FEDERAL RESERVE BANK
PHILADELPHIA

Working Papers
RESEARCH DEPARTMENT

Vacancy Chains

Michael W. L. Elsby
University of Edinburgh

Ryan Michaels
Federal Reserve Bank of Philadelphia Research Department

Axel Gottfries
University of Edinburgh

David Ratner
Board of Governors of the Federal Reserve System

WP 22-23

PUBLISHED
August 2022

ISSN: 1962-5361

DOI: https://doi.org/10.21799/frbp.wp.2022.23

Vacancy Chains

Michael W. L. Elsby Axel Gottfries

Ryan Michaels David Ratner*

May 2022

Abstract

Replacement hiring—recruitment that seeks to replace positions vacated by workers who quit—plays a central role in establishment dynamics. We document this phenomenon using rich microdata on U.S. establishments, which frequently report no net change in their employment, often for years at a time, despite facing substantial gross turnover in the form of quits. We devise a tractable model in which replacement hiring is driven by a novel structure of frictions, combining firm dynamics, on-the-job search, and investments into job creation that are sunk at the point of replacement. A key implication is the emergence of *vacancy chains*. Quantitatively, the model reconciles the incidence of replacement hiring with the large dispersion of labor productivity across establishments, and largely replicates the empirical volatility and persistence of job creation and, thereby, unemployment.

JEL codes: E32, J63, J64.

Keywords: Quits, replacement hiring, unemployment, vacancies, business cycles.

---

* Elsby and Gottfries: University of Edinburgh. Michaels: Federal Reserve Bank of Philadelphia. Ratner: Federal Reserve Board.

This paper has had a long gestation, with presentation of early iterations starting in 2014. It supersedes a previous working paper (Elsby, Michaels and Ratner 2019). We thank seminar participants at numerous institutions over many years for helpful comments, especially: John Bound, Charlie Brown, Simon Burgess, Jan Grobovsek, Pat Kline, Rasmus Lentz, David Romer, Benjamin Schoefer. Thanks also to Kathy Bauer and Jess Helfand at BLS. Trevor Dworetz provided excellent research assistance. All errors are our own.

This research was conducted with restricted access to Bureau of Labor Statistics (BLS) data. The views expressed here do not necessarily reflect the views of the BLS, the Federal Reserve Bank of Philadelphia, the staff and members of the Federal Reserve Board, or the Federal Reserve System as a whole. Philadelphia Fed working papers are free to download at https://philadelphiafed.org/research-and-data/publications/working-papers.

Elsby, Gottfries and Michaels gratefully acknowledge financial support from the UK Economic and Social Research Council (ESRC), Award reference ES/L009633/1.

What is a vacancy? After several decades of survey research dating back to the 1950s, the Bureau of Labor Statistics[1] (BLS) converged on a definition of a vacancy that includes, among other requirements, the notion that "a specific position exists and there is work available for that position." This definition, implemented at the inception of the Job Openings and Labor Turnover Survey in December 2000, has formed the basis of the leading source of vacancy data ever since, which in turn has become a central reference point for our understanding of labor markets.

In this paper, we argue that this notion of a vacancy has rich economic implications. The presence of a "position" connotes the presence of a sunk investment, be it in physical capital—an empty desk, an unused machine—or organizational capital—the blueprint of task allocations at an establishment. The crucial implication that we explore is that this sunk capital—or "position"—remains even after an employee quits. We show that this observation reconciles key features of both the micro- and macroeconomic behavior of labor markets, from the widespread incidence of inaction in employment adjustment and large disparities in productivity observed across plants, to the empirical volatility and persistence of labor market fluctuations in the aggregate.

We begin in section 1 by documenting a novel set of stylized facts of establishment dynamics using microdata underlying the Quarterly Census of Employment and Wages (QCEW) and the Job Openings and Labor Turnover Survey (JOLTS). These suggest a prominent role for *replacement hiring*—recruitment that replaces positions vacated by quits—in establishment dynamics. Establishments frequently report no *net* change in their employment, often for years at a time, despite facing substantial *gross* turnover in the form of quits. Furthermore, replacement hiring accounts for a large fraction of aggregate hires in the U.S. economy. Consistent with the BLS definition, the observation that establishments go to particular lengths to refill positions vacated by workers who quit further underscores the notion of a vacancy in which sunk investments loom large.

These novel stylized facts motivate a novel structure of frictions. Our point of departure is a class of canonical models that invoke the presence of a *gross* hiring cost as the primary constraint to labor demand (Bentolila and Bertola 1990; Hopenhayn and Rogerson 1993; Mortensen and Pissarides 1994). Motivated by the prominence of replacement hiring, we explore the implications of an additional friction whereby, per the BLS survey, firms must invest in *positions* to expand their workforce. Critically, the costs

[1] For further information on the evolution of these BLS surveys, see Elsby, Michaels and Ratner (2015).

of such investments are sunk at the point of replacement. And, echoing the data, the impetus to such replacement events is the presence of endogenous quits generated by *on-the-job search*. Together, these ingredients provide a prototype model of replacement hiring in which positions vacated by quits are frequently refilled.

To engage with the establishment-level stylized facts we document, in section 2 we embed this structure of frictions into a model of firm dynamics with on-the-job search. Firms operating a decreasing-returns technology face idiosyncratic shocks that drive changes in their desired employment. Heterogeneity induced by idiosyncratic shocks gives rise to a labor market characterized by an endogenous hierarchy of firms, ranked by the surplus they can offer their workers. Given the opportunity, workers searching on-the-job quit to firms that offer higher worker surpluses. Firms thus need to know the distribution of worker surpluses to infer their turnover and, thereby, their optimal demand for labor. This distribution, in turn, is implied by the aggregate consequences of firms' labor demand decisions. Labor market equilibrium thus involves the technical challenge of finding a fixed point for the distribution of worker surpluses. One of the contributions of this paper is that we are able to solve for labor market equilibrium in this environment.

This is aided, in part, by an approach developed in our companion paper (Elsby and Gottfries 2021). As in that paper, the environment gives rise to an equilibrium in which turnover is stratified by a single state variable, the marginal product of labor. Firms are thus ordered according to a hierarchy of marginal products. At the bottom of the hierarchy is a *natural wastage region*, on the interior of which firms neither hire nor fire, and at the lower limit of which firms shed workers into unemployment. At the top of the hierarchy is an *expansion region* in which firms expand their employment by investing in new positions. Expanding employment increases output, but it also increases turnover by retarding the marginal product. Resolution of this tradeoff yields a closed-form solution for the quit rate and, thereby, the hierarchy of marginal products in the expansion region.

The key novel theoretical contribution of the present paper is the addition of an intermediate *replacement region* in which firms hire solely to replace workers that quit. Intuitively, firms in the replacement region are neither productive enough to seek to invest in new positions nor unproductive enough to choose to shrink. Instead, since they have sunk an investment into a given stock of positions, they seek to maintain their employment. The presence of the replacement region thus underlies the realization of net inaction in the model, as employment is fixed for as long as a firm remains in the region. A firm's labor demand and turnover are more fundamentally intertwined in the

replacement region, since the sole impetus to a firm's hires is precisely its endogenous quit rate. We show that it nevertheless is possible to solve for both labor demand and equilibrium turnover analytically and, thereby, deliver a solution for the endogenous hierarchy of firms and workers across marginal products. A key insight is that constancy of firm employment in the replacement region recovers an inverse-proportionality relation between equilibrium quit and vacancy-filling rates that echoes an analogous result in the canonical Burdett and Mortensen (1998) model.

A critical implication of the replacement region is the emergence of *vacancy chains*. We show that the model admits a clean characterization of *chain length*—defined as the expected number of job postings induced by an initial position. Chains are initiated by the creation of new positions in the expansion region, propagate recursively as positions are filled from firms lower down the hierarchy of marginal products in the replacement region, and end when a position is filled from either the natural wastage region, or the unemployment pool. Consequently, a sufficient statistic for chain length is the measure of searchers in the replacement region, relative to the natural wastage region and unemployment. This, in turn, is captured by a relevant ratio of vacancy-filling rates.

Taken together, the model provides a parsimonious account for the empirical incidence of replacement hiring, based on one additional parameter—the sunk cost of investment into new positions. It delivers analytical solutions for the rate of separation into the unemployment pool, as well as the rates of hires, quits, vacancy-filling and, thereby, the distributions of offers and workers, at each marginal product. These solutions in turn inform a characterization of aggregate steady-state labor market equilibrium. As in the canonical Diamond, Mortensen and Pissarides paradigm, this is determined by a *Beveridge curve* condition for flow balance in unemployment, and a *job creation* condition that summarizes aggregate labor demand.

We turn to a quantitative assessment of the model in section 3. We explore a calibration that targets standard estimates of labor market stocks and flows, as well as the magnitudes of gross hiring costs and the average wage gains to on-the-job search. Crucially, we discipline the sunk cost of new job creation using one of the key stylized facts of replacement hiring that we document in section 1—specifically, a measure of the annual incidence of inaction in *net* employment adjustments across establishments. In the remaining sections of the paper, we show that the calibrated model is able to reconcile a striking array of empirical features of both the cross-sectional and the time-series behavior of labor markets.

First, we begin by exploring the implications of the calibrated model for the remainder of our stylized facts of replacement hiring. As in the data, the incidence of net inaction decays slowly over the window of adjustment—much more slowly than geometric decay (as implied by many standard models). And, as in the data, this is so despite the presence of substantial intervening gross turnover. There is thus considerable replacement hiring in the model: around half of aggregate hires are due to replacement, close to the empirical analogue of approximately 45 percent. Relatedly, the average length of hiring chains—the expected number of hires generated by each new position—is around two. The calibrated model is able to do considerable justice to the stylized facts that motivated it.

Second, we confront the model with available data on its additional cross-sectional implications. Most prominently, we find that the model demands a large sunk cost of new positions to rationalize the stylized facts of replacement hiring—as much as thirty times larger than conventional estimates of gross hiring costs. A corollary is that the calibrated model naturally generates considerable cross-sectional dispersion in marginal products. This dovetails with a large literature on "misallocation" and its origins (see, for example, the recent survey by Hopenhayn 2014). We find that the model generates around three-quarters of the estimates of cross-establishment dispersion in labor productivity reported by Bartelsman, Haltiwanger and Scarpetta (2013), and much more than an analogous model that suspends the sunk costs of expanding employment. Strikingly, the same friction required by the model to replicate the degree of replacement hiring in the data also reconciles the large degree of cross-sectional variance in labor productivity.

Finally, in section 4, we explore the model's implications for two enduring puzzles of aggregate labor market dynamics, namely their volatility and persistence. We begin by showing that the presence of replacement hiring in the model greatly amplifies labor market volatility. Intuitively, replacement hiring alters the feedback of job creation decisions across firms. Conventional gross hiring costs induce *negative* feedback: Increased job creation by other firms raises quit rates and, thereby, turnover costs, *reducing* the desired hiring of a given firm. Replacement hiring, by contrast, moderates this effect: The rise in turnover induced by increased vacancy posting by other firms now *raises* the desired hiring of a given firm, as it seeks to replace positions vacated by quits. This, in turn, further tightens the labor market and amplifies the equilibrium responses of unemployment and job-finding rates to aggregate shocks.

Quantitatively, we find that the calibrated model implies volatilities of labor market stocks and flows close to their empirical counterparts. The model thus provides one

resolution of Shimer's (2005) volatility puzzle. Replacement hiring is central to the microeconomic origins of this result: Suspending the frictions that give rise to replacement hiring (as in Elsby and Gottfries 2021) cuts labor market volatility by around a half. An alternative perspective on the same result is that vacancy chains in the model become *shorter* in recessions, accounting for a substantial fraction of the cyclical amplitude of aggregate hires and, thereby, labor market outcomes more generally.

Turning to our final application, we find that the same ingredients play a crucial role in the persistence of labor market dynamics. Establishing this result requires more than the usual ingenuity, however. Inferring the transition dynamics requires solution for a sequence through time of the distributions of marginal products across offers and workers. This challenge is aggravated by the presence of replacement hiring, for the same reason that it complicates steady-state solution: firms' hires in the replacement region are determined solely by time-varying, endogenous quit functions. However, we show that the same insights that inform the model's steady-state solution also render solution of its out-of-steady-state dynamics feasible. Specifically, we are able to distill the solution algorithm down to an outer loop for the time path of just a single scalar, labor market tightness, and a simple inner fixed-point problem for the distributions of job values.

Using this approach, we solve for the transition dynamics induced by a permanent MIT shock to aggregate labor productivity. The presence of replacement hiring contributes considerably to sluggishness in labor market dynamics: The half-life of unemployment generated by the calibrated model is over nine months; in a (re)calibrated model that suspends replacement hiring, the half-life is little more than one month. Furthermore, the origins of this additional persistence can be traced to sluggishness in hiring. Intuitively, vacancy chains naturally inherit persistent dynamics: They propagate recursively through the replacement region, and terminate when a position is filled from either the natural wastage region or the unemployment pool. The probability of the latter, and thereby the length of vacancy chains, is determined by the shares of workers occupying these regions, which are slow-moving state variables. Consistent with this, we find that all of the persistence in aggregate hires generated by the model can be traced to the persistence of vacancy chains.

To assess the persistence generated by the model, we estimate the dynamic responses of unemployment and job-finding rates to unanticipated changes in output per worker in both model and data. We find that the model reproduces the substantial and long-lived empirical rise (fall) in the unemployment (job-finding) rate following a negative

innovation to output per worker. By contrast, eliminating replacement hiring in the model restores near-jump dynamics, lacking the persistence observed in the data. Vacancy chains thus further provide a resolution to the puzzle of the persistence of labor market dynamics.

In closing, section 5 explores extensions to our baseline approach. It begins by examining the robustness of the quantitative results we report. Due to the parsimony of the model, much of the content that we emphasize can be traced to one additional parameter—the sunk cost of a new position. Accordingly, we examine perturbations to this parameter in line with the range of alternative moments of replacement hiring that we document in section 1. We find that our baseline calibration best captures the constellation of outcomes that we highlight: the replacement share of aggregate hires, the dispersion of labor productivity across establishments, and the volatility and persistence of labor market outcomes. However, we nonetheless find that reasonable variations in the degree of replacement hiring still imply empirically reasonable outcomes on these dimensions. The results we emphasize are thus quite robust.

In a further extension, we examine the structure of wage determination in the model. Since tractability is at a premium, the baseline environment maintains a simple model of *ex post* bargaining without offer matching (based on Elsby and Gottfries 2021). Since the jury is still out on the empirical prevalence of offer matching, we further examine the implications of a generalization of the sequential auctions approach of Postel-Vinay and Robin (2002) to our environment with firm dynamics. Although this obscures a clear mapping between wage outcomes in model and data, we show that it is essentially innocuous for the quantitative implications that we highlight. Labor market equilibrium takes a similar form, and implications for labor market quantities are essentially preserved.

