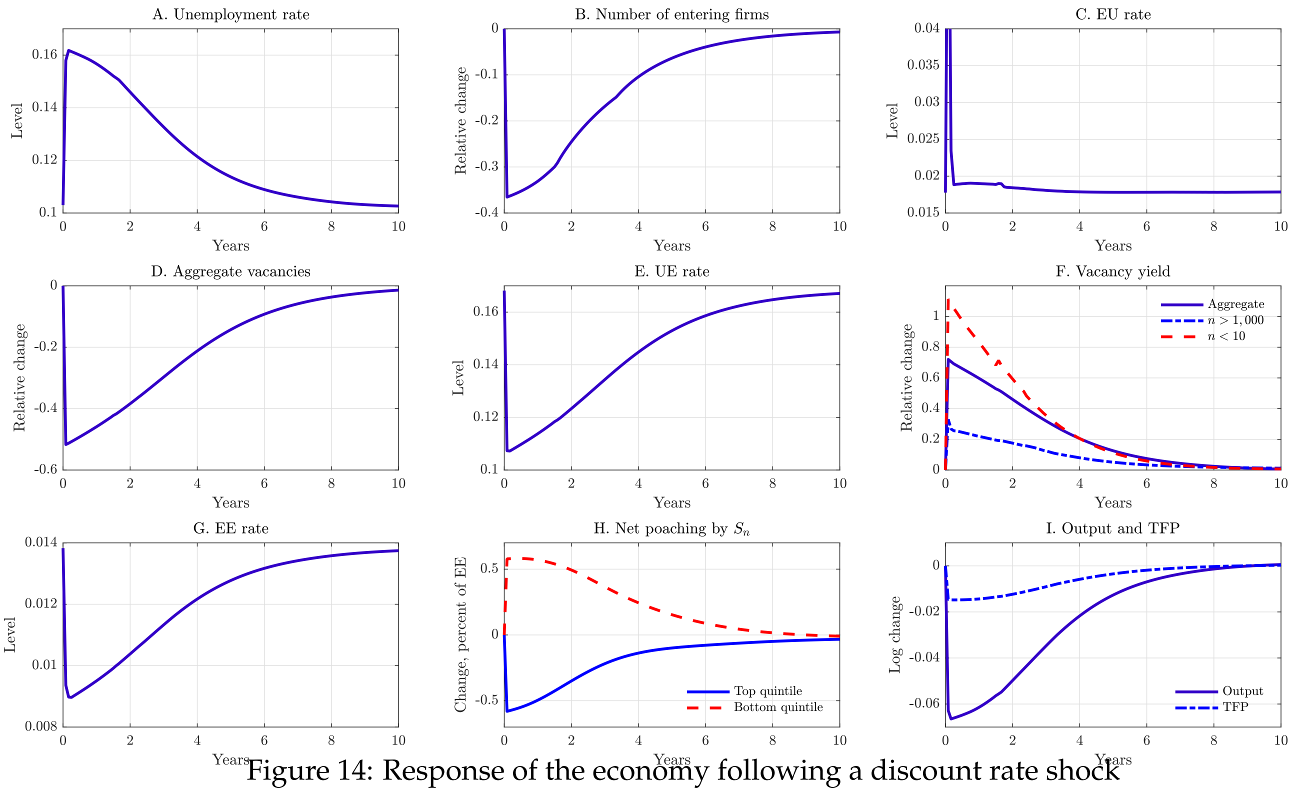


Figure 14: Response of the economy following a discount rate shock

distributed with shape parameter χ and scale normalized to 1. The measure of new firms at t satisfies:

$$M_t = \frac{\bar{S}_t^{\chi}}{\bar{S}_t^{\chi} + c_0^{\chi}}$$

We set $\chi = 0.05$ to match the observed decline in firm entry (-30 percent) following the discount rate shock.

Figure 14 describes the response of the model economy to the shock. The shock lowers the valuation of future revenues at all firms, and as a result, average surplus falls. Young firms have a disproportionate fraction of their revenues in the future so are especially hard hit, causing entry to collapse (Panel B). Marginal surplus also falls, which causes an endogenous spike in *EU* separations (Panel C). Both lead to a jump in unemployment (Panel A). Furthermore, declining marginal surplus reduces the return on vacancies (7), so aggregate vacancies collapse, job creation contracts, and *UE* hires decline (Panels D and E). This leads the jump in unemployment to persist. With less vacancies and more unemployed workers, the aggregate vacancy yield rises (Panel F). As in the data, the rise in the vacancy yield is more pronounced for small firms (Moscarini and Postel-Vinay, 2016). For small unproductive firms at the bottom of the ladder, unemployment is the main source of hiring, so as the pool of unemployed job seekers expands, the vacancy filling rate of these firms jumps (Figure 14F).

Quantitatively, the experiment matches key non-targeted moments: the *EE* rate falls about one third

(Panel G) and vacancies contract by 50 percent (Gavazza, Mongey, and Violante, 2018). The decline in output is 6.5 percent and 9 percent in the data (Fernald, 2014).

We now turn to the dynamics of the job ladder. In the aggregate, the job-to-job mobility rate drops upon impact and slowly recovers (Panel 14G). In the cross-section, the shift in the vacancy distribution away from high marginal surplus firms—whose poaching rates were most sensitive to their vacancy rates (recall Figure 9B)—causes poaching rates to collapse at high marginal surplus firms and, symmetrically, grow at low-marginal surplus firms (Panel H). This compositional effect reduces the probability that a worker moves from a low- to a high-marginal surplus firm, causing the observed 'failure' of the job ladder.

Throughout the recession and its protracted recovery, the slowdown of worker flows towards high-marginal surplus firms exacerbates the misallocation that arises from labor market frictions. This force grinds down aggregate TFP, and is responsible for about a quarter of the decline in output (Panel I). The recovery of aggregate productivity is sluggish, with the scars of the recession encoded in the slow moving dynamics of the distribution of employment across firms.

7 Conclusion

We have set out a new framework to jointly study firm entry and exit, job reallocation, and worker turnover, both through non-employment and through direct job-to-job transitions, in a frictional labor market. The novel feature of the environment, which makes the problem challenging in the presence of random on-the-job search, is diminishing returns to scale in the firm's technology –the hallmark of classic theories of the firm size distribution based on the idea of 'span of control'. By extending the contractual environment of Postel-Vinay and Robin (2002), we obtain a tractable 'joint value' representation that reduces a potentially unmanageable state space to a very parsimonious one. In contrast with search existing models with linear technology that display a job ladder in exogenous productivity, our model features an endogenous job ladder in *marginal surplus*. Canonical search models and competitive firm dynamics models are special cases of our environment.

We illustrated how to use a calibrated version of the model as a laboratory to shed light on the role of labor misallocation due to search and matching frictions in the cross-section, firm life cycle, and aggregate time series dimension.

Our framework is quite flexible and can be extended in a number of directions while retaining tractability (i.e., a parsimonious state space). For example, an isomorphic representation of our firm problem is *constant returns to scale* in production of a differentiated final good, which would yield a *decreasing marginal revenue* as under monopolistic competition. We can therefore easily accommodate imperfect substitutability in the goods market which is a key ingredient of trade models and macroeconomic models with nominal rigidities in goods markets.

The model can also integrate heterogeneity in the scale of production across firms (e.g. to allow for fast-growing 'gazelle' firms) and fixed heterogeneity in worker types to address sorting within and across firms. It is also straightforward to introduce firm-level amenities which have been documented to be important to describe sorting patterns in the data (Sorkin, 2018). Once wage determination is incorporated into the model —a task we left to future work— shocks to firm-specific or general human capital can also be accommodated to study earnings dynamics within firms and along worker careers.

In all these cases the joint-value representation remains valid and, as long as heterogeneity is discrete and of moderate dimension, the cardinality of the state vector remains manageable. As is, the model can be used to study the effect of most labor market policies (e.g., unemployment benefits, severance payments, firing taxes and hiring subsidies) on firm dynamics, worker reallocation and aggregate employment and output.

Stationary equilibrium and transitional dynamics can be computed very efficiently and, albeit not in the paper, incorporating aggregate shocks using the approach of Boppart, Krusell, and Mitman (2018) is straightforward which makes business cycles analysis possible.

In ongoing work (Bilal, Engbom, Mongey, and Violante, 2021), we show how one can combine this framework with a creative-destruction model of endogenous growth and revisit the nexus between the speed of technical change and aggregate employment (Aghion and Howitt, 1994) in the context of the recent growth slowdown and decline in labor market dynamism. Differently from the canonical models in the endogenous growth literature where optimal size is reached instantaneously, in this hybrid model building to the optimal productive capacity and replacing less productive incumbents is a slow process for an innovator, which requires poaching workers away from other firms. As a result, creative destruction can induce more labor misallocation.

In sum, with the introduction of a well defined notion of firm boundaries (through decreasing returns in technology or downward sloping demand) into a comprehensive model of frictional labor reallocation across firms, our framework can be potentially useful to study a number of questions in growth, business cycle analysis, labor and trade.

References

ACEMOGLU, D. (2001): "Good Jobs versus Bad Jobs," *Journal of Labor Economics*, 19(1), 1–21.

ACEMOGLU, D., AND W. B. HAWKINS (2014): "Search with Multi-Worker Firms," *Theoretical Economics*, 9(3), 583–628.

AGHION, P., AND P. HOWITT (1994): "Growth and Unemployment," *The Review of Economic Studies*, 61(3), 477–494.

ARELLANO, C., Y. BAI, AND P. KEHOE (2019): "Financial Frictions and Fluctuations in Volatility," *Journal of Political Economy*, 127(22990), 2049–2103.

AUDOLY, R. (2019): "Firm Dynamics and Random Search over the Business Cycle," Discussion paper, University College London.

BANSAL, R., AND A. YARON (2004): "Risks for the Long Run: A Potential Resolution of Asset Pricing Puzzles," *The Journal of Finance*, 59(4), 1481–1509.

BERTOLA, G., AND R. J. CABALLERO (1994): "Cross-Sectional Efficiency and Labour Hoarding in a Matching Model of Unemployment," *Review of Economic Studies*, 61(3), 435–56.

BILAL, A., N. ENGBOM, S. MONGEY, AND G. L. VIOLANTE (2019): "Firm and Worker Dynamics in a Frictional Labor Market," NBER Working Paper 26547, National Bureau of Economic Research.

——— (2021): "Labor Market Dynamics When Ideas are Getting Harder to Find," NBER Working Paper 29479, National Bureau of Economic Research.

BILAL, A., AND H. LHUILLIER (2021): "Outsourcing, Inequality and Aggregate Output," Discussion paper, Harvard University.

BOPPART, T., P. KRUSELL, AND K. MITMAN (2018): "Exploiting MIT Shocks in Heterogeneous-agent Economies: the Impulse Response as a Numerical Derivative," *Journal of Economic Dynamics and Control*, 89, 68–92.

BOROVICKOVÁ, K. (2016): "Job flows, Worker Flows and Labor Market Policies," Discussion paper, New York University.

BREKKE, AND OKSENDAL (1990): "The High Contact Principle as a Sufficiency Condition for Optimal Stopping," *Preprint series: Pure mathematics*.

BRÜGEMANN, B., P. GAUTIER, AND G. MENZIO (2018): "Intra Firm Bargaining and Shapley Values," *The Review of Economic Studies*, 86(2), 564–592.

BURDETT, K., AND D. MORTENSEN (1998): "Wage Differentials, Employer Size, and Unemployment," *International Economic Review*, 39(2), 257–273.

CLEMENTI, G., AND D. PALAZZO (2010): "Entry, Exit, Firm Dynamics, and Aggregate Fluctuations," Discussion paper, New York University.

COLES, M., AND D. MORTENSEN (2016): "Equilibrium Labor Turnover, Firm Growth, and Unemployment," *Econometrica*, 84, 347–363.

CONSTANTINIDES, G. M., AND A. GHOSH (2017): "Asset Pricing with Countercyclical Household Consumption Risk," *The Journal of Finance*, 72(1), 415–460.

DAVIS, S. J., R. J. FABERMAN, AND J. C. HALTIWANGER (2013): "The Establishment-Level Behavior of Vacancies and Hiring," *Quarterly Journal of Economics*, 128(2), 581–622.

DECKER, R. A., J. C. HALTIWANGER, R. S. JARMIN, AND J. MIRANDA (2020): "Changing Business Dynamism and Productivity: Shocks vs Responsiveness," *American Economic Review*, 110(24236), 3952–3990.

DIAMOND, P. A., AND E. MASKIN (1979): "An Equilibrium Analysis of Search and Breach of Contract, I: Steady States," *Bell Journal of Economics*, 10(1), 282–316.

ELSBY, M., AND A. GOTTFRIES (2021): "Firm Dynamics, On-the-Job Search and Labor Market Fluctuations," *Review of Economic Studies (forthcoming)*.

ELSBY, M. W. L., AND R. MICHAELS (2013): "Marginal Jobs, Heterogeneous Firms, and Unemployment Flows," *American Economic Journal: Macroeconomics*, 5(1), 1–48.

ELSBY, M. W. L., R. MICHAELS, AND D. RATNER (2019): "The Aggregate Effects of Labor Market Frictions," *Quantitative Economics*, 10(3), 803–852.

ENGBOM, N. (2017): "Firm and Worker Dynamics in an Aging Labor Market," Discussion paper, New York University.

FERNALD, J. G. (2014): "Productivity and Potential Output before, during, and after the Great Recession," in *NBER Macroeconomics Annual*, ed. by J. A. Parker, and M. Woodford, vol. 29, pp. 1–51. Chicago, IL: University of Chicago Press.

FUJITA, S., G. MOSCARINI, AND F. POSTEL-VINAY (2019): "Measuring Employer-to-Employer Reallocation," Discussion paper, Yale University.

FUJITA, S., AND M. NAKAJIMA (2016): "Worker Flows and Job Flows: A Quantitative Investigation," *Review of Economic Dynamics*, 22, 1–20.

GAVAZZA, A., S. MONGEY, AND G. L. VIOLANTE (2018): "Aggregate Recruiting Intensity," *American Economic Review*, 108(8), 2088–2127.

GOUIN-BONENFANT, E. (2018): "Productivity Dispersion, Between-Firm Competition, and the Labor Share," Discussion paper, University of California San Diego.

HALL, R. E. (2017): "High Discounts and High Unemployment," *American Economic Review*, 107(2), 305–30.

HALTIWANGER, JOHN ANG HYATT, H. R., E. MCENTARFER, AND M. STAIGER (2021): "Cyclical Worker Flows: Cleansing vs. Sullying," NBER Working Paper 28802, National Bureau of Economic Research.

HALTIWANGER, J. C., H. R. HYATT, L. B. KAHN, AND E. MCENTARFER (2018): "Cyclical Job Ladders by Firm Size and Firm Wage," *American Economic Journal: Macroeconomics*, 10(2), 52–85.

HAWKINS, W. B. (2015): "Bargaining with Commitment Between Workers and Large Firms," *Review of Economic Dynamics*, 18(2), 350–364.

HOPENHAYN, H., AND R. ROGERSON (1993): "Job Turnover and Policy Evaluation: A General Equilibrium Analysis," *Journal of Political Economy*, 101(5), 915–938.

HOPENHAYN, H. A. (1992): "Entry, Exit, and Firm Dynamics in Long Run Equilibrium," *Econometrica*, 60(5), 1127–50.

JAROSCH, G. (2021): "Searching for Job Security and the Consequences of Job Loss," Discussion paper, National Bureau of Economic Research.

KAAS, L. (2020): "Block-Recursive Equilibria in Heterogenous-Agent Models," Discussion Paper 8737, CESifo Working Paper.

KAAS, L., AND P. KIRCHER (2015): “Efficient Firm Dynamics in a Frictional Labor Market,” *American Economic Review*, 105(10), 3030–60.

KIYOTAKI, N., AND R. LAGOS (2007): “A Model of Job and Worker Flows,” *Journal of Political Economy*, 115(5), 770–819.

KLETTE, T., AND S. KORTUM (2004): “Innovating Firms and Aggregate Innovation,” *Journal of Political Economy*, vol. 112, no. 5](5), 986–1018.

LENTZ, R. (2015): “Optimal Employment Contracts with Hidden Search,” Discussion paper, University of Wisconsin-Madison.

LENTZ, R., AND D. MORTENSEN (2012): “Labor Market Friction, Firm Heterogeneity, and Aggregate Employment and Productivity,” Discussion paper, University of Wisconsin.

LINDENLAUB, I., AND F. POSTEL-VINAY (2016): “Multidimensional Sorting Under Random Search,” Discussion paper, Yale University.

LISE, J., AND J.-M. ROBIN (2017): “The Macrodynamics of Sorting between Workers and Firms,” *American Economic Review*, 107(4), 1104–1135.

LUCAS, R. E. (1978): “On the Size Distribution of Business Firms,” *The Bell Journal of Economics*, 9(2), 508.

LUTTMER, E. G. (2010): “Models of Growth and Firm Heterogeneity,” *Annual Review of Economics*, 2(1), 547–576.

LUTTMER, E. G. J. (2011): “On the Mechanics of Firm Growth,” *Review of Economic Studies*, 78(3), 1042–1068.

MACLEOD, W. B., AND J. M. MALCOMSON (1989): “Implicit Contracts, Incentive Compatibility, and Involuntary Unemployment,” *Econometrica: Journal of the Econometric Society*, pp. 447–480.

MENZIO, G., AND S. SHI (2011): “Efficient Search On the Job and the Business Cycle,” *Journal of Political Economy*, 119(3), 468–510.

MONGEY, S., AND G. L. VIOLANTE (2019): “Macro Recruiting Intensity from Micro Data,” NBER Working Papers 26231, National Bureau of Economic Research, Inc.

MOSCARINI, G., AND F. POSTEL-VINAY (2013): “Stochastic Search Equilibrium,” *Review of Economic Studies*, 80(4), 1545–1581.

——— (2016): “Did the Job Ladder Fail after the Great Recession?,” *Journal of Labor Economics*, 34(S1), S55–S93.

PESKIR, AND SHIRYAEV (2006): *Optimal Stopping and Free Boundary Problems*. Basel: Birkhauser.

PETRONGOLO, B., AND C. A. PISSARIDES (2001): “Looking into the black box: A survey of the matching function,” *Journal of Economic literature*, pp. 390–431.

PHAM (2009): *Continuous-Time Stochastic Control and Optimization with Financial Applications*. Berlin, Heidelberg: Springer.

POSTEL-VINAY, F., AND J.-M. ROBIN (2002): “Equilibrium Wage Dispersion with Worker and Employer Heterogeneity,” *Econometrica*, 70(6), 2295–2350.

POSTEL-VINAY, F., AND H. TURON (2010): “On-the-job Search, Productivity Shocks, and the Individual Earnings Process,” *International Economic Review*, 51(3), 599–629.

PUGSLEY, B. W., AND A. SAHIN (2019): “Grown-up Business Cycles,” *The Review of Financial Studies*, 32(3), 1102–1147.

SCHAAL, E. (2017): “Uncertainty and Unemployment,” *Econometrica*, 85(6), 1675–721.

SEDLÁČEK, P. (2020): “Lost Generations of Firms and Aggregate Labor Market Dynamics,” *Journal of Monetary Economics*, 111, 16–31.

SHIMER, R. (2005): “The Cyclical Behavior of Equilibrium Unemployment and Vacancies,” *American Economic Review*, 95(1), 25–49.

SIEMER, M. (2014): “Firm Entry and Employment Dynamics in the Great Recession,” Finance and Economics Discussion Series 2014-56, Board of Governors of the Federal Reserve System.

SORKIN, I. (2018): “Ranking Firms Using Revealed Preference,” *The Quarterly Journal of Economics*, 133(3), 1331–1393.

STOKEY (2009): *The Economics of Inaction : Stochastic Control Models with Fixed Costs*. Princeton, NJ: Princeton University Press.

STOLE, L. A., AND J. ZWIEBEL (1996): “Intrafirm Bargaining under Non-Binding Contracts,” *Review of Economic Studies*, 63(3), 375–410.

Appendix

Firm and Worker Dynamics in a Frictional Labor Market

Adrien Bilal, Niklas Engbom, Simon Mongey, Gianluca Violante

This Appendix is organized as follows. Section A provides intuition for how our assumptions **(A)** yield a tractable Bellman equation for joint value. Section B provides a characterization of the surplus function. Section C derives an alternative limit of the Bellman equation for the joint value as decreasing returns vanish. Section D derives the limiting behavior of our economy when frictions vanish. Section E illustrates identification of the model.

A Static Example

Set up. Consider a firm with decreasing returns to scale technology $y(z,n)$ such that $y(z,0)=0$. Suppose the firm starts with productivity z and $n=1$ worker. The current contract between the firm and the incumbent specifies a wage $w_1 \in (b, y(z,1))$, where $b=U$ is the value of unemployment. At this point, the incumbent worker does not have a credible threat to quit into unemployment nor the firm has a credible threat to fire the worker. Then, the labor market opens. For now we also assume that the firm has sunk the cost of a vacancy c. Later we explicitly consider the decision to post a vacancy.

A.1 UE hire

We describe how to obtain the 'UE hire' term in (1). Assume the firm's vacancy meets an unemployed worker. Four different cases can arise from the combination of hiring/not hiring and renegotiating/not renegotiating the wage with the incumbent. Our assumption on external negotiation **(A-EN)** requires that in all cases the take-leave wage offer of the firm to the outside worker is $w_2=b$. Our internal negotiation assumption **(A-IN)** requires that the joint value with and without renegotiation is the same and simply equals output $y(z,n)$. Let w_1^* be the incumbent wage after the internal negotiation.

If the firm hires the new worker, its profits are as follows:

$$\underbrace{y(z,2)-w_1-b}_{\text{Without renegotiation}} \quad , \quad \underbrace{y(z,2)-w_1^*-b}_{\text{With renegotiation}},$$

If the firm does not hire the new worker, its profits are

$$\underbrace{y(z,1) - w_1}_{\text{Without renegotiation}} \quad , \quad \underbrace{y(z,1) - w_1^*}_{\text{With renegotiation}}$$

We now describe which case occurs. This requires understanding when our mutual consent assumption **(A-MC)** coupled with limited commitment on layoffs **(A-LC)** bind. In particular, the firm may obtain a credible threat to trigger renegotiation of w_1. We focus first on when a hire occurs.

Hire. A hire *without* renegotiation occurs when the following two conditions hold:

$$\underbrace{y(z,2) - w_1 - b \geq y(z,1) - b}_{\text{No credible threat}} \quad , \quad \underbrace{y(z,2) - w_1 - b \geq y(z,1) - w_1}_{\text{Optimal to hire w/o renegotiation}} \tag{12}$$

The first condition illustrates that the threat to fire the incumbent worker is not credible, which under **(A-MC)** implies no renegotiation. Keeping the incumbent worker at w_1 and employing the outside worker at b delivers a higher value to the firm than the threat of "swapping": firing worker one and hiring the unemployed worker in his place. Given no renegotiation, the second condition ensures hiring is privately optimal for the firm.