We conclude by offering thoughts on future work. Returning to our motivating themes, there is more work to be done to understand the origins of replacement hiring: Are the investments embodied in job creation associated with physical, organizational, or some other form of capital? The increasing availability of rich worker-firm matched microdata shows promise in this regard, facilitating both the direct measurement of vacancy chains, as well as their correlates. Our hope is that the present paper will stimulate future research along these lines.

Related literature. This paper provides a set of new stylized facts on the prominence of replacement hiring, and draws out their implications using a new model of firm dynamics,

(random) job search both off- and on-the-job, and vacancy chains. The view of the labor market that emerges dovetails with prior work along three themes.

The first relates to the empirical literature on establishment dynamics pioneered in the early work of Davis and Haltiwanger (1992). More recently, Davis, Faberman and Haltiwanger (2012) have noted the importance of quits in driving a wedge between job flows and worker flows at the establishment level. Similarly, Burgess, Lane and Stevens (2001) document that both expanding and contracting employers experience considerable "churn" of workers, a point echoed more recently by Lazear and Spletzer (2012). Our work further highlights the prominence of replacement hiring in establishments' responses to quits, and thereby the link between worker and job flows. More closely related to our empirical results is the work of Faberman and Nagypal (2008), who use JOLTS microdata to show that quits at an establishment are often followed by vacancy posting and gross hiring, indicating the presence of replacement hiring. Mercan and Schoefer (2020) find that similar results hold in German administrative microdata. Relative to this literature, we document new evidence that reinforces an impression of pervasive replacement—most notably, the persistent prevalence of inaction in net employment changes over many years, and the substantial gross turnover experienced among such firms.

A second strand of related work is a recent stream of papers that have extended search and matching models to accommodate a notion of firm size (Elsby and Michaels 2013; Acemoglu and Hawkins 2014; Kaas and Kircher 2015; Gavazza, Mongey and Violante 2018). A handful of papers has begun to incorporate on-the-job search into these environments. An early contribution by Lentz and Mortensen (2012) studies implications for the dispersion of steady-state productivity and wages. Schaal (2017) provides a related model of *directed* search that gives rise to an equilibrium with a "block-recursive" structure first articulated by Menzio and Shi (2011) in a model without firm dynamics. This removes the dependence of firm turnover on the distribution of job values, aiding a complete characterization of the steady state and aggregate dynamics of the model. Most recently, Bilal, Engbom, Mongey and Violante (2019) study a model with random search that further endogenizes firm exit and quantitatively matches rich dynamics of worker flows and employment dynamics, as well as entry and exit, over firm lifecycles. Finally, our companion paper (Elsby and Gottfries 2021) develops analytical methods for solving for the steady state and transition dynamics of a model of firm dynamics and (random)

on-the-job search. Our key focus here—the prominence of replacement hiring, its origins, and its establishment-level and aggregate implications—is not taken up in these works.[2]

A third strand of related literature comprises work that explicitly incorporates a notion of a vacancy chain. This concept has a rich heritage in mathematical sociology, pioneered in the early work of White (1970), with applications to topics as diverse as the turnover of hermit crabs across shells, and of clergy across churches (Chase 1991). Our model provides several contributions relative to White's early work. First, we endogenize the birth and death of chains as manifestations of idiosyncratic shocks to labor demand across firms. Second, we articulate the central role of sunk investments into positions in generating the fixity of labor demand that propagates vacancy chains. Third, our model gives rise to an endogenous hierarchy of marginal products along which the chain evolves, endogenizing chain length. And, finally, we embed these ingredients into aggregate labor market equilibrium, elucidating the role of vacancy chains in the amplification and propagation of labor market dynamics.

Within economics, the literature on vacancy chains is much smaller. Akerlof, Rose and Yellen (1988) use the idea to explain the procyclicality of job-to-job quits (see also Contini and Revelli 1997). However, theirs is a model in which jobs are rationed, and the rate of unemployment is exogenous. This severs the link between the frictions that give rise to vacancy chains and their effect on the aggregate labor market. By contrast, another strand of literature uses models of on-the-job search to study implications for labor market equilibrium, but assumes for tractability that all job offers are accepted. Fujita and Nakajima (2016) invoke this in a firm dynamics setting to study worker and job flows over the cycle. Mercan and Schoefer (2020) invoke the same assumption in an extension of the homogeneous one-worker-firm Diamond, Mortensen and Pissarides model to allow for long-lived jobs, on-the-job search and, thereby, replacement hiring. Because all jobs are accepted, their model does not feature an endogenous hierarchy, and they show that chain length takes a particularly simple form. Both Mercan and Schoefer's model and ours share the prediction that vacancy chains amplify labor market responses. Finally, subsequent to early versions of this paper, Carrillo-Tudela, Clymo, and Coles (2021) have

[2] A further strand of related work studies the interaction of on-the-job search with business cycles in models with *linear* technologies. In addition to Menzio and Shi (2011), prominent contributions include Moscarini and Postel-Vinay (2013), Coles and Mortensen (2016), and Lise and Robin (2017). More recently, Audoly (2019) and Gouin-Bonenfant (2022) study related models that incorporate firm lifecycles. Krause and Lubik (2006) find that procyclicality of on-the-job search intensity can generate realistic variation in worker flows. Mukoyama et al. (2018), however, find weak evidence for cyclical search intensity.

extended the model of Coles and Mortensen (2016) to incorporate richer firm dynamics, as well as replacement hiring. Mirroring our environment, firms in their model experience endogenous quits from on-the-job search, and positions are costly to create, generating replacement hiring. A key difference is that firms in their model operate a linear technology, so that labor demand is bounded solely by the frictions facing the firm, rather than by decreasing returns. Nonetheless, they show that their model is able to match the dynamics of worker flows and employment growth across firm size and the firm lifecycle.

Importantly, relative to all these papers, the analytical characterization of (marginal) values, the offer and worker distributions of job values, and the length of hiring and vacancy chains in a model with replacement hiring are novel to the present paper.

1. Stylized facts on replacement hiring

In this section, we use establishment-level microdata to document a set of stylized facts on the interplay between establishment-level (net) employment adjustment and gross worker turnover. These suggest a prominent empirical role for *replacement hiring*. These facts will motivate the remainder of the paper, which sets out a model that accommodates these facts and draws out their ramifications.

1.1 Data

We use restricted-access microdata from the Quarterly Census of Employment and Wages (QCEW) and the Job Openings and Labor Turnover Survey (JOLTS) for the United States. Both sources permit longitudinal linking of establishments over time, thereby allowing an analysis of establishment dynamics.

Quarterly Census of Employment and Wages. The QCEW covers approximately 98 percent of employees on non-farm payrolls in the United States and territories, and is a near-census of non-agricultural workers in private establishments. The data are collected by the Bureau of Labor Statistics (BLS) in concert with State Employment Security Agencies, which run state Unemployment Insurance (UI) programs and cover all employers with employees covered by UI. Each month, firms are required to submit a count of employment and a quarterly compensation bill, which the BLS aggregates to form the QCEW. The BLS then links establishments in the QCEW over time to create the Longitudinal Database of Establishments (LDE).

We have been granted access to QCEW/LDE microdata for a subset of forty states, including Washington, DC. (Data-sharing agreements have not been signed by Florida, Illinois, Massachusetts, Michigan, Mississippi, New Hampshire, New York, Oregon, Pennsylvania, Wisconsin, and Wyoming.) These microdata permit longitudinal linking of establishments over time from the early 1990s through the second quarter of 2014.

We further restrict our samples to privately owned establishments[3] and to continuing establishments with positive employment in consecutive quarters. Specifically, we construct a set of overlapping quarter-to-quarter balanced panels that exclude births and deaths of establishments *within* the quarter. Note that we do not balance across quarters, so births in a given panel will appear as incumbents in the subsequent panel (if they survive). This eliminates about 2 percent of establishments.[4] As an example of the sample sizes involved, in the second quarter of 2014 our samples cover about 5 million establishments and 77 million workers.

We use these samples to track quarterly *net* changes in establishment employment through time. The BLS defines monthly employment as the count of employees on an establishment's payroll for the pay period encompassing the 12th of each month.[5] We follow BLS procedure by focusing on quarterly data and defining quarterly employment as employment in the third month of each quarter. Thus, the net employment change in, for example, the first quarter of a given year is the difference between employment in March of that year and in December of the previous year.

Job Openings and Labor Turnover Survey. The JOLTS data cover approximately 16,000 establishments per month. The sample is constructed from two subsamples: A certainty panel of establishments that are always included, and a rotating panel of establishments that are sampled for 24 months. We use JOLTS microdata from December 2000 through the middle of 2016.

[3] We exclude establishments in public administration (NAICS industry 92) and those that are not in a classified industry (NAICS code 99). Excluding privately owned unclassified establishments eliminates approximately 225,000 employees (about 0.1 percent of total employment) in approximately 190,000 establishments (about 2 percent of total establishments) in the published, aggregate QCEW data. These restrictions are consistent with those imposed in related literature (for example, Foote 1998).

[4] We also restrict attention to establishments that are not flagged as being a successor or predecessor of another establishment between quarters, to be more confident in continuing-establishment linkages. This accounts for approximately 0.1 percent of establishments in the second quarter of 2014.

[5] The count of workers includes all those receiving any pay during the pay period, including part-time workers and those on paid leave.

Crucially for our purposes, the JOLTS samples include rich data on *gross* worker turnover, measuring hires and separations, and their composition into quits and layoffs (and other types of separations), at the establishment level. As in the QCEW, employment is measured for the pay period including the 12th of each month. Gross flows of workers in JOLTS are measured as flows that accrue over the course of a month. Hires are the total number of additions to the establishment's payrolls.[6] Separations are split into three broad categories based on the reason for termination. *Quits* are defined as voluntary separations initiated by the employee (excluding retirements). *Layoffs and discharges* are defined as involuntary separations due to cause or business conditions. *Other separations* are defined to include retirements, transfers, deaths, or separations due to disability. Total separations are the sum of all three components.

We apply two adjustments to the raw JOLTS data. First, all empirical results are weighted using the sample weights provided by the BLS. Second, in cases where an establishment's employment deviates from that implied by its hires and separations, we follow Davis, Faberman, and Haltiwanger (2013) by adjusting an establishment's employment to be consistent with its reported gross flows.

1.2 Inaction over net employment changes

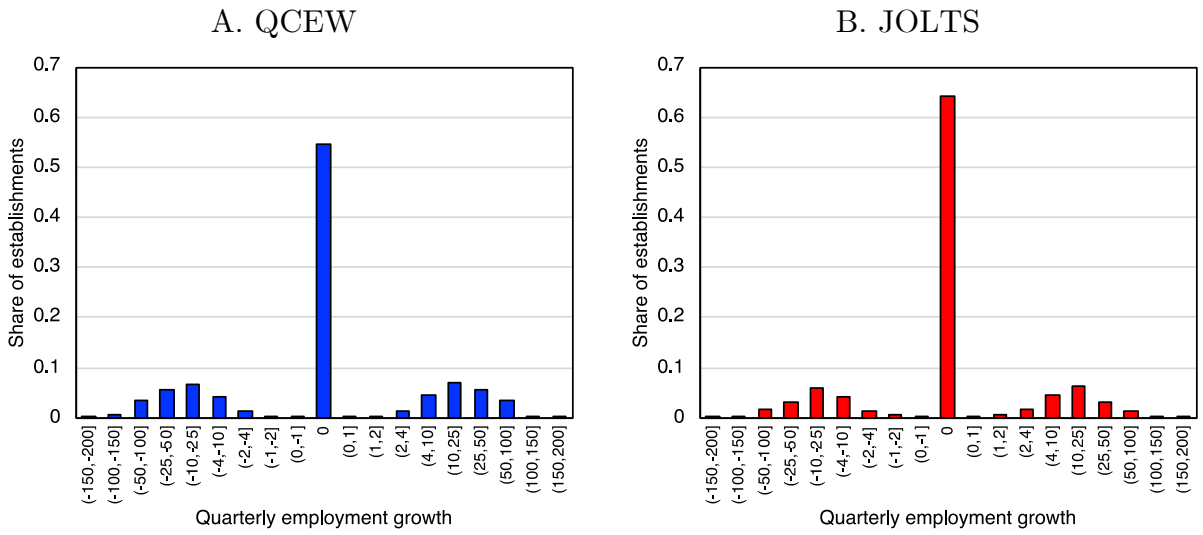
Our first fact is illustrated in Figure 1, which plots the distribution of quarterly employment growth at the establishment level using both the QCEW and JOLTS microdata.[7] This reiterates a long-recognized feature of establishment dynamics, namely that employment adjustment is marked by substantial *inaction* (Hamermesh 1989; Davis and Haltiwanger 1992). A large fraction of establishments—around 55 percent in the QCEW, and 65 percent in JOLTS—maintains the exact same employment level from one quarter to the next.

An underemphasized feature of this stylized fact, however, is that inaction is expressed over *net* changes in employment. This result stands in contrast to the implications of standard models of employment adjustment. Since the work of Oi (1962) and Nickell (1978), these models have stressed the role of costs to *gross* employment adjustments—that is, to hiring and firing workers. To the extent that such models generate inaction, it will be expressed at zero *gross* change in employment.

[6] These include both new hires and rehires, as well as part-time or full-time workers.

[7] Establishment growth is calculated as in Davis and Haltiwanger (1992).

Figure 1. Inaction over net employment changes

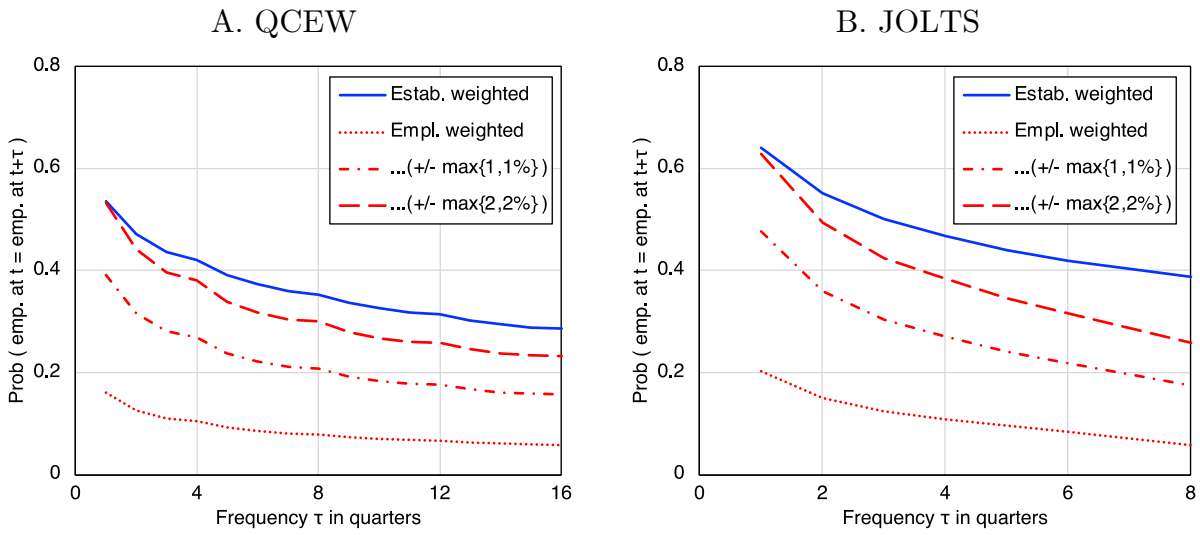


An important source of the wedge between net and gross changes is quits. Standard estimates suggest that the average rate of quits in the United States is substantial, on the order of 2 to 3 percent per month, according to employer reports in the JOLTS data (Davis et al. 2012), and job-to-job transitions in the Current Population Survey (Fallick and Fleischman 2004; Moscarini and Thomsson 2007). If such quits were evenly distributed across employers, standard models of gross adjustment costs would imply a mass point in the lower tail of the employment growth distribution, rather than at zero. Equivalently, it would imply that the mass of establishments reporting zero net change in employment had *replaced* a substantial fraction of their workforce over the quarter.