A hire *with* renegotiation occurs when the following two conditions hold:

$$\underbrace{y(z,2) - w_1 - b < y(z,1) - b}_{\text{Credible threat}} \quad , \quad \underbrace{y(z,2) - w_1^* - b > y(z,1) - w_1^*}_{\text{Optimal to hire w/ renegotiation}}. \tag{13}$$

The firm has now a credible threat to fire the incumbent worker according to **(A-LC)**. This is possible only under decreasing returns to scale: even though $w_1 < y(z,1)$, the first inequality in (13) implies $w_1 > y(z,2) - y(z,1)$, i.e. the incumbent wage is above its own marginal product. Employing the outside worker at b and keeping the incumbent worker at w_1 delivers a lower value than 'firing and swapping'. The second condition is necessary for hiring to be optimal under the renegotiated wage w_1^* to the incumbent worker.

Under the zero sum game assumption **(A-IN)**, the renegotiated wage w_1^* only redistributes value between the incumbent worker and the firm, but does not affect total value.[43] In addition, it must be individually rational, and so $w_1^* \in [b, y(z,2) - y(z,1)]$. Without further assumptions we cannot say where exactly the new wage lies within this interval, but we can nonetheless pin down allocations.

Rearranging the optimal hiring conditions, we observe that both are satisfied as long as

$$y(z,2) - y(z,1) > b. \tag{14}$$

Note that without internal renegotiation **(A-IN)**, the hiring condition would differ in the two cases. If wages could not be cut and the firm had a credible threat, the incumbent worker would be fired and the

[43] Two relevant cases that would violate this condition are (i) if worker's effort depends on the wage and enters the production function, and (ii) concave utility.

firm would always hire the unemployed worker ($y(z,1) > b$). As a result, to determine firm size next period, one would need to know the incumbent's wage to distinguish between the two cases (thus, in the general model with n workers, the whole wage distribution). Similarly, if a fraction of output were to be lost because of the internal negotiation, a violation of **(A-IN)**, the hiring conditions in (12) and (13) would differ and, again one would need to know wages to determine whether a hire occurs.

We can write inequality (14) in terms of joint value. Workers' values are simply equal to their wage w_i for $i \in \{1,2\}$. The firm's value is simply equal to its profits. The fact that wages are valued linearly by both worker and firm implies that the joint value $\Omega(z,n)$ is independent of wages:

$$\Omega\left(z,n\right) = \underbrace{y(z,n) - \sum_{i=1}^{n} w_i}_{\text{Firm value}} \quad + \quad \underbrace{\sum_{i=1}^{n} w_i}_{\text{Sum of workers' values}} \quad , \quad \text{for any } \left(w_i\right)_{i=1}^{n}.$$

Using the definition of joint value, equation (14) characterizes when the UE hire occurs:

$$\Omega\left(z,2\right) - \Omega\left(z,1\right) > U. \tag{15}$$

Thus, the decision of hiring from unemployment does not depend on wages, but only on productivity, size, and the value of unemployment $U = b$.

No hire. Consider now the cases where no hiring occurs. Recall that, once an unemployed worker is met, the firm has always a credible threat against the incumbent since $w_1 > b$. No hire with renegotiation therefore occurs when the following two conditions hold:

$$\underbrace{y(z,1) - b > y(z,1) - w_1}_{\text{Credible threat}} \quad , \quad \underbrace{y(z,1) - w_1^* \geq y(z,2) - w_1^* - b}_{\text{Optimal to not hire}} \tag{16}$$

After renegotiation the incumbent wage is driven down to w_1^*. Since this outcome represents a redistribution of value between firm and worker then, consistent with **(A-IN)**, the joint value remains $\Omega(z,1)$.[44] Finally, the no-hiring condition in (16) can be re-written as in (14) with the opposite inequality, $\Omega(z,2) - \Omega(z,1) \leq U$.

Combined. The firm hires from unemployment when the *marginal value* of the job seeker exceeds the value of unemployment:

$$\frac{\Omega\left(z,2\right) - \Omega\left(z,1\right)}{2-1} > U. \tag{17}$$

In addition, the joint value of the firm and its workers rises by $\frac{\Omega(z,2)-\Omega(z,1)}{2-1} - U$ when the hire occurs. This is exactly the *UE hire* term in the HJB equation (1). In the case of a hire, incumbent wages may or may not be renegotiated but this has no impact on whether hiring occurs, or how the joint value changes. When this condition fails, the firm does not hire, wages are renegotiated, but the joint value remains constant.

[44] The value before renegotiation was $\Omega\left(z,1\right) = z - w_1 + w_1 = z$. The joint value after renegotiation is $\Omega\left(z,1\right) = z - w_1^* + w_1^* = z$.

All decisions require knowledge of (z, n) only, but not of incumbents' wages.

We now generalize the UE hire case analyzed in the main text to a firm with multiple incumbents.

A.2 *UE hire* when the internal renegotiation involves multiple workers

It is sufficient to consider the case of two incumbent workers, $n = 2$. Without loss of generality, assume that the second worker is paid more than the first, $w_2 > w_1$. As in the approach taken earlier, suppose the firm has posted a vacancy that has met an unemployed worker. We have three cases to consider which illustrate how the firm may use a worker outside the firm to sequentially renegotiate wages of workers inside the firm.

First, the firm hires *without* renegotiation if:

$$\underbrace{y(z,3) - w_1 - w_2 - b > y(z,2) - w_1 - b}_{\text{No credible threat to } w_2} \quad , \quad \underbrace{y(z,3) - w_1 - w_2 - b > y(z,2) - w_1 - w_2}_{\text{Optimal to hire under } (w_1, w_2)} .$$

Hiring with current wages is preferred to replacing the most expensive incumbent—there is no credible threat—, and given no renegotiation, hiring is optimal. Since $w_2 > w_1$, no credible threat to worker 2 implies no credible threat to worker 1.

Second, the firm hires *with* renegotiation with worker 2 if:

$$\underbrace{y(z,2) - w_1 - b > y(z,3) - w_1 - w_2 - b > y(z,2) - w_2 - b}_{\text{Credible threat for worker 2 only}} \quad , \quad \underbrace{y(z,3) - w_1 - w_2^* - b > y(z,2) - w_1 - w_2^*}_{\text{Optimal to hire under } (w_1, w_2^*)} .$$

The threat is credible for worker 2, but is not for worker 1, and, conditional on renegotiating to (w_1, w_2^*), hiring is optimal.

Third, the firm hires *with* renegotiation with *both* workers if:

$$\underbrace{y(z,2) - w_1 - b > y(z,2) - w_2 - b > y(z,3) - w_1 - w_2 - b}_{\text{Credible threat for both workers}} \quad , \quad \underbrace{y(z,3) - w_1^* - w_2^* - b > y(z,2) - w_1^* - w_2^*}_{\text{Optimal to hire under } (w_1^*, w_2^*)} .$$

In all three cases, the optimal hiring condition can be written in terms of joint value as:

$$\frac{\Omega(z,3) - \Omega(z,2)}{3-2} > U. \tag{18}$$

This last inequality does not depend on the order of the internal negotiation between firm and workers. In conclusion, the distribution of wages among incumbents again determines the patterns of wage renegotiation, but is immaterial for the sufficient condition for hiring. Hiring occurs whenever the marginal value of adding a worker to the coalition exceeds the value of unemployment.

Assumption **(A-LC-c)** that was not present in the one worker example plays a role here. Suppose that the renegotiated wage for worker 2 is pushed all the way down to b, making her indifferent between staying and quitting. Worker 1 could transfer a negligible amount to worker 2 in exchange of her quitting, which would raise the firm's marginal product and, possibly, remove its own threat. This is problematic for the representation because in this latter case the hiring condition becomes $y(z,2) - w_1 - b > y(z,1) - w_1$, distinct from (18). Thus, to know whether a firm hires or not, one would need to know the wage distribution inside the firm. **(A-LC-c)** is sufficient to rule out transfers among workers and to prevent this scenario from happening.

Note that this transfer scheme between workers occurring during the internal negotiation changes the joint value, and hence one can think of **(A-LC-c)** as being subsumed into **(A-IN)** already.

In what follows we return to the case where the firm has only one worker.

A.3 *EE hire*

Now suppose that the worker matched with the firm's vacancy is currently employed at another firm with productivity z' and a single worker $n' = 1$. The situation is not that different from *UE hire*, except that the potential hire may have a better outside option in the form of the retention offer made to her by her current employer under **(A-EN)**. To see the similarity for now we fix this wage offer at $\overline{w}$. The same four cases described in section A.1 can arise, except with $\overline{w}$ playing the role of b.[45] We can therefore reason as before and jump to the result that hiring will occur if and only if the following counterpart to (15) holds:

$$\Omega(z,2) - \Omega(z,1) > \overline{w}.$$

We now determine the poached worker's outside option $\overline{w}$. The poached firm's willingness to pay is a wage $\widetilde{w}$ that makes it indifferent between retaining and releasing the worker: $y(z',1) - \widetilde{w} = 0$. Hence, the contacted worker switches to the new employer as long as the poaching firm offers $\overline{w} \geq \widetilde{w} = y(z',1)$. Bertrand competition between the two firms implies that the poaching firm offers $\overline{w} = y(z',1)$, which is exactly the marginal value of the worker at the poached firm. As in the case of *UE hire*, whether *EE hire* occurs can be summarized by joint values:

$$\frac{\Omega(z,2) - \Omega(z,1)}{2-1} > \frac{\Omega(z',1) - \Omega(z',0)}{1-0}. \tag{19}$$

The *EE hire* decision is entirely characterized by knowledge of the pair (z,n) for the two firms.[46] The value gain to the firm and its workers is the difference between the left-hand side and right-hand side of equation (19). This comparison of marginal values is precisely the *EE hire* term in the HJB equation (1).

Finally, this exercise explains the absence of a *EE quit* term in (1). The payment received by its

[45] Renegotiation will happen for different values of w_1 in the no hire case. Indeed, to establish the presence of a credible threat w_1 must be compared to $\overline{w}$ instead of b, but this has no allocative implications for the hiring decisions.

[46] The case when the firm meets a worker at a firm with $(z', 2, w_1, w_2)$ is similar. Suppose the firm meets worker 1. The poached firm has the additional option of cutting w_2, but this is inconsequential for the argument because it only redistributes value within the poached-from firm.

poached worker is equal to the poached coalition's willingness to pay, which is in turn exactly equal to the worker's marginal value to the coalition. The joint value of the poached coalition therefore does not change as it loses its worker. *EE quit* events play an important role in the dynamics of employment at the firm, but no role in the dynamics of $\Omega(z,n)$.

A.4 Vacancy posting

Up to this point we assumed that a meeting between a hiring firm and a job seeker had already occurred. We now turn to the vacancy posting decision and explain why **(A-VP)** is crucial for tractability.

Recall that in the hiring scenarios just analyzed, two cases arise when the firm can credibly force a wage cut: (i) when it hires and the incumbent wage is above the post-hire new marginal product; (ii) when hiring is not profitable, but the firm can credibly 'fire and swap', i.e. as long as the reservation wage of the external worker met through search is below the incumbent wage. The firm has therefore incentives to spend resources on vacancy posting for the sole purpose of transfering value between agents, a privately inefficient outcome. The amount spent would depend on the incumbent's wage, breaking the tractability of our representation. Private efficiency reinstates tractability.

We start with the firm's preferred vacancy policy. Without loss of generality, suppose firms only meet unemployed workers (hence, upon a meeting, the 'fire and swap' threat is always credible). The generalization to the case where the worker contacted can be either unemployed or employed is straightforward. Let v be the number of vacancies posted, $c(v)$ the associated cost, and qv the probability a single vacancy meets a single worker. If no meeting occurs, then as per **(A-MC)**, w_1 does not change so the value of the firm does not change. The firm maximizing the expected return from vacancy posting net of costs is:

$$\max_{v} \quad -c(v) + qv\Big[\max\Big\{\underbrace{y(z,2) - w_1' - b}_{\text{Hire (cases 1\&2)}},\ \underbrace{y(z,1) - b}_{\text{No hire (case 3)}}\Big\} - \Big(y(z,1) - w_1\Big)\Big] \quad ,$$

Following a meeting, three cases may occur. In **Case 1**, the firm hires and there is no renegotiation, $w_1' = w_1$. This case arises when the wage of the incumbent worker is low enough. Then, adding a second worker does not reduce the marginal product of labor down to the point where the firm has a credible layoff threat. In **Case 2**, the firm hires but the wage of the incumbent is renegotiated down to $w_1' = w_1^*$. In this case, diminishing marginal returns drive the marginal product of labor with two workers below the incumbent's initial wage. In **Case 3**, the firm is better off not hiring, but under the threat of swapping out the incumbent, renegotiates w_1 down to b. The firm's preferred vacancy policy v^f then equates marginal cost to marginal expected return:

$$c_v\left(v^f\right) = q\Big[\max\Big\{y(z,2) - w_1' - b\,,\, y(z,1) - b\Big\} - \Big(y(z,1) - w_1\Big)\Big]. \tag{20}$$

The first-order condition (20) highlights that the firm's preferred vacancy policy depends on the incumbent's wage w_1 because this wage determines the gains from forcing a renegotiation through vacancy posting. This dependence is a source of intractability because, in the general model with n workers, (20) would depend on the entire wage distribution inside the firm.

Our assumption **(A-VP)** ensures that firms do not post v^f, but instead post the privately efficient amount of vacancies which does not depend on worker wages. We now show how our micro-foundation **(A-VPI)** implements **(A-VP)**.

Case 1 – Hire without renegotiation. In this case the outcome is already *privately efficient*. The worker's value does not decrease ($w_1' = w_1$), and by the fact that a hire occurs, the firm's value must increase. We can also write the expected return as $qv\left[\Omega(z,2) - \Omega(z,1) - U\right]$. Since the return is independent of w_1, then the efficient vacancy policy is independent of w_1. The firm is choosing vacancies as if it were maximizing the joint surplus without having to appeal to additional assumptions.

In cases 2 and 3, the outcome is *privately inefficient* because the firm may profit from vacancies that, if met by a job seeker, deliver a credible threat to cut the incumbent's wage to $w_1' < w_1$.

Our assumption **(A-VPI)** allows the worker to correct for this over-posting by conceding a single pay cut in exchange for an alternative level of vacancies.[47] The firm will accept this wage cut and choose the worker's preferred vacancies if it delivers at least the value obtained under the firm's preferred vacancies v^f. We show that the worker's preferred package satisfying incentive compatibility restores efficiency in vacancy posting.

Case 2 – Hire with renegotiation. In this case, the incumbent's wage w_1 is high enough that the firm finds it profitable to raise the contact probability with an unemployed worker beyond what would be efficient. Although the hiring outcome is efficient ex-post, too much resources are spent on vacancies ex-ante. Let w_1^* be the renegotiated wage after a meeting. The worker chooses a package of vacancies and a wage cut in all states (v^w, x) that solves:

$$\max_{v^w,\, x}\; qv^w\Big(w_1^* - w_1\Big) - x \tag{21}$$

subject to

$$\begin{aligned} & qv^w\Big[\Big(y(z,2) - (w_1 - x) - b\Big) - \Big(y(z,1) - w_1\Big)\Big] - c(v^w) \\ \geq\;\; & qv^f\Big[\Big(y(z,2) - w_1^* - b\Big) \qquad\;\; - \Big(y(z,1) - w_1\Big)\Big] - c(v^f) \qquad\qquad (IC) \end{aligned}$$

[47] A pay cut regardless of the outcome of the search for a new worker maps exactly into a transfer from worker to firm, which is how we approach the proof. Promising *state-contingent* wage cuts that depend on who the firm meets or whether a meeting occurs is not possible given our assumption of what is verifiable and contractible. Even if these states were verifiable, the result would only be for the worker to offer a menu of wage-cuts across states. This would increase worker value but not change allocations, hence for consistency with the rest of our assumptions, we assume a single wage cut.

The worker anticipates that after a meeting their wage will be renegotiated to $w_1^* < w_1$. Given this wage cut, the worker seeks to limit the probability of this event by cutting back on vacancies. Incentive compatibility (IC) requires that as the worker cuts vacancies it also cuts its wage so that the firm accepts the proposed policy v^w over v^f.

The Pareto problem (21) yields the result that vacancy posting is independent of w_1. First, given the linear objective function, (IC) holds with equality. Thus, we can substitute out x. Second, the zero-sum game assumption **(A-IN)** implies that w_1^* is a renegotiated wage that only redistributes value and hence drops out. Third, all terms that do not depend on (x, v^w) are irrelevant to the worker's decision. Adopting the value notation, this leaves the following objective function:

$$\max_{v^w} qv^w \left[\Big(\Omega(z,2) - U \Big) - \Omega(z,1) \right] \; - \; c\left(v^w\right).$$

The decision can therefore be characterized by the *privately efficient return*, which is the change in joint value net of the cost of the new hire, $\Omega(z,2) - \Omega(z,1) - U$.

Case 3 – No hire with renegotiation. In this case the 'fire and swap' threat is credible. The incumbent's wage w_1 is high enough and the marginal product of an additional worker is below b. Replacing the return to hiring by the wage cut for the incumbent worker, the previous logic delivers

$$\max_{v^w} \; qv^w \left[\Omega(z,1) - \Omega(z,1) \right] \; - \; c\left(v^w\right) \qquad \Longrightarrow \qquad v^w = 0$$

Absent the transfer from worker to firm, the firm would post positive vacancies v^f even if the return from hiring is negative, i.e. $\Omega(z,2) - \Omega(z,1) < U$ to induce a wage cut, and v^f would depend on w_1. Under **(A-VPI)**, the worker takes a preemptive wage cut, and vacancies are zero, the efficient amount in this case.

Combined. Combining all three cases, privately efficient vacancies solve

$$\max_{v} qv \left[\max \left\{ \frac{\Omega(z,2) - \Omega(z,1)}{2-1} - U, 0 \right\} \right] - c(v).$$

Note three properties of this solution. First, the firm always hires when it meets an unemployed worker. Second, optimal vacancy posting equates the marginal gain in joint value to the marginal cost of a vacancy, and it only depends on (z,n). Third, this condition is the flip-side of the separation frontier. Expression (1) states that if $\Omega_n(z,n) > U$, then the firm will not separate with workers. The terms inside the max expression say that if this is true, then the firm will post vacancies.[48]

We conclude that under **(A-VPI)**, the joint value is sufficient to characterize the vacancy decision. The distribution of wages in the firm is immaterial.

[48] It is possible to determine the optimal wage cut x that delivers the efficient policy, but throughout the paper we focus on allocations only.

Multiple incumbents. When the firm employs more than one worker, the efficient transfer scheme can be implemented by randomly selecting a worker under threat to offer a package of wage-cuts and vacancies. In exchange, the firm posts the efficient number of vacancies. Under such a scheme, the initiating worker is strictly better off while the firm and the other workers are indifferent. We establish this case in detail in Appendix II.

A.5 Layoffs, quits, exit, entry

Having described most of the terms in the HJB (1), we conclude with the boundary conditions for exit, layoffs and the free entry condition.

Layoffs. Consider now a firm with $n = 2$ workers paid (w_1, w_2), and assume that $w_1 < y(z, 1)$ such that worker 1 is never under threat of layoff. The firm has a credible threat to fire worker 2 if

$$y(z, 1) - w_1 > y(z, 2) - w_1 - w_2.$$

Such a situation may occur if, for example, productivity has just declined. The firm has a credible threat to negotiate down to a wage level w_2^* such that $y(z, 1) - w_1 = y(z, 2) - w_1 - w_2^*$ and keep worker 2 employed. From the worker's perspective, it is individually rational to accept any wage w_2^* above b. Worker 2 is laid off if $y(z, 1) - w_1 > y(z, 2) - w_1 - b$. In terms of joint value, this can be written in exactly the form of the layoff frontier (2):

$$\frac{\Omega(z, 2) - \Omega(z, 1)}{2 - 1} < U.$$

The firm lays off workers until the marginal joint value of the worker is equal to the value of unemployment.[49] As noted earlier, this is the complement to the condition for posting vacancies. The special case with $n = 1$ of this scenario also arises in the one worker-one firm model with productivity shocks of Postel-Vinay and Turon (2010).

Quits to unemployment. Since in this static model workers will accept a renegotiated wage down to $w_i^* = b$, they will only quit at the point where the firm has a credible threat to lower wages below b. This is exactly the point at which the marginal value is equal to the value of unemployment. In this sense *layoffs* as described above are indistinguishable from quits to unemployment, as in any model with privately efficient separations. For ease of language all *endogenous UE* transitions are referred to as *layoffs*, and we use *quits* to refer only to *EE* transitions.

[49] Note that, when both workers are under threat, the particular order in which values of workers are reduced is immaterial to the condition $\Omega(z, 2) - \Omega(z, 1) < U$. One could for example lower the wages of both workers proportionally, increasing the value of the firm, but a worker *must* be fired if $y(z, 2) - w_1^* - b < y(z, 1) - w_1^*$ for any $w_1^* \geq b$.

Finally, recall that in the dynamic model unemployed job seekers are promised a wage that implements a value U to them. If events occur in the firm that reduce the continuation value to that worker below U (e.g., a negative productivity shock), the incumbent may have a credible threat to quit and renegotiate her wage to restore its value at U, or above it, depending on the details of the internal negotiation. However, such renegotiation is, again, only a transfer of value within the firm. Separations remain privately efficient even in the dynamic model.

Exit. Now consider the exit decision of a firm with one worker. The private value of exit to the firm is the scrap value $\vartheta > 0$. The firm therefore exits if and only if $y(z,1) - w_1^* < \vartheta$, where w_1^* is a possibly renegotiated wage contingent on the firm remaining in operation. If the profit from operating at the lowest possible renegotiated wage $w_1^* = b$ is greater than ϑ, then the firm will continue to operate. Hence, the firm exits if $y(z,1) - b < \vartheta$, and the renegotiated wage only affects the distribution of value.[50] The exit condition can be written as $\Omega(z,1) - U < \vartheta$, and in the general case of n workers is exactly the boundary condition in (1): $\Omega\left(z,n\right) - nU < \vartheta$.

Entry. Upon entry the firm has n_0 workers hired from unemployment. The private entry cost of the firm is c_0, so entry requires $\int y(z,n_0) d\Pi_0(z) - n_0 b > c_0$. Using $\Omega(n,z) = y(z,n)$ and $U = b$, this requires $\int \Omega\left(z,n_0\right) d\Pi_0(z) > c_0 + n_0 U$.