One simple explanation for the observed inaction over net changes is that, contrary to the preceding example, quits are not evenly distributed across establishments. It could be the case, for example, that establishments reporting zero change in employment are simply those "lucky" enough not to have experienced quits. However, among establishments in JOLTS that report no net change in employment, quit rates are on the order of 2.6 percent per quarter—lower than the average quarterly quit rate in JOLTS (of around 6 percent), but nonetheless substantial.

The combination of such nontrivial quit rates with observed inaction over net employment changes suggests that establishments frequently hire to replace exactly those workers that quit. We refer to this phenomenon as *replacement hiring*. In the remainder of this section, we explore several of its further implications.

Figure 2. The slow decay of inaction by frequency of adjustment



1.3 The slow decay of inaction by frequency of adjustment

We have established the prominence of net inaction, and its relation to the incidence of quits, at a quarterly frequency. We now explore these over longer horizons. Strikingly, our next fact suggests that net inaction is remarkably persistent, and that establishments maintain the same employment levels, often for years at a time.

To explore this, we utilize the panel dimension of the microdata, which allows one to track employment in continuing establishments over many quarters. Specifically, Figure 2 uses the QCEW and JOLTS microdata to plot the fraction of establishments that report the same employment level τ quarters ahead as a function of the frequency τ (blue solid line). Thus, for example, the 55-percent or so of establishments in the QCEW that report a zero net change in employment at a quarterly frequency in Figure 1A is replicated in the data point at $\tau = 1$ in Figure 2A.

Figure 2 reveals a striking result: net inaction rates decay very slowly by frequency τ. In the QCEW, the share of establishments reporting the same employment after one year ($\tau = 4$) is 42 percent, and is 35 percent after two years ($\tau = 8$). The analogous figures in the JOLTS microdata are slightly higher at 47 percent and 39 percent, respectively. The longer panel dimension of the QCEW further reveals that as many as 29 percent of establishments report the same level of employment as much as *four years* later. As a point of comparison, if establishments' rates of inaction were *independent* across frequency τ, inaction would decay geometrically with τ. The counterpart probabilities under this

hypothesis are essentially zero after just two years. The upshot is that U.S. establishments experience considerable rates of net employment inaction, often for years at a time.

This picture need not be representative of the average *employee's* establishment, however. Accordingly, Figure 2 also reports employment-weighted rates of inaction by frequency. These are naturally lower than their establishment-weighted counterparts, for the simple reason that larger establishments are less likely to report the *exact* same employment over time. However, much of this reduction can be traced to small employment changes. For example, while employment weighting reduces the one-quarter inaction rate closer to 20 percent, this rises to 40 percent or more once one includes small employment changes of one worker, or one percent of the workforce (whichever is larger). Widening the inaction window further to two workers, or up to two percent of the workforce, in turn raises estimated employment-weighted inaction rates back to the neighborhood of their establishment-weighted counterparts.

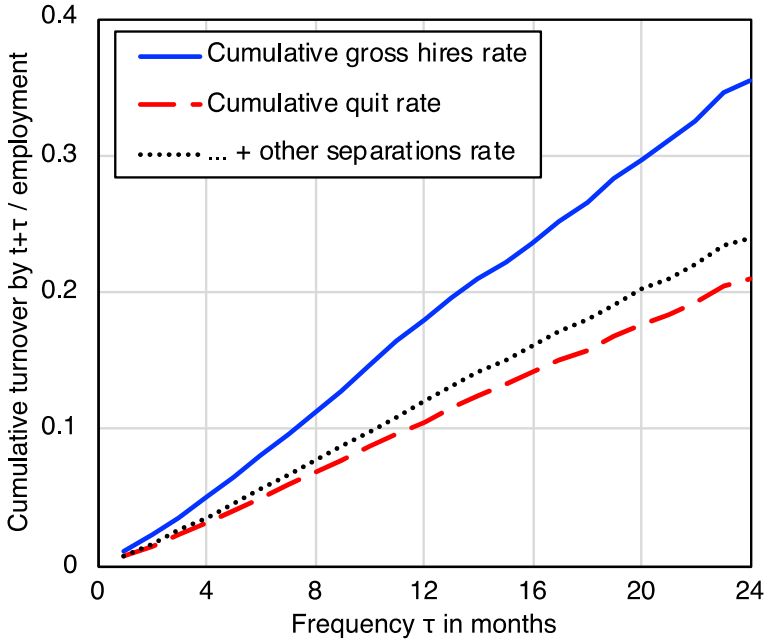
Importantly, the decay of estimated net inaction rates by frequency is only moderately more rapid after employment weighting. For example, four-year inaction probabilities in the QCEW remain at approximately 40 percent of their one-quarter counterparts for all of the inaction windows plotted in Figure 2. The striking persistence of establishment size suggests that employers have "reference" levels of employment which they maintain via replacement hiring.

1.4 Cumulative gross turnover in zero-growth establishments

Our third stylized fact returns to the question of how much gross turnover occurs at zero-growth establishments. We noted above that establishments that remain at the same employment level from one quarter to the next experience nontrivial quit rates, averaging 2.6 percent per quarter. We have also shown that net inaction is not merely prevalent at the quarterly frequency, but that establishments tend to maintain the same employment level for long periods, often years. Here, we explore whether these establishments that hold employment constant for long periods also experience substantial *cumulative* worker turnover, providing a sense for the magnitude of the intervening replacement hiring they implement to maintain their employment.

Specifically, in any given month of JOLTS microdata, we identify establishments that report the same employment level when surveyed τ months later. Among these establishments, we compute their cumulative rates of worker turnover over the course of

Figure 3. Cumulative gross turnover in zero-growth establishments (JOLTS)



the intervening τ months. Recalling that establishments included in the rotating panel element of the JOLTS sample are followed for 24 months, we implement this method for τs between one and 24 months.

Figure 3 reports the results of this exercise, pooled over all available months of JOLTS microdata. This reiterates the high-frequency results cited in our earlier discussion of Figure 1: Establishments reporting the same employment quarter-to-quarter also report gross hires (and, by definition, separations) equal to 3.6 percent of their workforce, of which 2.3 percent are reported as quits, and another 0.3 percent are reported as separations for other voluntary reasons.[8]

An important message of Figure 3, however, is that considerable gross worker turnover accumulates, almost linearly, in establishments with constant employment over longer frequencies. At a two-year horizon, over which nearly 40 percent of establishments report the same employment in the JOLTS microdata, gross hires in these establishments are replacing on average 35 percent of their workforce, around 25 percent of whom are

[8] We focus on quits and other separations since those are likely to be involuntary from the perspective of the employer. By contrast, some of the layoffs-cum-hires in Figure 3 may reflect recalls from temporary layoff. Estimates from Fujita and Moscarini (2017) imply that recalls account for around 20 percent of total hires, for example. And, of course, some separations categorized in layoffs and discharges are fires for cause, which may also be involuntary from the employer's perspective.

recorded as quits or other voluntary separations. Thus, the slow decay of net inaction depicted in Figure 3 occurs *despite* substantial gross worker turnover and is a further indication that many establishments engage in considerable degrees of replacement hiring.

1.5 Replacement hires are a large fraction of total hires

What fraction of *aggregate* hiring is accounted for by replacement hiring? We provide two perspectives on this question using the JOLTS microdata.

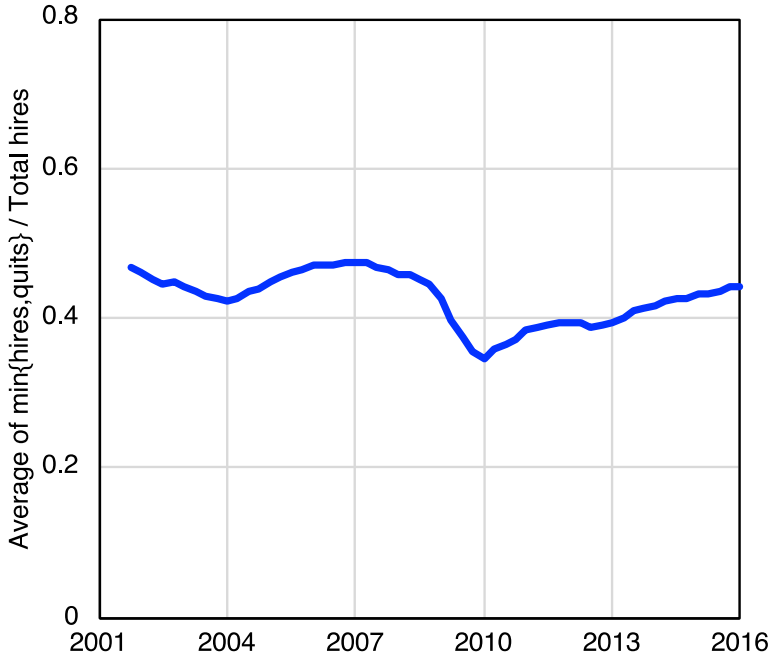
First, we consider a broader measure of replacements hires, defined as the minimum of an establishment's quits and its gross hires at a quarterly frequency. For instance, if an employer loses seven workers through quits in a quarter, but hires five, the number of replacement hires under this definition is five.[9] We then use this to compute aggregate replacement hiring as a fraction of aggregate hires. Figure 4 plots this replacement hiring rate for each quarter over the JOLTS sample period.

Figure 4 reveals that replacement hiring comprises a large fraction of total hiring by this measure, accounting for around 45 percent of all hires on average over the sample period. In addition, the ratio of replacement hires to total hires is procyclical, falling from a peak of nearly 48 percent in 2007 to close to 35 percent at its trough during the Great Recession.

A second perspective on the aggregate importance of replacement hiring returns to the stricter definitions explored earlier in this section. Specifically, we use the JOLTS microdata to compute the total number of hires accounted for by establishments that hold employment constant, for various inaction windows. Reiterating our earlier observation that large establishments are less likely to report the exact same employment over time, aggregate replacement hiring is modest by the strictest, literal definition of inaction, at just 7.5 percent of aggregate hires. As before, widening the inaction window to allow small changes greatly increases the estimated share of replacement hires, however. Hires among establishments that report net employment changes of less than one worker, or up to one percent of their workforce, account for over 25 percent of aggregate hires. Allowing employment changes of two workers, or up to two percent of the workforce, raises this further to nearly 40 percent of economy-wide hiring.

[9] This measure of replacement hires is related to those now reported in data from the U.S. Census Bureau. The Quarterly Workforce Indicators (QWI), a product of the LEHD program, defines replacement hires as the difference between gross hires at an establishment and its net employment growth.

Figure 4. A measure of replacement hiring as a fraction of total hires (JOLTS)



2. A model of vacancy chains

The striking persistence of establishment-level employment in the face of considerable gross worker turnover suggests that employers have reference levels of employment, to which they return routinely, often for years at a time, and do so via replacement hiring. In this section, we explore the economic implications of these stylized facts.

We argue that they call for a model with three ingredients: First, to map model outcomes to the preceding empirical results, it is necessary for the theory to accommodate multi-worker firms (or establishments). Second, to generate endogenous quits, and thereby provide an impetus to replacement hiring, the theory must incorporate on-the-job search, whereby employed workers contact, and sometimes transition to, alternative employers. Third, and most importantly, the theory must generate persistent reference levels of employment to which many firms seek to return when their workers quit. In what follows, we devise a model with these features and draw out its implications for cross-sectional labor market outcomes and aggregate labor market dynamics.

Several analytical challenges will arise as the model is developed. First, wage determination is complicated by the joint presence of multi-worker firms and of on-the-job search. Second, turnover is determined by the firm's position in an endogenous distribution of job values that both gives rise to firms' labor demand decisions and is

implied by aggregation of those same decisions. Equilibrium in steady state thus involves a fixed point in this distribution of job values and in a *sequence* of such distributions out of steady state. Our companion paper (Elsby and Gottfries 2021) develops an approach to address the first two challenges, which we review below. A third, and formidable, challenge is to solve for firms' choices of when to implement replacement hiring, which both depends on, and determines, turnover rates. A key contribution of this paper is to incorporate and solve the analytical challenge posed by replacement hiring.

2.1 Environment

Time is continuous, and the horizon infinite. The economy is populated by two sets of agents—firms and workers—that we now describe.

Firms. There is a unit measure of firms. Each firm employs a measure of workers, denoted n, to produce a flow of output, denoted y, according to an isoelastic production technology $y = xn^{\alpha}$, subject to decreasing returns, $\alpha \in (0,1)$. Idiosyncratic firm productivity x evolves according the geometric Brownian motion

$$\mathrm{d}x = \mu x \mathrm{d}t + \sigma x \mathrm{d}z, \tag{1}$$

where $\mathrm{d}z$ is the increment to a standard Wiener process. Idiosyncratic productivity x is the source of uncertainty to the firm, of shocks to its desired employment, and of *ex post* heterogeneity in productivity across firms.

Firms face frictions from two sources. The first is a conventional linear *gross* hiring cost, mirroring canonical models of firm dynamics (as in, for example, Bentolila and Bertola 1990, and Hopenhayn and Rogerson 1993). Specifically, denoting the firm's cumulative gross hires by H, and its flow increment over the time interval $\mathrm{d}t$ by $\mathrm{d}H$, the firm incurs a flow gross hiring cost of $c \cdot \mathrm{d}H$.

As we have emphasized, however, our innovation is to study a tractable environment that generates persistent reference levels of employment, endogenous quits and, thereby, replacement hiring. To accommodate this, we introduce a second friction that we refer to as a *net expansion cost*. Specifically, a *net* increment to the firm's total employment stock $\mathrm{d}n^{+}$ over the interval $\mathrm{d}t$ incurs an expansion cost equal to $C \cdot \mathrm{d}n^{+}$. The special case in which $C = 0$ corresponds to the environment studied in Elsby and Gottfries (2021).

An appealing interpretation of C dovetails with the notion of a *position* in the definition of a vacancy that motivated this paper. We noted that this definition evokes

the presence of a sunk investment that remains in place after the position is vacated. The net expansion cost C can be interpreted as the investment sunk into the creation of a new position.

This interpretation clarifies the novelty of the present environment. It is not the presence of net expansion costs *per se* that is central—indeed there are precedents for invoking such costs in prior work (see, for example, Cooper et al. 2007, 2015). Rather, it is their *interaction* with the presence of quits generated by on-the-job search that is novel. We will see that this interaction is central to a meaningful notion of replacement hiring, and thereby the emergence of vacancy chains in the model.

Turning to separations, these arise from two sources. First, each of the firm's n employees quits at rate δ. The determination of the quit rate δ is a crucial outcome of the model that we will return to below. Second, the firm can choose to shed additional employees—the increment of which we denote $\mathrm{d}S$—at zero cost. The law of motion for the firm's employment is thus

$$\mathrm{d}n = \mathrm{d}H - \mathrm{d}S - \delta n \mathrm{d}t. \tag{2}$$

Firms realize hires by posting vacancies. Given a vacancy-filling rate q, the firm posts the measure of vacancies v that fulfills its desired hires, $\mathrm{d}H = qv\mathrm{d}t$.[10] The aggregate measure of vacancies per firm is denoted V.

Workers. There is a unit measure of households, each of which comprises a measure L of workers. Each worker is risk neutral and is in one of two labor force states: A measure U of workers per household are unemployed, with the remaining measure $L - U$ employed. While unemployed, each worker receives a flow payoff b. While employed, they receive a wage w, the determination of which is described below.

Matching. Firms search for workers by posting vacancies. Workers search for jobs both while unemployed and while employed with relative search intensity s. The flow of contacts M is determined by a standard constant-returns-to-scale meeting function $M(U + s(L - U), V)$. Denoting labor market tightness by $\theta \equiv V/[U + s(L - U)]$, it follows that vacancies contact a searcher at rate $\chi(\theta) = M/V = M(1/\theta, 1)$, unemployed workers receive job offers at rate $\lambda(\theta) = M/[U + s(L - U)] = M(1, \theta)$, and employed workers

[10] This vacancy-posting policy is strictly optimal if, in addition to the frictions c and C described above, vacancy posting incurs an arbitrarily-small cost.

receive offers at rate $s\lambda(\theta)$. To economize on notation, we often will suppress dependence of the contact rates on tightness, except where necessary.

Wage setting. Since tractability is at a premium, we implement a wage protocol that is as simple as possible. Elsby and Gottfries (2021) provide such a model based on *ex post* wage bargaining without offer matching. It combines the insights of credible bargaining (Binmore et al. 1986) and multilateral bargaining (Bruegemann et al. 2018), in the presence of on-the-job search. We summarize that model briefly here. Later, in section 5, we provide a generalization of the sequential auctions model of Postel-Vinay and Robin (2002) to our multi-worker firm environment, and show that it preserves much of the structure and content of the solution described in the sections that follow. Since the bargaining model is particularly simple, and since it yields more natural implications for flow wages, for now we maintain that model as our baseline.[11]

Each $\mathrm{d}t$ period, after idiosyncratic productivity has been realized, and hires and separations resolved, the firm and its workers bargain over the flow wage, $w\mathrm{d}t$, according to the "Rolodex" game of Bruegemann et al. (2018). Workers in the firm are paid a common wage that corresponds to a marginal surplus sharing rule proposed by Stole and Zwiebel (1996). Threats of permanent severance are not credible (Binmore et al. 1986; Hall and Milgrom 2008), and so the relevant marginal surplus is the marginal *flow* surplus.

The result is a very simple wage equation. If, in the event of breakdown, the firm faces a flow cost of ω_f, and the worker a flow payoff of ω_e, the bargained wage is a simple affine function of the marginal product of labor,

$$w(n,x) = \frac{\beta}{1-\beta(1-\alpha)} x\alpha n^{\alpha-1} + \omega_0, \tag{3}$$

where $\beta \in (0,1)$ indexes worker bargaining power, and $\omega_0 \equiv \beta\omega_f + (1-\beta)\omega_e$. Wages rise in the marginal product, the payoffs to worker and firm from disruption and, due to the presence of decreasing returns, the inframarginal product.

Beyond its simplicity, this approach to wage setting has additional advantages. Since wages are continually renegotiated, the bargain over the current *flow* wage has a vanishingly small effect on the *present value* of the employment relationship to the worker,

[11] Possible motivations for the absence of offer matching include lack of verifiability of outside offers, and equal treatment constraints within the firm (see Mortensen 2003, section 5.1). Available empirical evidence on offer matching remains limited, but suggests only a modest propensity (see Brown and Medoff 1996; Bewley 1999; and Di Addario et al. 2020).

and thereby on workers' quit decisions. The protocol is thus not subject to Shimer's (2006) concern that the effects of wages on turnover can induce a nonconvexity in the bargaining set (Nagypal 2007; Gottfries 2019). In addition, the wage protocol provides a key source of tractability for the determination of turnover: Since all workers within a given firm receive a common wage, workers value job offers from each firm according to its *worker surplus*—the value it delivers in excess of unemployment—which we denote W. Turnover thus flows from low- to high-worker-surplus firms.

Turnover. This last implication can be used to simplify the determination of turnover, as embodied in the quit rate δ, and the vacancy-filling rate q, faced by each firm. Specifically, Elsby and Gottfries (2021) show how it gives rise to an m-*solution* in which turnover is stratified by a single state variable, the marginal product of labor $m \equiv x\alpha n^{\alpha-1}$. Under an m-solution, they show that the worker surplus W is uniquely determined by, and monotonically increasing in, the marginal product m. Worker turnover thus flows from low- to high-*marginal-product* firms, a considerable simplification.

Consider the quit rate δ. At rate $s\lambda$, each of a firm's employees is confronted with an outside offer. Under an m-solution, each contacted employee quits if the outside offer delivers a higher marginal product than the current firm. The quit rate faced by a firm with marginal product m is thus given by

$$\delta(m) = s\lambda[1 - F(m)], \tag{4}$$

where $F(\cdot)$ is the distribution function of marginal products among *job offers*.

Now consider the vacancy-filling rate q. At rate χ, each of a firm's vacancies contacts a searcher. With probability $\psi \equiv U/[U + s(L - U)]$, a contacted searcher is unemployed, and the vacancy is filled with certainty. (It is never optimal to post a vacancy unattractive to an unemployed searcher.) With probability $1 - \psi$, the contacted searcher is employed and is hired if the firm delivers a higher marginal product than the worker's existing firm. The vacancy-filling rate faced by a firm with marginal product m is thus given by

$$q(m) = \chi[\psi + (1 - \psi)G(m)], \tag{5}$$

where $G(\cdot)$ is the distribution function of marginal products among *employees*.

A fundamental implication of the environment is that the turnover rates in (4) and (5) are endogenous equilibrium outcomes. The distributions F and G both shape, and are shaped by, firms' decisions, due to the interaction of firm dynamics and on-the-job search. In what follows, we thus solve jointly for optimal labor demand and equilibrium turnover.

2.2 The firm's problem

We are now in a position to state the problem facing a firm. Given the preceding environment, the value of the firm Π satisfies

$$r\Pi(n,x)\mathrm{d}t = \max_{\mathrm{d}S\geq 0,\mathrm{d}H\geq 0}\left\{\left[xn^{\alpha} - wn - \delta n\Pi_n + \mu x\Pi_x + \frac{1}{2}\sigma^2 x^2\Pi_{xx}\right]\mathrm{d}t - \Pi_n \mathrm{d}S - (c-\Pi_n)\mathrm{d}H - C\mathrm{d}n^{+}\right\}, \tag{6}$$

where the firm's employment evolves according to (2), the flow of hires is delivered by posting the requisite vacancies, $\mathrm{d}H = qv\mathrm{d}t$, the wage w is given by (3), and the quit rate δ, and vacancy-filling rate q, respectively take the forms in (4) and (5).

The Bellman equation in (6) comprises the following. The firm receives the flow product xn^{α}, and pays each of its n employees a flow wage w. A flow δn of its employees quit, each of which is valued by the firm on the margin at Π_n. The firm's productivity x evolves according to the stochastic law of motion (1). By Ito's Lemma, the drift of productivity is valued on the margin at Π_x, and its instantaneous variance is valued in proportion to Π_{xx}. Each incremental separation forgoes the marginal value of labor Π_n. Each incremental hire generates value Π_n, but incurs cost c if the firm does not expand ($\mathrm{d}n^{+} = 0$), and $c + C$ if the firm expands its positions ($\mathrm{d}n^{+} > 0$). Given this, the firm chooses its flow of separations $\mathrm{d}S$ and hires $\mathrm{d}H$ to maximize its value.

Observe that positions in the model have a "use-it-or-lose-it" property: Positions that remain unfilled over a $\mathrm{d}t$ period lapse. This has the tractable implication that the firm's problem has just one endogenous state variable, employment n. The cost is that it rules out the realization of *delays* between vacation of a position, and replacement in the model. Empirically, though, vacancy-filling rates are high, with vacancy durations often a matter of weeks, or days, depending on the sector (Davis et al. 2013).[12] The gain in empirical content from accommodating delays in replacement in the model is thus limited relative to the significant gain in tractability afforded by our "use-it-or-lose-it" abstraction.

Returning to the firm's problem, note that optimal hires and separations satisfy (see Harrison and Taksar 1983)

[12] A counterpoint to this is that a firm with relatively low productivity may choose not to post a vacancy, but retain the position. However, we suspect that the longer a position remains vacant, the more difficult it becomes to insert a new worker into the position (at cost c) and resume production. We interpret C to include the costs of reactivating dormant positions and reintegrating them into the work flow of the firm.

$$\Pi_n \mathrm{d}S^* = 0, \quad (\Pi_n - c)(\mathrm{d}H^* - \delta n \mathrm{d}t)^- \mathrm{d}H^* = 0, \text{ and, } (\Pi_n - c - C)(\mathrm{d}H^* - \delta n \mathrm{d}t)^+ = 0. \quad (7)$$

The marginal value of a position Π_n is respectively set equal to: zero in the event of firing, $\mathrm{d}S^* > 0$; the gross hiring cost c in the event of partial replacement, $\mathrm{d}H^* \in (0, \delta n \mathrm{d}t)$; and the sum of the gross hiring cost and the expansion cost, $c + C$, in the event of expansion, $\mathrm{d}H^* > \delta n \mathrm{d}t$. Importantly for the results to come, note that there is replacement hiring, $\mathrm{d}H^* = \delta n \mathrm{d}t$, whenever $\Pi_n \in (c, c + C)$.

The maximized value of the firm therefore is given by

$$r\Pi(n, x) = xn^\alpha - wn - \delta n \min\{\Pi_n, c\} + \mu x \Pi_x + \frac{1}{2}\sigma^2 x^2 \Pi_{xx}. \quad (8)$$

The presence of on-the-job search implies that firms face the *turnover costs* $\delta n \min\{\Pi_n, c\}$. Each of the firm's n employees quits at rate δ. Among non-hiring firms, with $\Pi_n < c$, each quit is valued on the margin at Π_n. Among hiring firms, with $\Pi_n \geq c$, each quit is valued according to the *replacement cost*, equal to the gross hiring cost c.

Returning to the conditions for optimal hires and separations in (7), it remains to characterize the marginal value of labor. Recall that, under an m-solution, the quit rate is a function solely of the marginal product, $\delta(m)$ in (4). This in turn implies that the marginal value of labor also can be written as a function solely of the marginal product, $\Pi_n(n, x) \equiv J(m)$. Differentiating (8), and recalling the wage equation (3), yields

$$\begin{aligned} rJ(m) = (1 - \omega_1)m - \omega_0 - [\delta(m) - (1 - \alpha)m\delta'(m)] \min\{J(m), c\} \\ + \left[\mu + (1 - \alpha)\delta(m)\mathbf{1}_{\{\mathrm{d}n^* < 0\}}\right] mJ'(m) + \frac{1}{2}\sigma^2 m^2 J''(m), \end{aligned} \quad (9)$$

where $1 - \omega_1 \equiv (1 - \beta)/[1 - \beta(1 - \alpha)]$ is the firm's share of the marginal product. Optimality conditions for hires and separations provide boundary conditions for (9). We show that these are solved by a labor demand policy characterized by three[13] regions.

First, as in standard models of firm dynamics, there is a region of inaction $m \in (m_l, m_h)$ in which the marginal value of labor J lies in the interval $(0, c)$. Firms neither hire nor fire in this region, and their employment consequently decays as employees quit. For this reason, we refer to it as the *natural wastage region*. At its lower limit is a *layoff*

[13] As suggested by equation (7), it is possible for a fourth *partial replacement* region to exist, in which firms optimally replace only a fraction of their quits. Under plausible parameter values, however, we find that this region is degenerate. For simplicity, we present the simpler three-region case in the main text, and characterize the partial replacement region in Lemma 4 in the Appendix.

boundary m_l at which $J(m_l) = 0$, and firms shed workers into unemployment. At its upper limit is a *hiring boundary* m_h at which $J(m_h) = c$, and firms begin to hire.

Second, there is a *replacement region* in which firms hire to replace their quits, $\mathrm{d}H^* = \delta(m)n\mathrm{d}t$, for all $m \in (m_h, m_e)$. The presence of this region is the key innovation of the model that allows it to accommodate the evidence on replacement hiring in section 1. Accordingly, it will be the focus of the majority of attention in what follows. In the replacement region, the marginal value of labor J lies in the interval $(c, c + C)$—sufficient to induce gross hiring, but insufficient to induce net expansion. At its upper limit is the *expansion boundary* m_e at which $J(m_e) = c + C$, and firms begin to expand employment.

Finally, there is an *expansion region* in which optimal hires exceed quits, and the firm expands its employment, for all $m \in (m_e, m_u)$. Here, the marginal value of labor J is equal to a constant given by the sum of the gross hiring and expansion costs, $c + C$, and is supported by a quit rate that falls in m at an appropriate rate. The expansion region is the analogue of the hiring region highlighted by Elsby and Gottfries (2021). As they note, this region is degenerate in standard theories of firm dynamics: the firm hires until the marginal value of labor is brought down to the marginal cost ($c + C$ in this context). The fact that this region is nondegenerate here is a novel implication of the interaction of on-the-job search with firm dynamics. Its *upper boundary* is denoted m_u.