A.6 From static to dynamic

This static example showcases how to obtain every component of (1) from our set of assumptions. Appendix II generalizes this proof to the dynamic case. Two insights assist us. First, the proof begins with a discrete workforce. Here we are helped by continuous time, which removes complicated binomial probabilities of one, two, three, etc. incumbent workers meeting a competitor's vacancy. Second, we take the continuous workforce limit of the discrete workforce HJB equation. This limit delivers the joint value representation (1) in terms of the derivative of the joint value function rather than differences of values which, when moving up or down by one worker, are symmetric due to continuous differentiability.

B Characterization of surplus function

Here we prove the comparative statics on the surplus function $S(n,z)$ discussed in the main text.

[50] The firm has no credible threat to reduce w_1 if $y(z,1) - w_1 > \vartheta$. The firm can credibly threaten exit if $\vartheta \in (y(z,1) - w_1), y(z,1) - b)$, but in this case w_1 can be reduced to a point where this threat is no longer credible.

B.1 Hamilton Jacobi Bellman Variational Inequality for Total Value

Before characterizing these conditions, we note that the joint value representation (1) and smooth-pasting boundary conditions that define the exit and layoff boundaries (2) are derived from solving the following Hamilton-Jacobi-Bellman-Variational-Inequality (see Pham, 2009), which we present here for completeness. Its general formulation in terms of optimal switching between three regimes (operation, layoffs, exit) on the entire positive quadrant, can be written as the following system:

$$\begin{aligned} \max \Bigg\{ & -\rho\Omega\left(z,n\right) + \max_{v\geq 0} -\delta n[\Omega_n(z,n) - U] + qv\phi\left[\Omega_n(z,n) - U\right] \\ & + qv\left(1-\phi\right) \int \max\left[\Omega_n(z,n) - \Omega_n(z',n')\ ,\ 0\right] d\widetilde{H_n}\left(z',n'\right) + \mu(z)\Omega_z(z,n) + \frac{\sigma(z)^2}{2}\Omega_{zz}(z,n)\ ; \\ & \underbrace{\vartheta + nU - \Omega(z,n)}_{\text{Exit}}\ ;\ \underbrace{\max_{k\in[0,n]} \Omega(z,k) + (n-k)U - \Omega(z,n)}_{\text{Layoff}} \Bigg\} = 0 \quad , \quad \forall (z,n) \in \mathbb{R}^2_+ \end{aligned} \tag{22}$$

The HJBVI implies necessary "value-matching" and "smooth-pasting" boundary conditions: see Brekke and Oksendal (1990), Peskir and Shiryaev (2006) and Stokey (2009).

Value matching conditions are standard, and simply state that the value function must be continuous at the exit and separation boundaries. Smooth pasting conditions obtain only when coalitions are actually crossing the exit or layoff boundaries. Intuitively, coalitions can then take an interior first-order optimality condition when they choose the stopping boundary. Thus, smooth pasting obtains either when there is volatility, or when the drift pushes coalitions outside of the continuation region.

Combining these observations, for exit, we have a value matching condition that holds for the entire boundary $n_E^*(z)$, and a smooth pasting condition in the n direction that holds only where the drift is negative and firms actually exit. We have a smooth pasting condition in the z direction that holds for the entire boundary $n_E^*(z)$ because there is volatility in the z direction. We collect these conditions in Conditions (iii) in Section 4.1.

B.2 S is increasing in n

The no-endogenous-separations condition $S_n \geq 0$ implies that the surplus is increasing in n.

B.3 S is increasing in z

Re-write the problem in terms of $x = \log z$. Denote with a slight abuse of notation $y(x,n) = y(e^x, n)$.

$$\begin{aligned}
\rho S(x,n) &= \max_{v\geq 0} y(x,n) - nb - c(v) \\
&+ [q\phi v - \delta n] S_n(x,n) + q(1-\phi) v \mathcal{H}(S_n(x,n)) \\
&+ \mu S_x(x,n) + \frac{\sigma^2}{2} S_{xx}(x,n)
\end{aligned}$$

where we integrated by parts, and denoted $\mathcal{H}(s) = \int_0^s H_n(r)dr$. Denote $\zeta(x,n) = S_x(x,n)$. Differentiate the Bellman equation w.r.t. x and use the envelope theorem to obtain

$$\begin{aligned}
\rho\zeta(x,n) &= y_x(x,n) \\
&+ \Big\{ \big[q\phi + q(1-\phi)H_n(S_n(x,n))\big] v^*(x,n) - \delta n \Big\} \zeta_n(x,n) \\
&+ \mu \zeta_x(x,n) + \frac{\sigma^2}{2} \zeta_{xx}(x,n)
\end{aligned}$$

Now consider the stochastic process defined by

$$\begin{aligned}
dx_t &= \mu dt + \sigma dW_t \\
dn_t &= \Big\{ \big[q(1-\phi)H_n(S_n(x_t,n_t)) + q\phi\big] v^*(x_t,n_t) - \delta n_t \Big\} dt
\end{aligned} \tag{23}$$

This correponds to the true stochastic process for productivity, but a hypothetical process for employment, that in general differes from the realized one. We can now use the Feynman-Kac formula (Pham 2009) to go back to the sequential formulation:

$$\zeta(x,n) = \mathbb{E}\left[\int_0^T e^{-\rho t} y_x(x_t,n_t) + e^{-\rho T}\zeta(x_T,n_T) \,\Big|\, x_0 = x, n_0 = n, \{x_t, n_t\} \text{ follows (23)} \right]$$

and where T is the hitting time of either the separation of exit region. By assumption, $y_x > 0$, so the contribution of the first part is always positive. On the exit region, smooth-pasting requires that $\zeta = 0$. In the interior of the separation region, $\zeta = 0$. Under our regularity assumption, we thus get $\zeta = 0$ on the layoff boundary. Thus,

$$\zeta(x,n) = \mathbb{E}\left[\int_0^T e^{-\rho t} y_x(x_t,n_t) dt \,\Big|\, x_0 = x, n_0 = n, \{x_t, n_t\} \text{ follows (23)} \right] > 0$$

which concludes the proof.

B.4 S is concave in n

Denote $s(z,n) = S_n(z,n)$. Differentiate the Bellman equation w.r.t. n on the interior of the domain, use the envelope theorem and integrate by parts to obtain:

$$\begin{aligned} (\rho+\delta)s(z,n) &= y_n(z,n) - b \\ &+ \Big\{ \big[q\phi + q(1-\phi)H_n(s(z,n))\big] v^*(z,n) - \delta n \Big\} s_n(z,n) \\ &+ \mu(z)s_z(z,n) + \frac{\sigma^2(z)}{2} s_{zz}(z,n) \end{aligned}$$

Recall that

$$(1+\gamma)\bar{c}[v^*(z,n)]^\gamma = q\phi s(z,n) + q(1-\phi)\mathcal{H}(s(z,n)) \tag{24}$$

In particular, differentiating w.r.t. n,

$$\gamma(1+\gamma)\bar{c}[v^*(z,n)]^{\gamma-1}v_n^*(z,n) = \big[q\phi + q(1-\phi)H_n(s(z,n))\big]s_n(z,n)$$

and so

$$\gamma\frac{v_n^*(z,n)}{v^*(z,n)} = \frac{\phi + (1-\phi)H_n(s(z,n))}{\phi + (1-\phi)\overline{H}(s(z,n))} \frac{s_n(z,n)}{s(z,n)}$$

where $\overline{H}(s) = \frac{\mathcal{H}(s)}{s} \leq 1$. Now denote $\zeta(z,n) = s_n(z,n) = S_{nn}(z,n)$. Differentiate the recursion for s w.r.t. n to obtain

$$\begin{aligned} &\left(\rho + 2\delta - q(1-\phi)H_n'(s(z,n)v^*(z,n)s_n(z,n) - q[\phi + (1-\phi)H_n(s(z,n))v_n^*(z,n)\right)\zeta(z,n) \\ =\; & y_{nn}(z,n) \\ +\; & \Big\{ \big[\lambda\phi + \lambda(1-\phi)H_n(s(z,n)\big] v^*(z,n) - \delta n \Big\} \zeta_n(z,n) \\ +\; & \mu(z)\zeta_z(z,n) + \frac{\sigma^2(z)}{2}\zeta_{zz}(z,n) \end{aligned}$$

Now define the "effective discount rate"

$$\begin{aligned} R(z,n,s_n(z,n)) &= \rho + 2\delta - q(1-\phi)H_n'(s(z,n)v^*(z,n)s_n(z,n) - q[\phi + (1-\phi)H_n(s(z,n))]v_n^*(z,n) \\ &= \rho + 2\delta - qv^*(z,n)s_n(z,n)\underbrace{\left\{(1-\phi)H_n'(s(z,n)) + \frac{\phi + (1-\phi)H_n(s(z,n))}{\gamma s(z,n)}\frac{\phi + (1-\phi)H_n(s(z,n))}{\phi + (1-\phi)\overline{H}(s(z,n))}\right\}}_{\equiv P(z,n)>0} \end{aligned}$$

where the second equality uses the expression for v_n^* derived above. Define the stochastic process

$$\begin{aligned} dz_t &= \mu(z_t)dt + \sigma(z_t)dW_t \\ dn_t &= \Big\{ \big[q(1-\phi)H_n(S_n(z_t,n_t)) + q\phi\big] v^*(z_t,n_t) - \delta n_t \Big\} dt \end{aligned} \tag{25}$$

As before, we can use the Feynman-Kac formula to obtain

$$\zeta(z,n) \;=\; \mathbb{E}\Bigg[\int_0^T e^{-\int_0^t R(z_\tau,n_\tau,\zeta(z_\tau,n_\tau))d\tau} y_{nn}(z_t,n_t)dt + e^{-\int_0^T R(z_\tau,n_\tau,\zeta(z_\tau,n_\tau))d\tau T}\zeta(z_T,n_T) \\ \Bigg| \; z_0 = z, n_0 = n, \{z_t,n_t\} \text{ follows (25)}\Bigg]$$

for T the first hitting time of the exit/separation region. The contribution of the first term is always negative. Note that ζ enters in the effective discount rate. Inside the separation region and in the exit regions, $\zeta = 0$. We restrict attention to twice continuously differentiable functions, so $\zeta = 0$ on the exit and separation frontiers. Then

$$\zeta(z,n) \;=\; \mathbb{E}\Bigg[\int_0^T e^{-\int_0^t R(z_\tau,n_\tau,\zeta(z_\tau,n_\tau))d\tau} y_{nn}(z_t,n_t)dt \;\Bigg|\; z_0 = z, n_0 = n, \{z_t,n_t\} \text{ follows (25)}\Bigg] < 0$$

which concludes the proof.