In what follows, we provide for each of these regions analytical solutions for optimal labor demand, the marginal value of the firm J, as well as the equilibrium layoff rate ς, quit rate δ, gross hiring rate η, and vacancy-filling rate q. The latter are based on aggregation results that are especially simple when aggregate labor demand is stationary. Applying Ito's Lemma, this is ensured by the following condition on the instantaneous drift and variance,

$$\mu + \frac{1}{2}\sigma^2 \frac{\alpha}{1-\alpha} = 0. \tag{10}$$

This assumption is made solely in the interest of simplicity, by abstracting from growth, and is maintained throughout the following analyses.[14] We now characterize each of these three regions, and show how they lead to the emergence of vacancy chains in the model.

[14] Analogously, one could study a balanced-growth environment by letting c, C, and ω_0 grow at the rate of aggregate productivity.

2.3 Chain creation and destruction

The structure of labor demand implies that new chains are formed in the expansion region where firms grow by creating new positions, and end when a worker is hired from the natural wastage region, or unemployment. We begin by describing this process of creation and destruction of chains by characterizing the natural wastage and expansion regions. We will see that these two regions share a convenient analytical property that aids solution for labor demand and turnover.

Natural wastage region. We begin with the natural wastage region, $m \in (m_l, m_h)$. Since turnover flows from low- to high-m firms, and since all hiring occurs at firms with marginal products of m_h or higher, the quit rate is maximal and invariant in the natural wastage region: $\delta(m) = s\lambda$, and $\delta'(m) = 0$, for all $m \in (m_l, m_h)$. As emphasized in Elsby and Gottfries (2021), the recursion for the marginal value of labor $J(m)$ in (9) can thus be decoupled from the quit rate $\delta(m)$,

$$(r + s\lambda)J(m) = (1 - \omega_1)m - \omega_0 + [\mu + (1 - \alpha)s\lambda]mJ'(m) + \frac{1}{2}\sigma^2 m^2 J''(m). \tag{11}$$

This is a canonical firm dynamics problem (Bentolila and Bertola 1990; Abel and Eberly 1996). Smooth-pasting and super-contact conditions for optimal hires and fires provide boundary conditions for (11) and, thereby, a solution for the marginal value $J(m)$. In turn, solutions for layoff and vacancy-filling rates can be recovered from the Fokker-Planck (Kolmogorov Forward) equation for the flow of workers across marginal products. The following Lemma summarizes.

Lemma 1 *In the natural wastage region, (i) the firm's marginal value is given by*

$$J(m) = \frac{(1 - \omega_1)m}{\rho(1)} - \frac{\omega_0}{\rho(0)} + J_1 m^{\gamma_1} + J_2 m^{\gamma_2}. \tag{12}$$

The coefficients J_1 and J_2, and the boundaries m_l and m_h, are known implicit functions (provided in the appendix) of the parameters of the firm's problem; $\gamma_1 < 0$ and $\gamma_2 > 1$ are the roots of the fundamental quadratic,

$$\rho(\gamma) = -\frac{1}{2}\sigma^2\gamma^2 - \left[\mu - \frac{1}{2}\sigma^2 + (1 - \alpha)s\lambda\right]\gamma + r + s\lambda = 0. \tag{13}$$

(ii) The separation rate into unemployment, quit rate, and gross hiring rate are given by

$$\varsigma = \frac{\sigma^2/2}{1-\alpha} m_l g(m_l),\ \delta(m) = s\lambda,\ \textit{and},\ \eta(m) = 0. \tag{14}$$

(iii) The vacancy-filling rate is given by

$$q(m) = \chi\psi\left(\frac{m}{m_l}\right)^{\frac{1-\alpha}{\sigma^2/2}s\lambda}. \tag{15}$$

Lemma 1 reiterates standard results. The marginal value of labor $J(m)$ comprises affine terms that capture the marginal value of the firm's current workforce in perpetuity, and nonlinear terms that capture the values of the options to hire workers in favorable future states and to fire them in adverse future states. Separations into unemployment arise at a lower reflecting layoff boundary m_l. There, a density $g(m_l)$ of workers receives shocks to their log marginal product of instantaneous variance σ^2. Negative innovations induce firms to shed employees until their marginal product is reflected back to m_l, at a rate determined by the elasticity of labor demand $1/(1-\alpha)$. Finally, constancy of the quit rate $\delta(m) = s\lambda$ and a zero gross hiring rate $\eta(m) = 0$ imply that the marginal product evolves according to a geometric Brownian motion,

$$\mathrm{d}m = [\mu + (1-\alpha)s\lambda]m\mathrm{d}t + \sigma m\mathrm{d}z, \text{ for all } m \in (m_l, m_h). \tag{16}$$

A standard implication is that the implied stationary worker distribution $G(m)$, and thereby the vacancy-filling rate $q(m)$, take the form of a power law.

Expansion region. Now consider the expansion region comprising firms with marginal products $m \in (m_e, m_u)$ that choose to expand their employment. This case echoes results identified in Elsby and Gottfries (2021). In a simpler environment, they show that a decoupling of the marginal value $J(m)$ and quit rate $\delta(m)$, analogous to that in the natural wastage region, holds for hiring firms. A similar logic applies in the present environment for expanding firms.

Specifically, recall that the optimality condition in (7) stipulates that the marginal value of an expanding firm be set equal to the sum of the gross hiring and net expansion costs. It follows that $J(m) = c + C$, and thus $J'(m) = 0 = J''(m)$, for all $m \in (m_e, m_u)$. Inserting these into the recursion for the marginal value in (9) yields a simple differential equation for the quit rate $\delta(m)$ that is decoupled from the marginal value $J(m)$,

$$r(c + C) = (1-\omega_1)m - \omega_0 - [\delta(m) - (1-\alpha)m\delta'(m)]c. \tag{17}$$

Boundary conditions for the latter are provided by the solution for $\delta(m_e)$ from the replacement region (to be provided shortly), and by the fact that the quit rate falls to zero at the upper boundary $\delta(m_u) = 0$. The implied solution for the equilibrium quit rate in turn provides solutions for hiring and vacancy-filling rates in the expansion region. The following lemma summarizes.

Lemma 2 *In the expansion region, (i) the firm's marginal value is given by* $\mathrm{J}(\mathrm{m}) = \mathrm{c} + \mathrm{C}$. *(ii) The quit rate is given by*

$$\delta(m) = \delta(m_e) + \frac{1}{c}\left\{\frac{(1-\omega_1)(m-m_e)}{\alpha} - \left[\frac{(1-\omega_1)m_e}{\alpha} - \omega_0 - [r+\delta(m_e)]c - rC\right]\left[\left(\frac{m}{m_e}\right)^{\frac{1}{1-\alpha}} - 1\right]\right\}. \tag{18}$$

(iii) The gross hiring rate is given by

$$\eta(m) = -\frac{\sigma^2/2}{1-\alpha}\frac{m\delta'(m)}{\delta(m)}. \tag{19}$$

(iv) The vacancy-filling rate is given by

$$q(m) = q(m_e)\exp\left[\frac{1-\alpha}{\sigma^2/2}\int_{m_e}^{m}\frac{\delta(\tilde{m})}{\tilde{m}}\mathrm{d}\tilde{m}\right]. \tag{20}$$

An expanding firm faces a subtle tradeoff. On the one hand, it can create additional positions at a marginal cost C, and hire workers into them at a further marginal cost of c, generating a flow of output of $(1-\omega_1)m - \omega_0$ in (17). On the other hand, this generates costs of replacement, $[\delta(m) - (1-\alpha)m\delta'(m)]c$ in (17): It raises the measure of workers needing to be replaced, at a marginal cost of $\delta(m)c$; and, due to decreasing returns, it retards the firm's marginal product m, raising the firm's quit rate and inducing a marginal replacement cost of $-(1-\alpha)m\delta'(m)c > 0$. The equilibrium quit rate $\delta(m)$ in (18) declines with the marginal product at a rate that just balances these forces. Put more simply, firms can expand employment by two means—recruitment and retention. Equation (17) stipulates that, in equilibrium, firms must be indifferent between these, yielding the solution for the quit rate in (18).

This in turn provides the key to solutions for the equilibrium hiring rate $\eta(m)$ in (19), and the equilibrium vacancy-filling rate $q(m)$ in (20). Interestingly, these are related to the quit rate $\delta(m)$ according to a hierarchy of elasticities: The hiring rate is proportional

to (minus) the elasticity of the quit rate in (19); and the quit rate is proportional to the elasticity of the vacancy-filling rate, which can be seen by differentiating (20) to obtain

$$\delta(m) = \frac{\sigma^2/2}{1-\alpha}\frac{mq'(m)}{q(m)}. \tag{21}$$

The structure of these solutions reflects the fact that turnover is stratified by the marginal product, flowing from low-m firms to high-m firms. Accordingly, vacancies filled at some marginal product m reflect quits from all *lower* ms, as in (20). Similarly, quits at a given m become hires at all *higher* ms, as can be confirmed by integrating (19).

The upshot is that, since the quit rate $\delta(m)$ is strictly decreasing and concave and reaches zero at the upper boundary m_u, the hiring rate $\eta(m)$ is strictly increasing and asymptotes to infinity as the marginal product approaches m_u. As a consequence, the rate of net employment growth at a given marginal product m, $\eta(m) - \delta(m)$, is strictly increasing in the expansion region, such that the marginal product obeys a stochastic law of motion with endogenous gradual mean reversion,

$$\mathrm{d}m = \{\mu + [\eta(m) - \delta(m)]\}m\mathrm{d}t + \sigma m\mathrm{d}z. \tag{22}$$

Intuitively, positive innovations to the marginal product induce an increased rate of hiring and a decreased rate of turnover, so that the marginal product reverts downward in expectation. These forces in turn give rise to a thinning of the tail of the stationary worker distribution $G(m)$ implied by (20) relative to the power laws that emerge in the preceding regions. Indeed, in the limit as m approaches m_u, infinite mean reversion implies that there can be no density of workers at the upper boundary, $g(m_u) = 0$.

2.4 Chain propagation

The novel implication of our environment, and the key focus of our attention, is the presence of a *replacement region*, $m \in (m_h, m_e)$, in which firms hire to replace their quits. The replacement region is the means by which chains are propagated in the model. Mirroring our motivating intuition, this region emerges from the presence of a sunk investment into *positions*, as captured by the unit expansion cost C. Firms with a marginal value J in excess of the gross hiring cost c, but less than the sum of gross hiring and expansion costs $c + C$, maintain their existing employment stock, $\mathrm{d}n^* = 0$, and do so via replacement hiring, $\mathrm{d}H^* = \delta n\mathrm{d}t$.

The distinctive nature of the replacement region brings with it a distinctive analytical challenge, however. We have seen that solution of the natural wastage and expansion

regions is simplified by a convenient decoupling of the solutions for the marginal value $J(m)$ and the quit rate $\delta(m)$. By contrast, in the replacement region, the evolutions of the marginal value and the quit rate are fundamentally intertwined. To see how, note that the recursion for the marginal value of labor in (9) now takes the form

$$rJ(m) = (1-\omega_1)m - \omega_0 - [\delta(m) - (1-\alpha)m\delta'(m)]c + \mu m J'(m) + \frac{1}{2}\sigma^2 m^2 J''(m), \quad (23)$$

for all $m \in (m_h, m_e)$. Intuitively, the firm's marginal product m determines not only its marginal flow payoff, $(1-\omega_1)m - \omega_0$, but also its quit rate $\delta(m)$, and thereby its marginal replacement costs, $[\delta(m) - (1-\alpha)m\delta'(m)]c$. Optimal labor demand and equilibrium turnover must therefore be solved jointly in the replacement region.

An important analytical contribution of this paper is to show that it is nonetheless possible to characterize model outcomes in the replacement region. Proposition 1 summarizes.

Proposition 1 *In the replacement region, (i) the firm's marginal value is given by*

$$J(m) = \frac{(1-\omega_1)m}{\varrho(1)} - \frac{\omega_0}{\varrho(0)} - \mathcal{J}_0(m) + \mathcal{J}_1 m^{\tilde{\gamma}_1} + \mathcal{J}_2 m^{\tilde{\gamma}_2}, \quad (24)$$

where

$$\mathcal{J}_0(m) = c\frac{1-\alpha}{\sigma^2/2}\int_{m_h}^{m}\left[\varpi\left(\frac{m}{\tilde{m}}\right)^{\tilde{\gamma}_1} + (1-\varpi)\left(\frac{m}{\tilde{m}}\right)^{\tilde{\gamma}_2}\right]\frac{\delta(\tilde{m})}{\tilde{m}}\mathrm{d}\tilde{m}, \quad (25)$$

and is strictly increasing in m*, and* $\varpi \equiv [1/(1-\alpha) - \tilde{\gamma}_1]/(\tilde{\gamma}_2 - \tilde{\gamma}_1) \in (0,1)$ *is a weight. The coefficients* $\mathcal{J}_1$ *and* $\mathcal{J}_2$*, and the boundaries* m_h *and* m_e*, are known implicit functions (provided in the appendix) of the parameters of the firm's problem;* $\tilde{\gamma}_1 < 0$ *and* $\tilde{\gamma}_2 > 1$ *are the roots of the fundamental quadratic*

$$\varrho(\tilde{\gamma}) = -\frac{1}{2}\sigma^2\tilde{\gamma}^2 - \left(\mu - \frac{1}{2}\sigma^2\right)\tilde{\gamma} + r = 0. \quad (26)$$

(ii) The quit rate and the gross hiring rate are equal, and are given by

$$\delta(m) = s\lambda\left[1 + \frac{1-\alpha}{\sigma^2/2}s\lambda\ln\left(\frac{m}{m_h}\right)\right]^{-1} = \eta(m). \quad (27)$$

(iii) The vacancy-filling rate is given by

$$q(m) = q(m_h)\frac{s\lambda}{\delta(m)}. \quad (28)$$

Several aspects of Proposition 1 are novel. Consider first (24), which provides a solution for the marginal value of labor $J(m)$ for any given quit rate $\delta(m)$. The expression in (24) differs from its counterpart in the natural wastage region (12) in two respects. First, the terms that capture the marginal value of the current workforce and the option values of future employment adjustment are subject to a different effective discount rate, captured by the difference between $\varrho(\cdot)$ in (26), and $\rho(\cdot)$ in (13). Since firms replace quits in the replacement region, there is no longer drift in the marginal product induced by natural wastage, and future flows of value are no longer additionally discounted by $s\lambda$.

More fundamentally, a second difference is that the marginal value in the replacement region includes an additional term, $\mathcal{J}_0(m)$, that reflects the expected discounted value of the firm's marginal replacement costs, which rise with the marginal product m. Although a higher m is associated with a lower rate of turnover $\delta(m)$, this is dominated by an opposing force whereby firms with higher ms expect to continue replacing quits, and incurring replacement costs, over longer durations.

To complete the solution to the firm's problem, Proposition 1 provides further insights into equilibrium turnover. Recall that, in the replacement hiring region, one cannot solve for optimal labor demand and equilibrium turnover in isolation: they cannot be decoupled. Nevertheless, by using the fact that hiring and quit rates are equal in this region, and applying the Fokker-Planck (Kolmogorov Forward) equation to characterize the flow of workers across marginal products m, one can solve for the equilibrium quit and vacancy-filling rates in (27) and (28) and, using (4) and (5), the equilibrium offer and worker distributions, $F(m)$ and $G(m)$.