B.5 S is supermodular in $(\log z, n)$

Denote again $s(x,n) = S_n(x,n)$, where $x = \log z$. Recall that

$$\begin{aligned}(\rho+\delta)s(x,n) \;&=\; y_n(x,n) - b \\ &+\; \Big\{ \big[q\phi + q(1-\phi)H_n(s(x,n)\big] v^*(x,n) - \delta n \Big\} s_n(x,n) \\ &+\; \mu s_x(x,n) + \frac{\sigma^2}{2} s_{xx}(x,n)\end{aligned}$$

and that

$$(1+\gamma)\bar{c}[v^*(x,n)]^\gamma = q\phi s(x,n) + q(1-\phi)\mathcal{H}(s(x,n))$$

In particular, differentiating w.r.t. x,

$$\gamma \frac{v_x^*(x,n)}{v^*(x,n)} = \frac{\phi + (1-\phi)H_n(s(x,n))}{\phi + (1-\phi)\overline{H}(s(x,n))} \frac{s_x(x,n)}{s(x,n)}$$

Now denote $\zeta(x,n) = s_x(x,n) = S_{xn}(x,n)$. Differentiate the recursion for $s(x,n)$ w.r.t. x to obtain

$$\begin{aligned}&\Big(\rho + \delta - q(1-\phi)H_n'(s(x,n)v^*(x,n)s_x(x,n) - q[\phi + (1-\phi)H_n(s(x,n))v_x^*(x,n)\Big)\zeta(x,n) \\ =\;& y_{nx}(x,n) \\ +\;& \Big\{ \big[\lambda\phi + \lambda(1-\phi)H_n(s(x,n)\big] v^*(x,n) - \delta n \Big\} \zeta_n(x,n) + \mu\zeta_x(x,n) + \frac{\sigma^2}{2}\zeta_{xx}(x,n)\end{aligned}$$

As before, define the "effective discount rate"

$$\begin{aligned} R(x,n,s_x(x,n)) &= \rho+\delta-q(1-\phi)H_n'(s(x,n)v^*(x,n)s_x(x,n)-q[\phi+(1-\phi)H_n(s(x,n))]v_x^*(x,n) \\ &= \rho+\delta-qv^*(x,n)s_x(x,n)\underbrace{\left\{(1-\phi)H_n'(s(x,n))+\frac{\phi+(1-\phi)H_n(s(x,n))}{\gamma s(x,n)}\frac{\phi+(1-\phi)H_n(s(x,n))}{\phi+(1-\phi)\overline{H}(s(x,n))}\right\}}_{\equiv P(x,n)>0} \end{aligned}$$

where the second equality uses the expression for v_n^* derived above. As before, define the stochastic process

$$\begin{aligned} dx_t &= \mu dt+\sigma dW_t \\ dn_t &= \Big\{\big[q(1-\phi)H_n(S_n(e^{x_t},n_t))+q\phi\big]v^*(x_t,n_t)-\delta n_t\Big\}dt \end{aligned} \tag{26}$$

As before, we can use the Feynman-Kac formula to obtain

$$\begin{aligned} \zeta(x,n) &= \mathbb{E}\Bigg[\int_0^T e^{-\int_0^t R(x_\tau,n_\tau,\zeta(x_\tau,n_\tau))d\tau}y_{nx}(x_t,n_t)dt+e^{-\int_0^T R(x_\tau,n_\tau,\zeta(x_\tau,n_\tau))d\tau T}\zeta(x_T,n_T) \\ &\qquad\qquad \Bigg| \; x_0=z,n_0=n,\{x_t,n_t\}\text{ follows (26)}\Bigg] \end{aligned}$$

for T the first hitting time of the exit/separation region. The contribution of the first term is always positive. Inside the separation region and in the exit regions, $\zeta=0$. We restrict attention to twice continuously differentiable functions, so $\zeta=0$ on the exit and separation frontiers. Then

$$\zeta(x,n) = \mathbb{E}\left[\int_0^T e^{-\int_0^t R(x_\tau,n_\tau,\zeta(x_\tau,n_\tau))d\tau}y_{nx}(x_t,n_t)dt \;\Bigg|\; x_0=z,n_0=n,\{x_t,n_t\}\text{ follows (26)}\right]>0$$

which concludes the proof.

B.6 Net employment growth

Denote again $s(z,n)=S_n(z,n)$. Net employment growth in the continuation region is

$$\frac{dn_t}{dt} = q\Big[\phi+(1-\phi)H_n(s(z,n))\Big]v^*(z,n)-\lambda^E(1-H_v(s(z,n)))n-\delta n\equiv g(z,n)$$

Using the expression the optimal vacancy condition $v^*(z,n)$ in (24):

$$\begin{aligned} g(z,n) &= \frac{q^{1+1/\gamma}}{[(1+\gamma)\bar{c}]^{1/\gamma}}\Big(\phi+(1-\phi)H_n(s(z,n))\Big)\Big(\phi s(z,n)+(1-\phi)\mathcal{H}(s(z,n))\Big)^{1/\gamma} \\ &\quad -\lambda^E(1-H_v(s(z,n)))n-\delta n \end{aligned}$$

From the previous comparative statics on $S(z,n)$, it is straightforward to see that $g(z,n)$ is increasing in $\log z$ and decreasing in n.

C Alternative CRS limit

Consider the surplus equation (4). Assume $\alpha = 1$, $\vartheta = 0$ and vacancy costs homogeneous of degree one in (v,n). Let $\nu = v/n$. In this case, the joint surplus is linear in n, $S(z,n) = \hat{S}(z)n$, where

$$\begin{aligned}(\rho+\delta)\hat{S}(z) &= \max_{\nu\geq 0}\; y(z) - b + q(\theta)\nu\left[\phi\bar{S}(z) + (1-\phi)\int_0^{\hat{S}(z)}\left(\hat{S}(z) - S'\right)dH_n(S')\right] - c(\nu) \\ &\quad + \mu(z)\hat{S}_z(z) + \frac{\sigma^2(z)}{2}\hat{S}_{zz}(z)\end{aligned} \tag{27}$$

where, once again, $H_n(S') = H(z)$ and the *marginal* surplus still depends only on exogenous productivity z. The model continues to behaves like a one-worker-one-firm model for all *firm decisions*, up to rescaling vacancies by size. This economy, however, produces different worker dynamics from the limiting one described in Section 4.3.1 since gross hires now depend on firm size.

D Frictionless limits

D.1 Setup

Frictional problem. Start by recalling the Bellman equation for the joint surplus in the frictional case:

$$\begin{aligned}\rho S(z,n) &= \max_{v}\, y(z,n) - nb - c(v) - \delta S_n(z,n) \\ &+ \quad q(\theta)v\left\{\phi S_n + (1-\phi)\int_0^{S_n} H_n(s)ds\right\} \\ &+ \quad (\mathbb{L}S)(z,n) \\ \text{s.t.} &\qquad S(z,n)\geq 0,\; S_n(z,n)\geq 0\end{aligned} \tag{28}$$

where H_n is the employment-weighted cumulative distribution function of marginal surpluses. $\mathbb{L}$ is the differential operator that encodes the continuation value from productivity shocks, $(\mathbb{L}S)(z,n) = \mu(z)S_z(z,n) + \frac{\sigma(z)^2}{2}S_{zz}(z,n)$. Recall that $\phi = \frac{u}{u+\xi(1-u)}$ is the probability that a vacancy meets an unemployed worker, and q is the vacancy meeting rate.

Inside the continuation region, the density function $h(z,n)$ of the distribution of firms by productivity and size is determined by the stationary KFE

$$0 = -\frac{\partial}{\partial n}\Big(h(z,n)g(z,n)\Big) + (\mathbb{L}^* h)(z,n)$$

where $\mathbb{L}^*$ is the formal adjoint of the operator $\mathbb{L}$, and $g(z,n)$ is the growth rate of employment

$$g(z,n) = q(\theta)v^*(z,n)\Big[\phi + (1-\phi)H_n(S_n(z,n))\Big] - \xi\lambda^U n\Big[1 - H_v(S_n(z,n)),\Big] \tag{29}$$

where λ^U is the meeting rate from unemployment, and ξ the relative search efficiency of the employed.

The mass of entrant firms $\mathtt{m}_0$ is determined by the free-entry condition

$$c_e = \mathbb{E}_0[\max\{S(z,n_0),0\}] \tag{30}$$

where n_0 is initial employment which is a parameter, and $\mathbb{E}_0$ is the expectation operator under the productivity distribution for entrants $\Pi_0(z)$. Note that the surplus is a function of $\mathtt{m}_0$ through the vacancy meeting rate $q(\theta)$, since θ is increasing in $\mathtt{m}_0$.

For ease of exposition, and without loss of generality, we make three additional assumptions. First, we consider isoelastic vacancy cost functions

$$c(v) = \frac{\bar{c}}{1+\gamma}v^{1+\gamma},$$

and normalize $\bar{c} = 1$, but the result does not depend on the particular functional form nor on the normalization. Also, we specialize to a Cobb-Douglas matching function $\mathtt{m}(s,v) = As^{\beta}v^{1-\beta}$, where A is match efficiency, a proxy for labor market frictions. Third, we set to zero exogenous separations to unemployment $\delta = 0$, but endogenous separations when $S(n,z) < 0$ still occur, and we denote by Δ the aggregate endogenous separation rate.

To ease notation, we write $B \approx C$ for a first-order Taylor expansion. We also denote $||S_n|| = \mathbb{E}^{SS}\left[S_n^{1/\gamma}\right]^{\gamma}$, where $\mathbb{E}^{SS}$ denotes the expectation under the steady-state distribution of marginal surpluses. This is also the Lebesgue $(1/\gamma)$-norm of S_n under the steady-state probability measure.

Finally, we note that in characterizing the limits we make use of the fact that both $\mathtt{m}_0$ and $\mathtt{v}$ must remain finite: infinite entry and vacancy costs would violate the economy's resource constraint.

Comparative statics. We describe behavior of the economy in the limit when match efficiency $A \to \infty$. We do so for two different configurations of the economy:

1. No on-the-job-search: $\xi = 0$

2. On-the-job search: $\xi > 0$

D.2 No on-the-job search

Since $\xi = 0$, $\phi = 1$. From (28), the FOC for vacancies gives

$$v^*(z,n) \quad = \quad \left(qS_n\right)^{1/\gamma}. \tag{31}$$

Using this optimality condition in the value function of hiring firms:

$$\begin{aligned} \rho S(z,n) \quad &= \quad y(z,n) - nb + \frac{\gamma}{1+\gamma} \cdot q(\theta)^{\frac{1}{1+\gamma}} S_n^{\frac{1}{1+\gamma}} + (\mathbb{L}S)(z,n) \\ \text{s.t.} \quad & \quad S(z,n) \geq 0,\ S_n(z,n) \geq 0 \end{aligned}$$

which now only depends on $q(\theta)$ as the sole aggregate. Hence, free-entry (30) uniquely pins down $q(\theta)$ to the same number no matter what value A takes. Therefore, the value function always satisfies the same Bellman equation, irrespective of A. Hence, throughout the state space, at any given (n,z), marginal surpluses $S_n(z,n)$ remain the same as A varies. Moreover, since the value $S(z,n)$ is independent from A, so are all the decisions by firms. As a result, the endogenous separation rate Δ always remains the same – and in particular, finite.

Aggregates in the limit We now study how aggregates $\mathtt{v}, \mathtt{u}, \theta$ evolve along this limiting path. Given the matching function these determine all other equilibrium objects: q, λ^U, λ^E.

Integrating both sides of the FOC for vacancies under the firm distribution, and using the matching function which implies that $q = A\theta^{-\beta}$, aggregate vacancies are

$$\mathtt{v} = \mathtt{m}_0 q^{\frac{1}{\gamma}} ||S_n||^{\frac{1}{\gamma}} = \mathtt{m}_0 A^{\frac{1}{\gamma}} \theta^{-\frac{\beta}{\gamma}} ||S_n||^{\frac{1}{\gamma}}$$

Since q remains constant, and $\mathtt{v}$ and $\mathtt{m}_0$ are finite in the limit, then the first equality implies that $||S_n||$ remains finite in the limit.

In the limit, the unemployment rate is $\mathtt{u} \approx \frac{\Delta}{\lambda^U}$. The matching function implies $\lambda^U = A\theta^{1-\beta}$. Combined, the unemployment rate is $\mathtt{u} \approx \Delta A^{-1} \theta^{-(1-\beta)}$. Combining these expressions with the expression for aggregate vacancies $\mathtt{v}$, tightness satisifies

$$\theta = \frac{\mathtt{v}}{\mathtt{u}} \approx \frac{\mathtt{m}_0 A^{\frac{1}{\gamma}} \theta^{-\frac{\beta}{\gamma}} ||S_n||^{\frac{1}{\gamma}}}{\Delta A \theta^{1-\beta}}$$

so that

$$\theta^{\beta \frac{1+\gamma}{\gamma}} \approx \left(\frac{\mathtt{m}_0}{\Delta}\right) ||S_n||^{\frac{1}{\gamma}} A^{\frac{1+\gamma}{\gamma}}.$$

Since $\mathtt{m}_0$, Δ, and $||S_n||$ are finite, θ diverges with A. Therefore, λ_U diverges as well. On the worker side, since λ_U, diverges to infinity, $\mathtt{u}$ goes to zero. On the firm side, $\mathtt{m}_0$ remains finite, but changes such that q remains constant and vacancies remain finite.

Invariant distribution of marginal surpluses We now turn to the invariant distribution $h(z,n)$. After substituting optimal vacancies into (29) evaluated at $\xi = 1 - \phi = 0$, one obtains that the growth of employment in the hiring region is:

$$g(z,n) = q\Big(qS_n(z,n)\Big)^{\frac{1}{\gamma}}.$$

Since $S_n(z,n)$ remains constant throughout the state space, then employment growth in the hiring region remains constant throughout the state space. The firm loses no workers to employment because there is no on-the-job search. Since $S_n(z,n)$ and $U = b/\rho$ both stay unchanged, then the employment outflows to unemployment are still unchanged. Since $S(z,n)$ is unchanged, then the exit decision is also unchanged.

Hence, the law of motion of employment is independent of A and the steady-state distribution $h(z,n)$ is also independent from A. Therefore the values of firms $S(z,n)$ are the same across the state space and the relative mass of firms at each (z,n) is unchanged, despite higher but finite mass of entrants $\mathtt{m}_0$.

D.3 On-the-job search

We now turn to the case in which on-the-job search remains positive at some fixed value $\xi > 0$. We follow the same logic as before, with some additional steps due to on-the-job search.

Consider (28) written in terms of the return on a vacancy $R(S_n)$

$$\begin{aligned} \rho S(z,n) &= \max_v \; y(z,n) - nb - c(v) + q(\theta)vR(S_n) + (\mathbb{L}S)(z,n) \\ \text{s.t.} \quad & S(z,n) \geq \vartheta, \; S_n(z,n) \geq 0 \end{aligned}$$

where

$$R(S_n) = \phi S_n + (1-\phi)\int_0^{S_n} H_n(s)ds \tag{32}$$

The growth of employment is

$$g(z,n) = qv\Big[\phi + (1-\phi)\,H_n\left(S_n(z,n)\right)\Big] - \xi\lambda^U n\Big[1 - H_v\left(S_n(z,n)\right)\Big] \tag{33}$$

Aggregates in the limit We restrict attention to the economically meaningful case in which (1) output and aggregate vacancies remains finite and strictly positive in the limit, and (2) the rate at which workers separate into unemployment remains finite in the limit. These restrictions are equivalent to a guess and verify strategy, in which we guess that (1-2) hold and then verify those conditions.

Consider first meeting rates. Because some measure $\mathtt{n}$ of employed jobseekers are always present regardless of A, effective search effort $\mathtt{s} = \mathtt{u} + \xi\mathtt{n}$ remains finite and positive even if $\mathtt{u}$ goes to zero. By (1), vacancies also remain finite. Combined, these imply that market tightness $\theta = \mathtt{v}/\mathtt{s}$ remains finite. Since $q = A\theta^{-\beta}$ and $\lambda^U = A\theta^{1-\beta}$, then both meeting rates diverge to infinity at the same rate as A.[51]

Consider unemployment and aggregate vacancies. (2) requires that the rate at which workers separate into unemployment is a positive constant Δ in the limit. Since $\mathtt{u} \approx \frac{\Delta}{\lambda^U}$, and λ_U diverges, then the unemployment rate converges to zero. Since the unemployment rate converges to zero, then ϕ also converges to zero and thus $\mathtt{s} = \xi$. Firm level and aggregate vacancies are given by

$$v = q^{\frac{1}{\gamma}} R\left(S_n\right)^{\frac{1}{\gamma}} \qquad , \qquad \mathtt{v} = \mathtt{m}_0 q^{\frac{1}{\gamma}} \left|\left|R\left(S_n\right)\right|\right|^{\frac{1}{\gamma}}. \tag{34}$$

(1) implies that both aggregate vacancies $\mathtt{v}$ and the mass of entering firms $\mathtt{m}_0$ remain finite. Since $\mathtt{v}$ is finite and $\mathtt{m}_0$ is finite, while q diverges at the same rate as A, then $\gamma > 0$ requires $\left|\left|R\left(S_n\right)\right|\right|$ must go to zero at the same rate as A goes to infinity.

Invariant distribution of marginal surpluses We now show that the distribution of marginal surpluses degenerates to a single value on the support of the invariant distribution.

First, we use (34) to express firm level vacancies as a share of aggregate vacancies, where that share is determined by the firms' return on a vacancy relative to the average return:

$$v = \frac{1}{m_0}\left(\frac{R\left(S_n\right)}{\left|\left|R\left(S_n\right)\right|\right|}\right)^{\frac{1}{\gamma}} \mathtt{v} = \frac{1}{m_0}\left(\frac{R\left(S_n\right)}{\left|\left|R\left(S_n\right)\right|\right|}\right)^{\frac{1}{\gamma}} \left(\frac{\lambda^U \xi}{q}\right) \tag{35}$$

where the second equality uses $q = A(\mathtt{v}/\xi)^{-\beta}$, and $\lambda^U = A(\mathtt{v}/\xi)^{1-\beta}$, which jointly imply that $\mathtt{v} = \lambda^U \xi / q$. Now consider the expression for growth of employment inside the continuation region (33), under the limiting case of $\phi = 0$:

$$g(z,n) \approx qvH_n\left(S_n(z,n)\right) - \xi\lambda^U n\Big[1 - H_v\left(S_n(z,n)\right)\Big]$$

Substituting in the expression for firm vacancies (35) and collecting $\lambda^U\xi$ terms:

$$g(z,n) \approx \lambda^U \xi \left\{\frac{1}{\mathtt{m}_0}\left(\frac{R(S_n)}{||R(S_n)||}\right)^{\frac{1}{\gamma}} H_n(S_n) - n\Big[1 - H_v(S_n)\Big]\right\}.$$

Since λ^U diverges but growth must remain finite on the support of the invariant distribution, the term

[51] Strictly speaking, free-entry then ensures that θ is pinned down to a strictly positive value. This proof is more lengthy but does not require any additional assumptions and is available upon request.

in braces must be equal to zero in the limit:

$$\frac{1}{\mathtt{m}_0}\left(\frac{R(S_n)}{||R(S_n)||}\right)^{\frac{1}{\gamma}} H_n(S_n) = n\Big[1 - H_v(S_n)\Big] \tag{36}$$

Using this relation we can show that the distribution of marginal surplus converges point-wise to a degenerate limiting distribution H_n^∞, i.e. for every z there is a unique $n^*(z)$.

We proceed by contradiction. Suppose that H_n converges to a limiting distribution H_n^∞ that is non-degenerate.[52] Consider a firm at the top of the distribution, such that $1 - H_v(S_n) = 0$. The probability that the firm loses a worker is zero, so the right-hand side is zero. However, by the supposition that H_n is non-degenerate, then $R(S_n)$ in (32) converges to a non-zero value, since the firm can increase its value by poaching from workers below it on the ladder. Since there is some $R(S_n)$ that is non-zero, then $||R(S_n)||$ also converges to a non-zero value. Therefore the right hand side of (36) is zero, but the left hand side is positive which violates the above equality, a contradiction. Hence, in the limit H_n^∞ must be degenerate, and marginal surpluses of firms converge to a common limit which we denote S_n^*.

Since the limiting distribution H_n^∞ is degenerate at every z, the invariant joint distribution of employment and productivity lines up along a strip $\{z, n^*(z)\}$ implicitly defined by $S_n(n^*(z), z) = S_n^*$. Since $S_{nn} < 0$ and $S_{zn} > 0$, $n^*(z)$ is strictly increasing.

Unique value for S_n^* on the limiting strip We now show that the unique equilibrium value of the marginal surplus, S_n^*, is zero. The first step of the proof is to express the marginal surplus as the present discounted value of flow marginal products $y_n - b$. Second, we show that these marginal products would be equal to $-b$ should the marginal surplus S_n^* be any strictly positive value. We then conclude that $S_n^* = 0$.

For the first step, we start by maximizing out vacancies in the HJB in (28) to obtain

$$\rho S(z,n) = y(z,n) - bn + \frac{\gamma}{1+\gamma}\left(q\phi S_n(z,n) + q(1-\phi)\int_0^{S_n(z,n)} H_n(s)ds\right)^{\frac{1+\gamma}{\gamma}} + (\mathbb{L}S)(z,n). \tag{37}$$

Differentiate (37) with respect to n to obtain a HJB for the marginal surplus,

$$\rho S_n(z,n) = y_n(z,n) - b + \left(q\phi S_n(z,n) + q(1-\phi)\int_0^{S_n(z,n)} H_n(s)ds\right)^{\frac{1}{\gamma}} \Big(q\phi + q(1-\phi)H_n(S_n(z,n))\Big) S_{nn}(z,n) + (\mathbb{L}S_n)(z,n). \tag{38}$$

[52] So the probability measure of S_n in the cross-section would converge in distribution to a non-degenerate limit.

where the size process associated with marginal surplus dynamics is:

$$\hat{g}(z,n) = \left(q\phi S_n(z,n) + q(1-\phi) \int_0^{S_n(z,n)} H_n(s)ds \right)^{\frac{1}{\gamma}} \Big(q\phi + q(1-\phi) H_n(S_n(z,n)) \Big). \tag{39}$$

Note that $\hat{g}(z,n) \geq 0$, with equality if and only if $S_n(z,n) = 0$.

The HJB (38) therefore re-writes as

$$\rho S_n(z,n) = y_n(z,n) - b + \hat{g}(z,n) S_{nn}(z,n) + (\mathbb{L}S_n)(z,n) \tag{40}$$

Using the Feyman-Kac formula, we obtain a sequential representation of (40)

$$S_n(z,n) = \mathbb{E}_0 \left[\int_0^\infty e^{-\rho t} \big(y_n(z_t, \hat{n}_t) - b \big) dt \middle| z_0 = z, \hat{n}_0 = n \right] \tag{41}$$

where z_t follows the actual productivity process, and $\hat{n}_t$ follows (39).

We are now ready for the second step of the proof. Suppose for a contradiction that $S_n^* > 0$. Then $\hat{g}(z, n^*(z)) = +\infty$ because the first parenthesis in (39) is strictly positive ($q\phi \to \frac{\Delta}{\theta\xi} \in (0, +\infty)$), and the second parenthesis in (39) is infinite since $H_n(S_n^*) = 1$ and $q \to +\infty$. Similarly, for any $n \geq n^*(z)$ such that $S_n(z,n) > 0$, $\hat{g}(z,n) = +\infty$. Together, these observations imply that $\hat{n}_t = +\infty$ for any $t > 0$, starting from $n^*(z)$ at $t = 0$. Intuitively, if the marginal surplus from hiring were to remain always strictly positive, given the infinite meeting rate firms would keep growing without bound, a contradiction.