To see how, note that a first consequence of hires being equal to separations in the replacement region is that the flow of workers across marginal products has zero drift. Workers thus diffuse randomly across (log) marginal products and, in steady state, thereby are distributed *uniformly* over $\ln m$ in the replacement region,

$$G(m) \propto \ln m + \text{constant, for all } m \in (m_h, m_e). \tag{29}$$

Equivalently, the worker density obeys a simple power law, $g(m) \propto m^{-1}$. The latter can be confirmed by using (5), and combining (27) and (28).

A second corollary of hires being equal to separations in the replacement region is that the vacancy-filling rate $q(m)$ must be inversely proportional to the quit rate $\delta(m)$. This result has a direct analogue in Burdett and Mortensen's (1998) classic model of wage posting with on-the-job search. There, the same relation holds, but with jobs stratified by

an endogenous distribution of *wages*, as opposed to marginal products. The key intuition here is that, mirroring Burdett and Mortensen (1998), firm employment is held constant in the replacement region, and thus recruitment must exactly offset quits. The inverse proportionality of $q(m)$ and $\delta(m)$ follows. Combining with (29), and using (5), delivers the solution for the quit rate stated in (27), and completes the solution provided by Proposition 1.

We have noted that the implied form of the worker distribution $G(m)$ is log-uniform. Proposition 1 likewise implies a form for the offer distribution $F(m)$. It turns out that, among offers, the log marginal product assumes a generalized Pareto distribution with location $\ln m_h$, scale $(\sigma^2/2)/[(1-\alpha)s\lambda]$, and shape equal to one. Specifically, one can use (27) and (4) to write

$$F(m) = \frac{\frac{1-\alpha}{\sigma^2/2} s\lambda \ln\left(\frac{m}{m_h}\right)}{1 + \frac{1-\alpha}{\sigma^2/2} s\lambda \ln\left(\frac{m}{m_h}\right)}, \text{ for all } m \in (m_h, m_e). \tag{30}$$

To conclude this subsection, we return to the idea that the replacement region is critical in propagating vacancy chains in the model. We can now demonstrate this point more formally.

White (1970) underscored a defining characteristic of a vacancy chain—that it has *length*, defined as the number of times an initial vacancy is (re)posted. Interestingly, the present model admits a particularly clean characterization of chain length. Proposition 2 considers two interpretations: *Hiring chains*, with length ℓ_H summarizing the expected number of hires generated by an initial hire; and *vacancy chains*, with length ℓ_V summarizing the expected number of vacancies induced by an initial vacancy.

Proposition 2 *The expected remaining lengths of hiring and vacancy chains at* m *are given respectively by*

$$\ell_H(m) = 1 + \ln\frac{q(m)}{q(m_h)}, \text{ and } \ell_V(m) = \frac{q(m)}{q(m_h)}, \text{ for all } m \in (m_h, m_u). \tag{31}$$

Chain length in the model is thus intimately related to the vacancy-filling rate $q(m)$. Chains propagate recursively when positions are filled from lower-m firms in the replacement region. The cumulative sum of these replacement events is summarized by the ratio of $q(m)$ to $q(m_h)$. Chain lengths thus decay as they progress down the hierarchy

of marginal products. Given the form of the solution for $q(m)$ implied by (28) and (29), hiring chains decay at a double-logarithmic rate. Vacancy chains decay more quickly as m declines, at a logarithmic rate, since each consecutive vacancy is filled with progressively lower probability, inducing fewer subsequent vacancies down the chain.

2.5 Steady-state equilibrium

The preceding results characterize equilibrium labor market outcomes for a given job-finding rate λ. Recall from the matching structure that the latter, in turn, is determined by labor market tightness θ. It remains to determine tightness and aggregate unemployment. In a steady-state equilibrium, these solve two conditions analogous to those in the canonical Mortensen and Pissarides (1994) model. First, a *Beveridge curve* sets inflows into unemployment, $\varsigma(\theta)(L-U)$, equal to outflows from unemployment, $\lambda(\theta)U$, as follows,

$$U_{BC}(\theta) = \frac{\varsigma(\theta)}{\varsigma(\theta)+\lambda(\theta)}L. \tag{32}$$

Second, a *job creation* condition summarizes aggregate labor demand,

$$U_{JC}(\theta) = L - \frac{X}{\int m^{1/(1-\alpha)} g(m;\theta)\mathrm{d}m}, \tag{33}$$

where $X \equiv \mathbb{E}\left[(\alpha x)^{1/(1-\alpha)}\right]$ determines the level of aggregate labor demand. Note that the aggregate stationarity condition in (10) ensures that the latter is constant over time and, thus, that aggregate labor market equilibrium is stationary. The ratio on the right-hand side of (33) is simply the expectation of employment across firms, $\mathbb{E}[n]$. To see this, recall that for every firm, $nm^{1/(1-\alpha)} \equiv (\alpha x)^{1/(1-\alpha)}$. The result then follows by calculating the mean of each side of the latter and observing that $g(m;\theta)$ is the *employment-weighted* density of marginal products.

3. Quantitative exploration

We have characterized a model that captures all the qualitative ingredients necessary to engage with the stylized facts documented in section 1. Multi-worker firms facing idiosyncratic shocks generate firm dynamics. On-the-job search generates endogenous quits. And, the creation of new positions involves a sunk investment—the expansion cost C—that generates replacement hiring and vacancy chains.

Table 1. Parameters and targeted moments of calibrated model (monthly frequency)

Parameter		Value	Reason / Moment	Model	Target
A. Externally calibrated					
ω_0	Flow breakdown payoff	0.948	Normalization	—	—
r	Discount rate	0.004	Annual real interest rate	0.05	0.05
α	Returns to scale	0.64	Cooper et al. (2007, 2015)	—	—
L	Labor force	21.28	Average firm size	20	20
B. Internally calibrated					
c	Gross hiring cost	1.051	Hiring costs / Monthly pay	1	1
C	Net expansion cost	35.03	12-month net inaction rate	0.271	0.271
σ	Std. dev. x shocks	0.175	Unemployment rate	0.06	0.06
X	Job creation curve shifter	217.4	U-to-E rate	0.25	0.25
A	Matching efficiency	1.236	Vacancy rate	0.025	0.025
ϵ	Matching elasticity	0.324	Beveridge curve elasticity	-1	-1
s	Employed search intensity	0.202	E-to-E rate	0.032	0.032
β	Worker bargaining power	0.052	Avg. job-to-job wage gain	0.08	0.08
ϖ	Elasticity of ω_0 to p	1.014	Elasticity of ς to Y/N	-3.6	-3.6

Notes. The rationale and source for each targeted moment are explained in the main text.

We now explore the model's ability to reconcile the stylized facts of section 1 from a quantitative perspective. To do so, we study a calibration informed by just one moment of the data on replacement hiring—the annual rate of net employment inaction. We then confront the calibrated model with an array of nontargeted outcomes, including the remaining moments of replacement hiring documented in section 1; the degree and persistence of markers of cross-sectional productivity dispersion; and the amplitude and persistence of aggregate labor market stocks and flows.

3.1 Calibration

Table 1 summarizes our calibration strategy. We begin by applying a normalization. Observe that, in a "frictionless" economy in which both the gross hiring cost and the net expansion cost are eliminated, $c = C = 0$, marginal products are equalized across firms at $m^* \equiv \omega_0/(1-\omega_1)$, implying a common wage $w^* \equiv \omega_0/(1-\beta)$. We normalize the latter

to one by setting $\omega_0 \equiv 1 - \beta$. The implication is that all flow variables reported in Table 1 are to be interpreted in units of frictionless wages.

We then externally calibrate the discount rate r to replicate an annual real interest rate of 5 percent; the returns to scale parameter α to mirror the estimates of Cooper et al. (2007, 2015); and the labor force L to yield an average firm size of 20. Panel A of Table 1 summarizes.

Internally calibrated parameters, together with target moments, are then reported in Panel B of Table 1. Although calibration of many of these parameters is in principle informed by all target moments, in what follows we provide an intuitive mapping between the parameters and the moments most naturally associated with them.

Key ingredients to the model are the firm-side frictions, encapsulated in the gross hiring cost c and the net expansion cost C. As we have emphasized, the former is the more conventional of the two. We choose c to correspond to a month of average wages. As noted by Manning (2011), although the literature provides relatively few estimates of the magnitude of hiring costs, a striking feature of the available estimates is that they broadly align with those reported in Oi's (1962) influential work: Hiring costs are mostly composed of training (rather than recruiting) costs, consistent with our choice to model hiring (rather than vacancy) costs; and their magnitude corresponds to around one month's pay. Most recently, Gavazza et al. (2018) report similar estimates of hiring costs compiled by human resources professionals.

Central to the model's ability to generate replacement hiring and, thereby, engage with the stylized facts that motivate the model, is the net expansion cost C. Accordingly, we use one of these motivating facts to discipline C, namely the annual rate of net inaction. From the range of estimates reported in Figure 2, we target the employment-weighted annual net inaction rate in the JOLTS, where inaction is defined as being within one worker, or up to one percent, of the initial employment level. This yields a target of 27.1 percent. Later, we explore alternative calibrations of the net expansion cost.

Next, we target labor market stocks and flows. Idiosyncratic shocks, as captured by the instantaneous standard deviation of log productivity σ, drive unemployment inflows. We therefore use σ to target the unemployment inflow rate, ς in (14), such that the steady-state unemployment rate is 6 percent. Turning to unemployment outflows, we choose the job creation curve shifter, X in (33), to generate a steady-state unemployment-to-employment transition rate λ equal to 25 percent, consistent with data from the Current Population Survey gross flows data.

We then map the latter outcomes to vacancy data by specifying a matching function. In common with much of the literature, we use a Cobb-Douglas matching technology,

$$M(U + s(L - U), V) = A[U + s(L - U)]^{\epsilon} V^{1-\epsilon}. \tag{34}$$

We choose matching efficiency A to generate a steady-state vacancy rate of 2.5 percent, and the matching elasticity ϵ such that the steady-state Beveridge curve relation between unemployment and vacancies has an elasticity of minus one. Both targets are broadly consistent with JOLTS data on job openings.

It remains to calibrate the model parameters that govern on-the-job search and wages. We select the search intensity of the employed s to replicate a monthly job-to-job transition rate of 3.2 percent, consistent with the estimates of Moscarini and Thomsson (2007). Relatedly, we choose worker bargaining power β to replicate an average wage increase upon a job-to-job transition of 8 log points, consistent with Barlevy (2008).

The latter fully describes the parameters that determine steady-state equilibrium. However, we also will be interested in how the model responds to aggregate shocks. Specifically, we will explore the effects of modifying the production function to $y = pxn^{\alpha}$, subject to changes in aggregate labor productivity p. Since it is natural that the outside options of firm and worker in the wage bargain will vary with aggregate labor productivity, we allow the flow breakdown payoff ω_0 to vary with p,

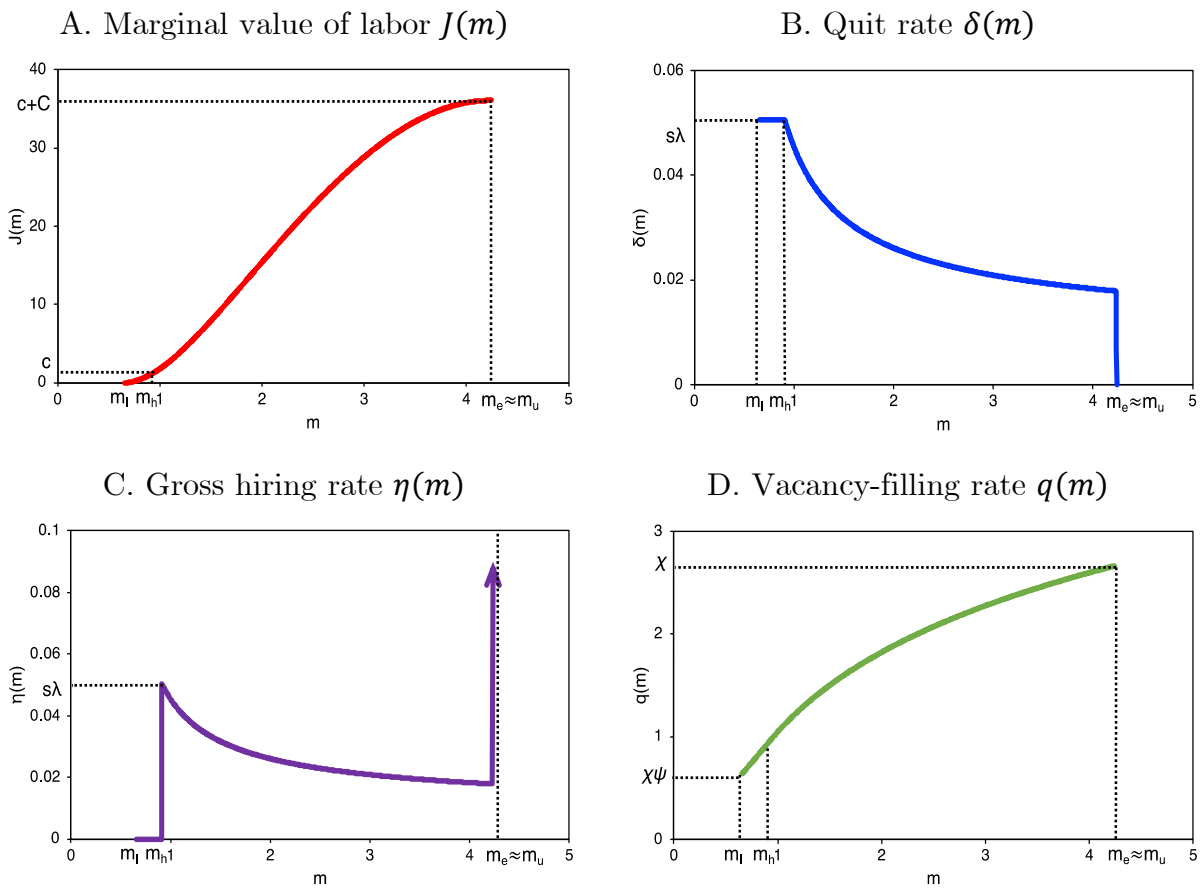
$$w(m) = \frac{\beta}{1 - \beta(1 - \alpha)} m + p^{\varpi} \omega_0. \tag{35}$$

Since our focus will be on the model's implications for *job creation*, we set the elasticity ϖ to pin down the (counter)cyclicality of *job destruction*. Specifically, we set ϖ such that the steady-state elasticity of ς with respect to output per worker in the model replicates the empirical relative standard deviation of the employment-to-unemployment transition rate of 3.6. This yields a ϖ close to one.[15]

This completes the calibration. In what follows, we explore its implications for a range of nontargeted moments. Before we do so, we first highlight key properties of the calibrated model.