Under our Inada condition, we obtain, $y_n(z_t, \hat{n}_t) - b = -b < 0$ for any $t > 0$. Using (41), we obtain $S_n(z, n^*(z)) < 0$ a contradiction. Thus, it must be that $S_n^* = 0$.

Optimal size Our goal is now to characterize the optimal size $n^*(z)$. Our strategy is to leverage that the marginal surplus is zero $S_n^* = 0$. We connect the marginal surplus to the static net marginal product of labor $y_n - b$ using the sequential representation (41). This expression relates the marginal surplus to the present discounted value of all future net marginal products. To operationalize this idea, we split the time integral in (41) in several components: a component that captures the immediate future, and a continuation value. We define our candidate optimal size, $\bar{n}(z)$, to be such that $y_n(z, \bar{n}(z)) = b$. To show that $n^*(z) = \bar{n}(z)$, we proceed by contradiction.

Suppose first that $n^*(z) < \bar{n}(z)$, and so $y_n(z, n^*(z)) - b > 0$. Let $\epsilon > 0$, and rewrite (41) as

$$S_n(z, n^*(z)) = \mathbb{E}_0 \left[\int_0^\epsilon e^{-\rho t} \big(y_n(z_t, \hat{n}_t) - b \big) dt \middle| z_0 = z, \hat{n}_0 = n^*(z) \right] + \mathbb{E}_0 \left[e^{-\rho\epsilon} S_n(z_\epsilon, \hat{n}_\epsilon) \middle| z_0 = z, \hat{n}_0 = n^*(z) \right] \tag{42}$$

The first component is just the integral of the marginal product in a small time interval $[0, \epsilon]$. Recall that the marginal product is positive at $t = 0$ by assumption. Thus, by continuity, it must be that the marginal product $y_n(z_t, \hat{n}_t) - b$ is positive for all $t \leq \epsilon$ when ϵ is small enough. Therefore the first component of (42) is strictly positive. The second component is also always positive because $S_n \geq 0$. Thus (42) implies $S_n^* = S_n(z, n^*(z)) > 0$, a contradiction with $S_n^* = 0$. Therefore, we obtain that $n^*(z) \geq \bar{n}(z)$.

Suppose next for a contradiction that $n^*(z) > \bar{n}(z)$, and so $y_n(z, n^*(z)) - b < 0$. Our strategy for this inequality mirrors our previous one. We must however split the integral into three components rather than two to deal with continuation values.

Set $\hat{n}_0 = n^*(z)$, and let $\epsilon > 0$ be small enough. We define the stopping time

$$T = \inf\{t \geq \epsilon : \hat{n}_t = n^*(z_t)\}.$$

Recall that $\hat{g} \geq 0$. Therefore, $\hat{n}_t \geq n^*(z) > \bar{n}(z)$ for any $t > 0$. In addition, by definition of the stopping time T, we also have that $\hat{n}_t > n^*(z_t) > \bar{n}(z_t)$ for all $\epsilon \leq t \leq T$.

Then return to (41) evaluated at $(z, n^*(z))$. Write

$$\begin{aligned} S_n(z, n^*(z)) = \mathbb{E}_0 &\left[\int_0^\epsilon e^{-\rho t} \big(y_n(z_t, \hat{n}_t) - b\big) dt \middle| z_0 = z, \hat{n}_0 = n^*(z) \right] \\ + \mathbb{E}_0 &\left[\int_\epsilon^T e^{-\rho t} \big(y_n(z_t, \hat{n}_t) - b\big) dt \middle| z_0 = z, \hat{n}_0 = n^*(z) \right] \\ + \mathbb{E}_0 &\left[e^{-\rho T} S_n(z_T, n^*(z_T) \middle| z_0 = z, \hat{n}_0 = n^*(z) \right] \end{aligned}$$

Similarly to our previous argument, the first component is strictly negative when ϵ is small enough, by continuity. The second component is strictly negative because, by definition of T, $\hat{n}_t > n^*(z_t) > \bar{n}(z_t)$ for all $\epsilon \leq t \leq T$. The third component is zero since $S_n(z_T, n^*(z_T)) = S_n^* = 0$. Therefore, we obtain that $0 = S_n^* < 0$, a contradiction.

We conclude that $n^*(z) = \bar{n}(z) = \arg\max_n y(z, n) - bn$.

Limiting value function Return to the surplus equation (37). Evaluating at $(z, n^*(z))$, the vacancy return component is equal to zero because $S_n^* = 0$. Therefore,

$$\begin{aligned} \rho S(z, n^*(z)) \quad &= \quad y(z, n^*(z)) - n^*(z)b + \mu(z) S_z(z, n^*(z)) + \frac{\sigma(z)^2}{2} S_{zz}(z, n^*(z)) \\ \text{s.t.} \qquad &S(z, n^*(z)) \geq \vartheta \end{aligned}$$

To arrive at our representation in the main text, we must show that the partial derivatives, e.g. $\frac{\partial S}{\partial z}(z, n^*(z))$, are equal to the total derivatives $\frac{dS(z,n^*(z))}{dz}$. To that end, note that, in the limit

$$\begin{aligned}\frac{dS(z,n^*(z))}{dz} &= \frac{\partial S}{\partial z}(z,n^*(z)) + \frac{\partial S}{\partial n}(z,n^*(z))\frac{dn^*(z)}{dz} \\ &= \frac{\partial S}{\partial z}(z,n^*(z)) + S_n^*\frac{dn^*(z)}{dz}\end{aligned}$$

Because $S_n^* = 0$, we obtain

$$\frac{dS(z,n^*(z))}{dz} = \frac{\partial S}{\partial z}(z,n^*(z))$$

Therefore, in the limit, with a slight abuse of notation, the surplus can be described by the value function evaluated on the strip, $S(z) := S(z, n^*(z))$ which evolves according to

$$\rho S(z) = y(z,n^*(z)) - n^*(z)b + \mu(z)S_z(z) + \frac{\sigma(z)^2}{2}S_{zz}(z)$$

and an exit cut-off determined by $S(z) = \vartheta$. This proves equation (10) in the main text.

Finally, the free-entry condition $\int S(z,n_0)\Pi_0(z)dz = 0$ pins down the mass of firms $\mathtt{m}_0$ and thus market tightness θ along the limit.

E Identification

To illustrate the identification of the model's parameters more formally, we conduct two exercises. First, we show how the minimum distance of the objective function changes as we move each parameter ψ_i in steps in a wide range around ψ_i^*, letting the other parameters $\boldsymbol{\psi}_{-i}^*$ adjust to minimize the distance criterion function. We argue that the model is identified if $\mathcal{G}(\psi_i, \boldsymbol{\psi}_{-i}^*)$ plotted as a function of ψ_i, traces a steep "U" with a minimum at ψ_i^*. Figure E.1 plots this exercise and gives us confidence that our parameter vector is well identified.

Second, in the main text we discussed how each parameter is especially informed by a particular moment, despite the model being jointly identified. To support this argument, Figure E.2 plots each of the eight moments as a function of the corresponding parameter in Panel C of Table 1, keeping all other parameters at their estimated values. All panels in the figure show significant variation in the moment of interest as a function of its respective parameter.

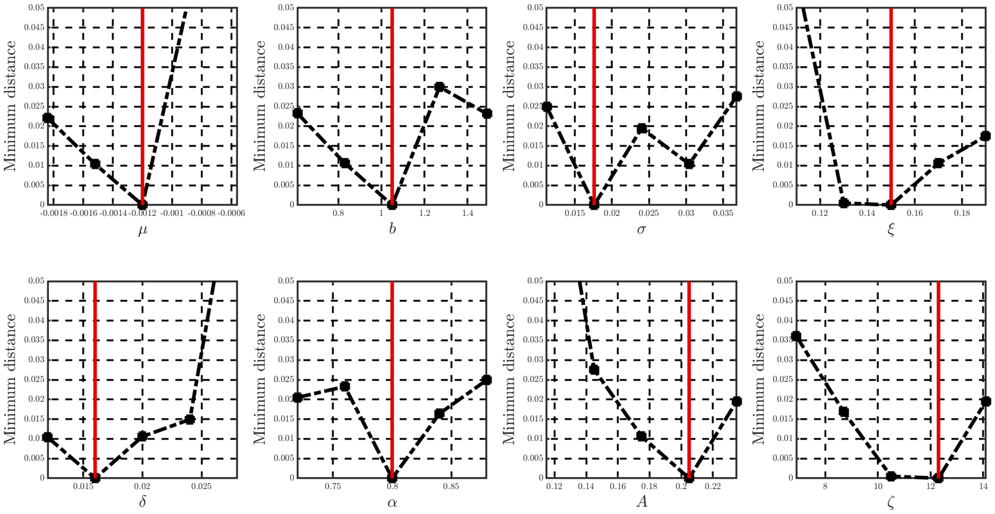


Figure E.1: Minimum distance as function of each parameter

Notes For each parameter $\psi_i \in \{\mu, \ldots, b\}$, the black line plots the minimum distance function $\psi_i \mapsto \mathcal{G}(\psi_i, \boldsymbol{\psi}^*_{-i})$, where $\boldsymbol{\psi}^*_{-i}$ adjusts to minimize the distance criterion. The red vertical line marks the estimated value ψ^*_i listed in Table 1

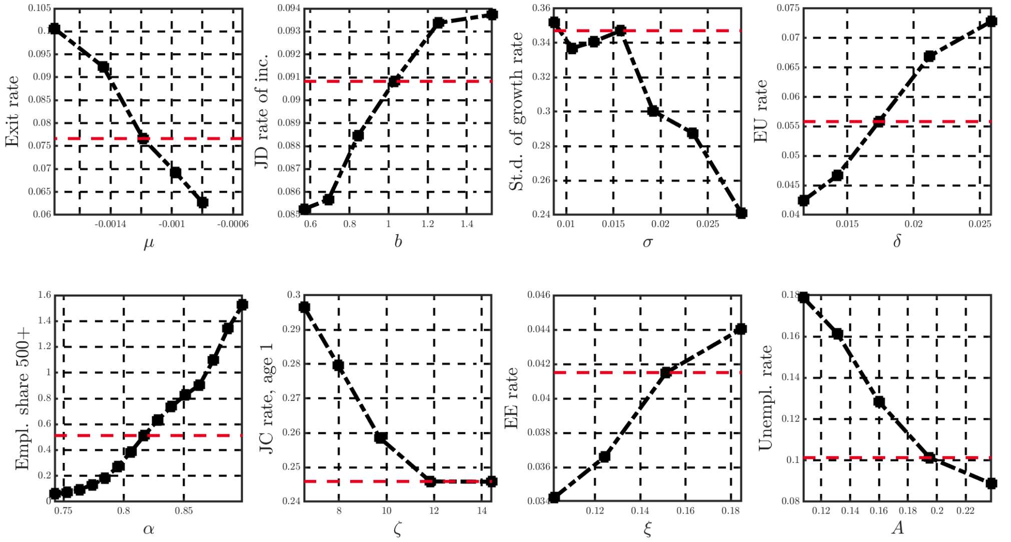


Figure E.2: Each targeted moment against each parameter

Notes This figure plots the relationship between each parameter $\psi_i \in \{\mu, \ldots, b\}$ and the moment aligned with the parameter in Table 1. For each panel, the x-axis plots alternative values of the parameter. The y-axis plots the change in the corresponding moment in the steady state of the model obtained when all other parameters are as in Table 1.