[15] The value for worker bargaining power β of 0.052 implies a passthrough of productivity into wages that lies toward the lower end of the empirical estimates surveyed by Manning (2011). Although consistent with some prominent contributions to that literature (see, for example, Card et al. 2018), some recent estimates suggest a passthrough of 0.2 to 0.4 (Kline et al. 2019). Note, however, that the calibration of ϖ implies a near-unit passthrough of *aggregate* productivity into wages. Passthrough in the model thus depends on the source of variation in productivity.

Figure 5. Model outcomes



Notes. Parameter values are based on the model calibrated as described in Table 1.

3.2 Properties of the calibrated model

The most striking message of Table 1 is that the net expansion cost implied by the target moments is an order of magnitude larger than the gross hiring cost, $C \gg c$; indeed, over thirty times larger. In this way, the model provides a barometer for the quantitative significance of the replacement hiring documented in section 1 for labor market frictions. Viewed through the lens of the model, the requisite frictions are large, much larger than the conventional gross hiring costs that have informed canonical models. Echoing one motivation for our analysis, this elucidates the microeconomic origins of labor market frictions.

Figure 5 reinforces the quantitative significance of net expansion costs in the model. It plots the equilibrium outcomes characterized in the preceding section under the calibration in Table 1. A consequence of the magnitude of C is that the replacement region, $m \in (m_h, m_e)$, is much larger than its natural wastage and expansion counterparts.

This impression foreshadows several results that we confirm in the ensuing subsections. First, firms will tend to spend considerable episodes of time in the replacement region, naturally giving rise to many of the stylized facts documented in section 1. Second, the implied dispersion of marginal products m generated by the frictions that bring about replacement hiring is substantial. We will see that this goes some way to rationalizing the large dispersion in productivity across firms observed in available data.

Figure 5 conveys several other important facets of the calibrated model. First, in the replacement region, the hiring rate, and therefore the quit rate, is declining in the marginal product m (Panels B and C). Intuitively, more productive firms need to replace a smaller fraction of their workforce in this region. Nonetheless, note that *net* employment growth is weakly *increasing* in m, since contracting (expanding) firms shrink (grow) more when a lower (higher) m is realized.

Second, the vacancy filling rate is increasing in m (Panel D). Davis et al. (2013) study the related notion of a vacancy yield, defined as the share of current vacancies filled over a subsequent month, which can be measured directly in JOLTS microdata. Since net employment growth also is (weakly) increasing in m, the model predicts a positive association between vacancy yields and net employment growth (similar to Elsby and Gottfries 2021), and consistent with the findings of Davis et al. (2013).

Finally, the calibration implies an expansion region such that $m_e \approx m_u$. Indeed, for numerical purposes, m_e can be thought of (approximately) as a *reflecting* barrier: if the simulated path of m exceeds m_e, it is likely to revert back below m_e. As a result, firms that expand on net will in fact undertake sizable adjustments: a majority of gross hires will occur among firms that expand, even though the expansion region is rather narrow.

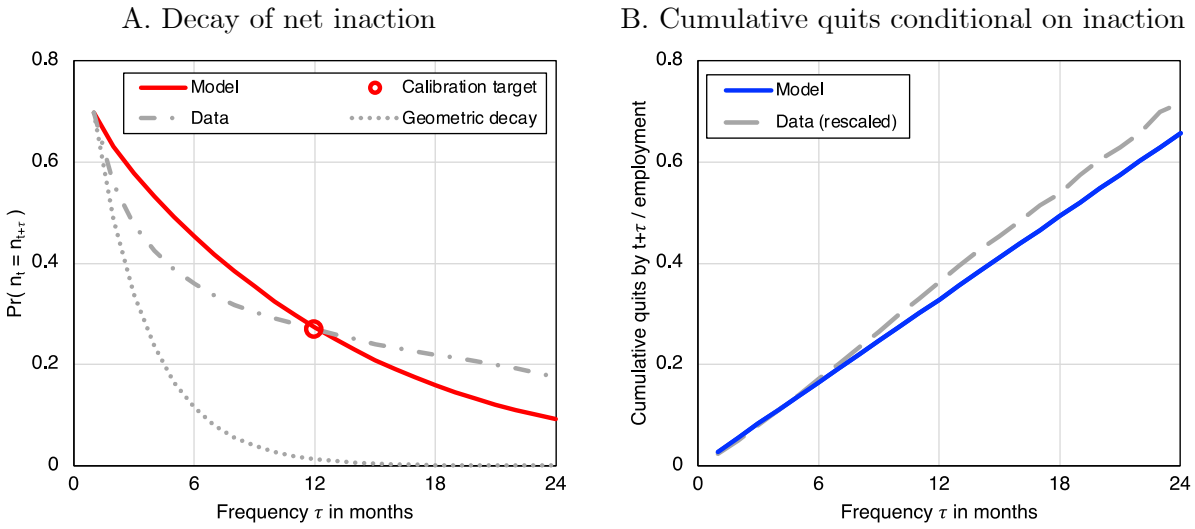
We now confront the calibrated model with a range of nontargeted moments motivated by the equilibrium outcomes in Figure 5.

3.3 Replacement hiring and vacancy chains

We begin by returning to the stylized facts documented in section 1 above. Figure 6 contrasts the empirical indicators of replacement hiring in Figures 2 and 3 with their model-implied analogues.

Consider first the decay of net inaction in Panel A. Recall that the calibration summarized in Table 1 targets only one of these moments—the *annual* employment-

Figure 6. Indicators of replacement hiring: Model versus data



Notes. Panel A: Data correspond to the employment-weighted net inaction rate in JOLTS from Figure 2 (inaction defined as being within one worker, or up to one percent, of initial employment). Panel B: Data correspond to the quits series from Figure 3, adjusted for the discrete drop in the empirical quit rate at zero growth due to small integer-sized establishments. We rescale by the ratio of the empirical quit rate at zero growth to that in adjacent growth bins (+/- 0.1 percent) in JOLTS data reported in Davis et al. (2012).

weighted net inaction rate in JOLTS (where inaction is defined as being within one worker, or up to one percent, of initial employment). This is depicted by the hollow circle in Figure 6A. The remaining net inaction rates by horizon are not targeted, and therefore provide a sense of the model's ability to match this dimension of replacement hiring in the data.

Two aspects of Figure 6A are reassuring. First, the model does a decent job of broadly matching the slope of the decay of net inaction by horizon. Indeed, it almost exactly replicates the monthly net inaction rate. At the same time, there are inevitable discrepancies, including underpredicting net inaction at longer horizons. Second, as in the data, the decay of inaction in the model is much slower than geometric. Employment adjustment in the model is thus very far from being independent across time.

Panel B of Figure 6 then reports cumulative quits conditional on net employment inaction by horizon in model and data. Recall from Figure 3 that this displayed a sizeable, near-linear relationship in the JOLTS data. Figure 6B compares the latter with its analogue in the model, subject to one adjustment. In the data, small, integer-sized establishments are disproportionately likely to report both zero growth, and zero quits, inducing a discrete drop in the empirical quit rate at zero net growth. For tractability, employment is continuous in the model, and so misses this feature of the data.

Accordingly, we rescale the data by the ratio of the empirical quit rate in near-zero growth bins (+/- 0.1 percent) to its rate at zero growth using estimates from Davis et al. (2012). Since only establishments of more significant size can report growth in the range of +/- 0.1 percent, this rescaling helps adjust for the integer constraint on smaller employers. Figure 6B reveals that, viewed this way, model and data line up well.

Next, mirroring our analysis of the prominence of replacement hiring as a share of total hires in Figure 4, we compute the same statistic implied by the model. Figure 4 suggested an empirical replacement share of hires on the order of 45 percent. The analogous measure in the calibrated model is of a similar magnitude: 51 percent.

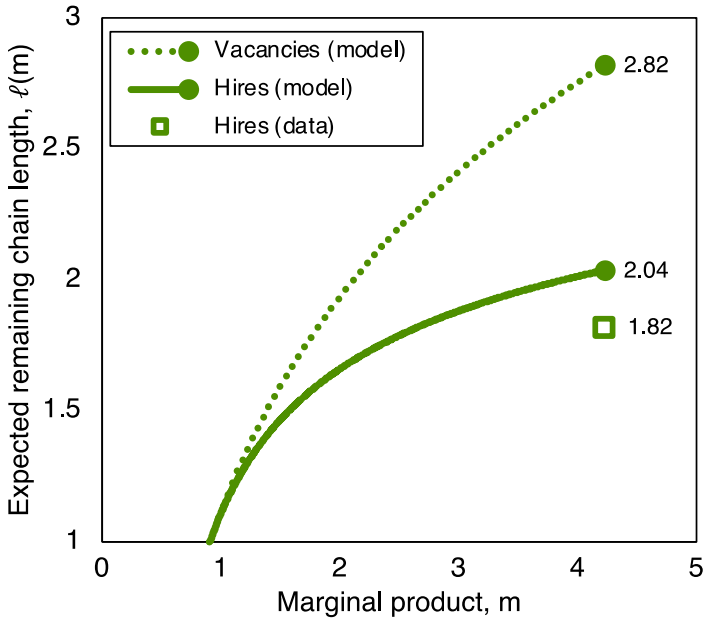
The replacement share is closely related to the length of hiring chains in the model. A convenient implication of the fact that $m_e \approx m_u$ given the calibrated parameters is that the replacement share of aggregate hires is well-approximated by $[\ell_H(m_e) - 1]/\ell_H(m_e)$, where $\ell_H(m_e)$ is the length of a hiring chain initiated at m_e: If a hire at m_e generates $\ell_H(m_e)$ *total* hires, $\ell_H(m_e) - 1$ must be for replacement, and the ratio of the latter to the former is the replacement share. Recall from Proposition 2 that $\ell_H(m)$ can in turn be recovered directly from the solution for the vacancy-filling rate $q(m)$. Figure 7 uses this to illustrate the expected remaining chain length as a function of the marginal product.

Figure 7 provides an additional perspective on replacement hiring in the model. Echoing the forgoing result that about half of hires are due to replacement, Figure 7 confirms the flipside implication that each new hire in the expansion region sets off a hiring chain with an expected length of approximately two (specifically, 2.04) in the model: Each job created generates on average double the number of hires. Likewise, when viewed through the lens of the model, the empirical replacement share of around 0.45 implies a hiring chain length of 1.82 in the data.

Figure 7 also reiterates two further properties of the chains generated by the model. First, vacancy chains are somewhat longer than hiring chains—each new vacancy created gives rise to 2.82 vacancies in total. Second, the expected remaining chain length decays as the chain progresses down the hierarchy of marginal products, reaching its lower bound of one at the lower limit of the replacement region, m_h.

Taken together, these results suggest that the model provides a compelling account of the key features of replacement hiring in the data that motivated our analysis. This lends further credence to the calibration in Table 1—specifically, the target moment for the net expansion cost C gives rise to results that line up not only with the decay of inaction, but also with other nontargeted empirical indicators of replacement hiring.

Figure 7. Chain length: Model versus data



Notes. Parameter values are based on the model calibrated as described in Table 1. The hiring chain length implied by the data is inferred from the empirical replacement share of hires of approximately 0.45 in Figure 4.

3.4 Dispersion in labor productivity

Viewed through the model, these indicators of replacement hiring imply considerable equilibrium dispersion in labor productivity across firms. A first glimpse of this is provided in Figure 5, which reveals that the marginal product differs as much as fourfold across firms in the calibrated model. It is natural, therefore, to explore the quantitative magnitude of productivity dispersion in the model in relation to the data.

Labor productivity dispersion in the model emerges primarily because the model requires considerable additional frictions in the form of the net expansion cost C to rationalize the stylized facts of replacement hiring. As reported in Table 1, the requisite friction is over thirty times larger than estimates of conventional gross hiring costs.

To illustrate this point, Table 2 reports the standard deviation of log average labor productivity under the baseline calibration in Table 1 and contrasts it with an alternative model in which we suspend the expansion cost. This provides a sense of the extent of productivity dispersion accounted for by C in the model.

Table 2. Productivity dispersion: Model versus data

Moment		Model		Data
		Baseline	$C = 0$	
Log average labor productivity	Std. dev.	0.44	0.18	0.58
	Autocorr. (annual)	0.36	0.03	—
Employment growth (annual)	Std. dev.	0.33	0.31	0.39
	Autocorr. (biannual)	0.19	0.02	0.16

Notes. Baseline parameter values are based on the model calibrated as described in Table 1. The $C = 0$ model is recalibrated to replicate all target moments in Table 1, aside from the annual net inaction rate. Data are from the following sources. The standard deviation of log average labor productivity is taken from Bartelsman et al. (2013). For employment growth: the standard deviation is based on data from Haltiwanger et al. (2013); the biannual autocorrelation is taken from Bloom (2009).

A key first impression of Table 2 is that productivity dispersion in the baseline calibrated model bears a close resemblance to related moments in the data. The cross-sectional standard deviation of log productivity implied by the model comes out at 0.44. As a point of comparison, Bartelsman et al. (2013) report a cross-establishment standard deviation of log revenue labor productivity of 0.58 for the United States. The suggestion, then, is that the same friction required by the model to replicate the degree of replacement hiring in the U.S. economy also implies an empirically-reasonable degree of productivity dispersion.

This feature of the model contrasts interestingly with recent work that has sought to identify the origins of productivity dispersion. In their recent study of the related question of dispersion in average *capital* productivity, David and Venkateswaran (2019) point out an impediment to accounts of productivity dispersion based on (convex) adjustment costs. Specifically, the degree of adjustment costs required to generate observed dispersion in productivity typically implies excessive persistence in measures of firm growth (investment in their case).

It is natural to ask, then, whether the same critique applies to our calibrated model. The lower panel of Table 2 reports measures of the dispersion and, crucially, persistence of employment growth implied by the model. Strikingly, both measures again resemble their empirical counterparts. The standard deviation of annual employment growth comes out at 0.33 in the model. The analogous estimate based on data from the Longitudinal Business Database provided by Haltiwanger et al. (2013) is 0.39. More importantly, the

biannual autocorrelation of employment growth in the model is 0.19. Bloom (2009) reports an analogous estimate from Compustat data of 0.16. The model thus generates realistic productivity dispersion without implying excessive persistence in firm-level employment growth.

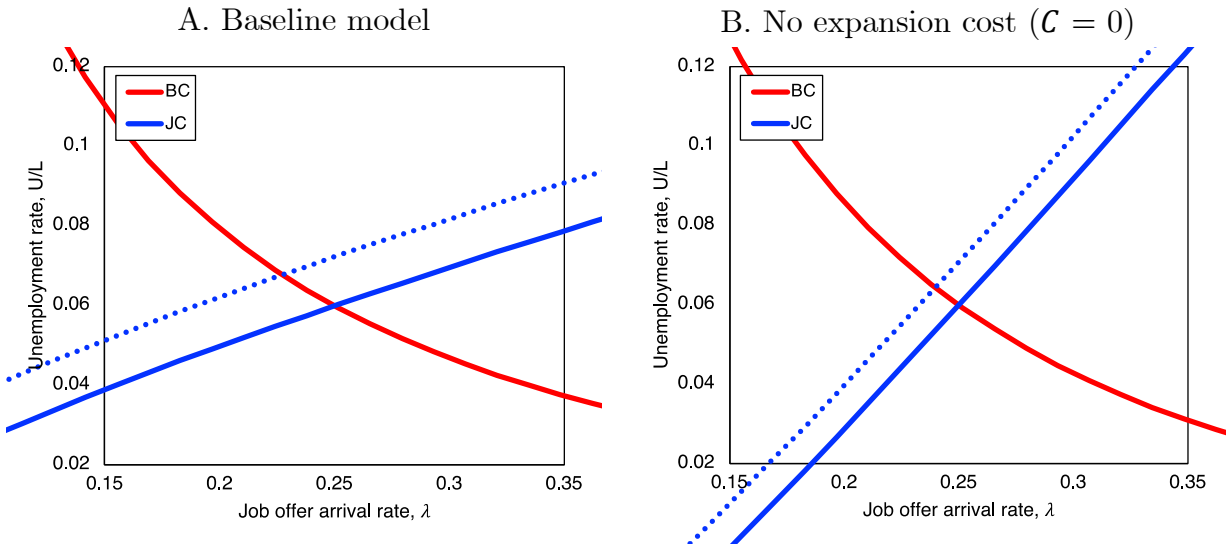
The large degree of productivity dispersion in the calibrated model is, in turn, quite persistent across firms: Table 2 reports an annual autocorrelation of log labor productivity of 0.36. Although we are unaware of an empirical analogue for the United States, a sense of magnitudes is provided by Lentz and Mortensen (2012) who report annual autocorrelations on the order of 0.5 in firm-level data for Denmark. Again, the model's implications are of a reasonable order of magnitude relative to these data.

Table 2 then explores the origins of productivity dispersion in the model. As an alternative, we suspend the expansion cost and recalibrate to replicate all target moments in Table 1, aside from the annual net inaction rate (which is zero given $C = 0$). Table 2 reveals that this version of the model delivers both considerably lower magnitude and persistence of productivity dispersion. Although able to generate realistic dispersion in employment growth, the $C = 0$ recalibrated model implies a standard deviation of log productivity of only 0.18, and very limited autocorrelation of productivity and employment growth, far short of empirical counterparts. This underscores the key role of replacement hiring in generating realistic productivity dispersion in the model.

4. Volatility and persistence of labor market dynamics

The preceding analyses have confirmed that the calibrated model is able to reconcile an array of relevant empirical properties of cross-sectional establishment dynamics, from a range of indicators of replacement hiring, to the level and persistence of cross-establishment dispersion in productivity and employment growth. We now explore the macroeconomic implications of the calibrated model for aggregate labor market dynamics. Two enduring puzzles of the latter have been the empirical volatility and persistence of the rates of unemployment and job finding. In what follows, we address each of these puzzles in turn. We will see that, viewed through the model, replacement hiring, and the vacancy chains that result, can play a crucial role in their resolution.

Figure 8. Comparative steady states of calibrated model



Notes. Based on the model calibrated as in Table 1. The figure illustrates the steady-state response to a one-percent decline in aggregate labor productivity.

4.1 Volatility

Figure 8 provides a sense of the calibrated model's implications for labor market volatility and its economic origins. It depicts an analysis of comparative steady states induced by a one-percent decline in aggregate labor productivity p—recall, a firm's production function is now taken to be $y = pxn^{\alpha}$—and juxtaposes it with the $C = 0$ recalibration that suspends replacement hiring. Steady-state equilibrium is determined by the intersection of the Beveridge curve condition (32), and the job creation condition (33). Figure 8 thus illustrates these for each model variant.

Figure 8 delivers two key messages. First, the baseline model with replacement hiring generates considerably greater volatility relative to its $C = 0$ counterpart, especially in regards to the job finding rate λ. Second, the origins of this additional volatility can be traced primarily to the slope of the job creation condition, which is substantially shallower in the presence of replacement hiring.

The intuition is as follows. The slope of the job creation condition summarizes the feedback of job creation decisions across firms—the extent to which a rise in other firms' job creation, as captured by λ, chokes off labor demand. Absent replacement hiring, rises in λ increase firms' turnover costs and depress their labor demand. The key is that replacement hiring partially neutralizes this channel: Rises in λ have no effect on the labor

Table 3. Amplitude of labor market stocks and flows: Model versus data

Moment	Model		Data
	Baseline	$C = 0$	Relative sd.
A. Response relative to output per worker			
Unemployment rate	14.0	8.1	14.0
Vacancy rate	14.0	8.1	12.5
U-to-E rate	11.3	5.0	11.6
E-to-U rate	3.6	3.6	3.6
E-to-E/Quit rate	8.1	4.8	CPS: 5.7 DFH-JOLTS: 8.5
	Baseline	$C = 0$	Semi-elasticity
B. Response relative to unemployment rate			
Average wage	-1.3	-1.7	≈ -1

Notes. Model outcomes are (absolute values of) steady-state elasticities with respect to output per worker in Panel A, and steady-state semi-elasticities with respect to the unemployment rate in Panel B. Data in Panel A are based on an update and extension of the empirical results of Shimer (2005) for the period 1994 to 2016. Vacancies are measured using Barnichon's (2010) Composite Help-Wanted Index (which terminates in 2016). The quit rate is measured using Fallick and Fleischman's (2004) CPS-based estimates of the job-to-job transition rate (available from 1994), and Davis, Faberman and Haltiwanger's (DFH, 2012) synthetic JOLTS-BED-based measure (available from 1990Q2 to 2010Q2), extended from 2010Q2 using JOLTS data. Data reported in Panel B are a summary of Solon et al. (1994) and Elsby et al. (2016).

demand of firms in the replacement region, as they continue to hire to replace their quits. Thus, the net expansion of some firms does not crowd out hiring in the replacement region. Rather, since expansion is fueled partly by poaching employed workers, this hiring is propagated through the replacement region via hiring, and vacancy, chains. The upshot is that the negative feedback in job creation decisions is weakened, and the job creation condition flattens, amplifying the response of the job-finding rate λ, and thereby the unemployment rate, to adverse shifts in aggregate labor demand.

Table 3 provides a quantitative sense of the degree of labor market volatility generated by the calibrated model. For each model variant, it reports the (absolute value) of the steady-state elasticities of labor market outcomes relative to output per worker. These are then confronted with empirical results from an update and extension of Shimer (2005) to include Barnichon's (2010) Composite Help-Wanted Index (available up to 2016), Fallick and Fleischman's (2004) data on the CPS job-to-job transition rate (available from 1994), and Davis, Faberman and Haltiwanger's (2012) JOLTS-BED

synthetic quit rate series. Since the latter is available up to 2010Q2, we project their series forward to 2016 using its relation with the JOLTS quit rate over their common sample period. Following Shimer (2005), Table 3 uses these 1994-2016 data to report the relative standard deviations of quarterly log-detrended outcomes with respect to output per worker. This reiterates the calibration target of a relative volatility of the E-to-U rate equal to 3.6.

The results in Table 3 for the baseline model with replacement hiring are striking. The model almost exactly replicates the unconditional volatilities of unemployment and vacancies in the data. But even more starkly, the same is true of the job-finding rate from unemployment: As in the data, the U-to-E rate in the model is over 10 times more volatile than average labor productivity. With some justification, the model can claim to provide a resolution of Shimer's (2005) well-known volatility puzzle.

Table 3 also confirms the visual impression of Figure 8 that replacement hiring plays a central role in the volatility of labor market outcomes. Suspending the expansion cost ($C = 0$) and recalibrating lowers the volatility of the job-finding rate from unemployment by half, and the volatilities of the unemployment and vacancy rates by nearly a half.

Table 3 highlights two further aspects of the baseline calibrated model. First, the model implies an E-to-E (quit) rate that is a little over 8 times more volatile than average labor productivity. Although this overstates the volatility of the CPS job-to-job transition rate as reported by Fallick and Fleischman (2004), it is very much in line with the volatility of Davis, Faberman and Haltiwanger's JOLTS-BED quit rate. Second, the model implies a semi-elasticity of average wages with respect to the unemployment rate of -1.3. The latter almost exactly replicates the influential estimates of the procyclicality of real wages reported by Solon et al. (1994). Broadly speaking, real wages are about as procyclical as aggregate employment, in model as in data.

Stepping back, the message of this exercise is that the same ingredients that reconcile the array of evidence for replacement hiring in the cross-section also give rise to the considerable volatility in labor market outcomes we observe in the time series, and do so while replicating the observed procyclicality of real wages.

4.2 Persistence

We turn next to a further quantitative property of the calibrated model, namely its implications for the *persistence* of labor market outcomes, and contrast these with the well-known sluggish dynamics of job finding and unemployment in the data.

Doing so requires a solution for the transition dynamics of the model, however, implementation of which is easier said than done. Recall that a central challenge of the environment is that equilibrium involves solution for the distributions of job values among offers, and across workers, summarized by $F(m)$ and $G(m)$. In steady state, the solution to this challenge is provided by Lemmas 1 and 2, and Proposition 1. Out of steady state, however, solution for the transition dynamics of the model requires solution for the *dynamic path* of these distributions.

To infer the model's implications for the persistence of labor market dynamics, then, we begin by providing a method for inferring the model's transition dynamics in response to an MIT shock to aggregate productivity. We then confront the results of this exercise with the empirical dynamics of rates of job finding and unemployment. Strikingly, we will see that the calibrated model goes a considerable way toward accounting for the observed persistence in empirical labor market dynamics.

Solution method. In simpler models of firm dynamics, the approach to solving for transition dynamics would be to solve backward for labor demand using the out-of-steady-state analogue of the Bellman equation (9), and to solve forward for the worker distribution using the flows of workers across firms summarized by the Fokker-Planck (Kolmogorov Forward) equation. The interaction of on-the-job-search, firm dynamics and, especially, replacement hiring, greatly frustrates this scheme, however. To see how, consider first the out-of-steady-state Bellman equation for the marginal value of the firm,

$$\begin{aligned} rJ_t(m) = (1-\omega_1)m - \omega_0 - [\delta_t(m) - (1-\alpha)m\delta_t'(m)]\min\{J_t(m), c\} \\ + \left[\mu + (1-\alpha)\delta_t(m)\mathbf{1}_{\{\mathrm{d}n^*<0\}}\right]mJ_t'(m) + \frac{1}{2}\sigma^2 m^2 J_t''(m) + \frac{\partial J_t(m)}{\partial t}. \end{aligned} \tag{36}$$

Equation (36) suggests that, in general, the solution for the time path of $J_t(m)$ would require knowledge of the time path of equilibrium quit *functions*, $\delta_t(m)$. One can verify that the same information is, in principle, also required to solve the Fokker-Planck equation that describes the flow of workers across firms.

The key to our approach is to note that the same insights that inform the model's steady-state solution also greatly simplify solution for its transition dynamics. Specifically, using the results of section 2, we show that it is possible to solve for the transition dynamics armed only with knowledge of a sequence of *scalars*: the job-finding rate λ_t.

This result is especially straightforward in the natural wastage and expansion regions. In the former, the quit rate is equal to $s\lambda_t$, and so directly follows from knowledge of λ_t.

In the expansion region, the firm's marginal value is equal to a constant, the sum of gross hiring and net expansion costs, $J_t(m) = c + C$ for all $m \in (m_{et}, m_{ut})$. As in steady state, an implication is that $J_t'(m) = 0 = J_t''(m)$ in the expansion region. Crucially, the out-of-steady-state capital gains also are zero in the expansion region, $\partial J_t(m)/\partial t = 0$ for all $m \in (m_{et}, m_{ut})$. It follows from (36) that the same decoupling result that aids steady-state solution for the quit rate in the expansion region also simplifies its solution out of steady state: Specifically, $\delta_t(m)$ shares the same functional form as in Lemma 2 for all $m \in (m_{et}, m_{ut})$, and thus is known up to the path of two scalars, m_{et} and λ_t.

Matters are more complicated in the replacement region, however. There, the absence of a decoupling of marginal value $J_t(m)$ and quit rate $\delta_t(m)$ that frustrated the steady-state solution in turn impedes solution for the transition dynamics. Again, though, the insights of section 2 provide a clue: Recall that the defining feature of the replacement region is that the hiring rate equals the quit rate, $\eta_t(m) = \delta_t(m)$. An implication is that the steady-state inverse-proportionality of the quit and vacancy-filling rates identified in Proposition 1 also holds out of steady state,[16]

$$q_t(m) = q(m_{ht})\frac{s\lambda_t}{\delta_t(m)}, \text{ for all } m \in (m_{ht}, m_{et}) \text{ and } t. \tag{37}$$

The quit rate in the replacement region is thus known up to the path of λ_t and $q_t(m)$.

The upshot is that, for any sequence of the job-finding rate λ_t, we can formulate a fixed-point problem: On one hand, the Bellman equation (36) can be solved backward for a given sequence of filling rates $q_t(m)$ (in the replacement region) to yield a sequence of boundaries $\{m_{lt}, m_{ht}, m_{et}, m_{ut}\}$. On the other hand, the Fokker-Planck equation can be solved forward for a given sequence of boundaries $\{m_{lt}, m_{ht}, m_{et}, m_{ut}\}$ to yield a sequence of worker distributions $G_t(m)$ and, thereby, filling rates $q_t(m)$.

The model thus delivers an algorithm for feasible solution of its transition dynamics: In an outer loop, we conjecture a time path for the job offer arrival rate, λ_t. In an inner loop, we then solve a fixed point for the time paths of the boundaries $\{m_{lt}, m_{ht}, m_{et}, m_{ut}\}$, and the vacancy-filling rate $q_t(m)$, using the out-of-steady-state Bellman and Fokker-Planck equations. Finally, we return to the outer loop, and iterate over the path of λ_t

[16] The hiring rate at m is equal to the ratio of the total measure of hires to the total measure of employees at m, $\eta_t(m) = q_t(m)f_t(m)V_t/[g_t(m)(L - U_t)]$. Using (4) and (5), and the matching structure, yields $\eta_t(m) = -q_t(m)\delta_t'(m)/q_t'(m)$, which in turn must equal $\delta_t(m)$ in the replacement region. (37) follows.

until a measure of excess labor demand at each point in time is reduced to zero (up to numerical error). The Appendix provides further detail.

Transition dynamics. The fruits of this algorithm are depicted in Figure 9, which plots the transition dynamics induced by an unanticipated one-percent, permanent decline in aggregate labor productivity p. These results suggest the following narrative.

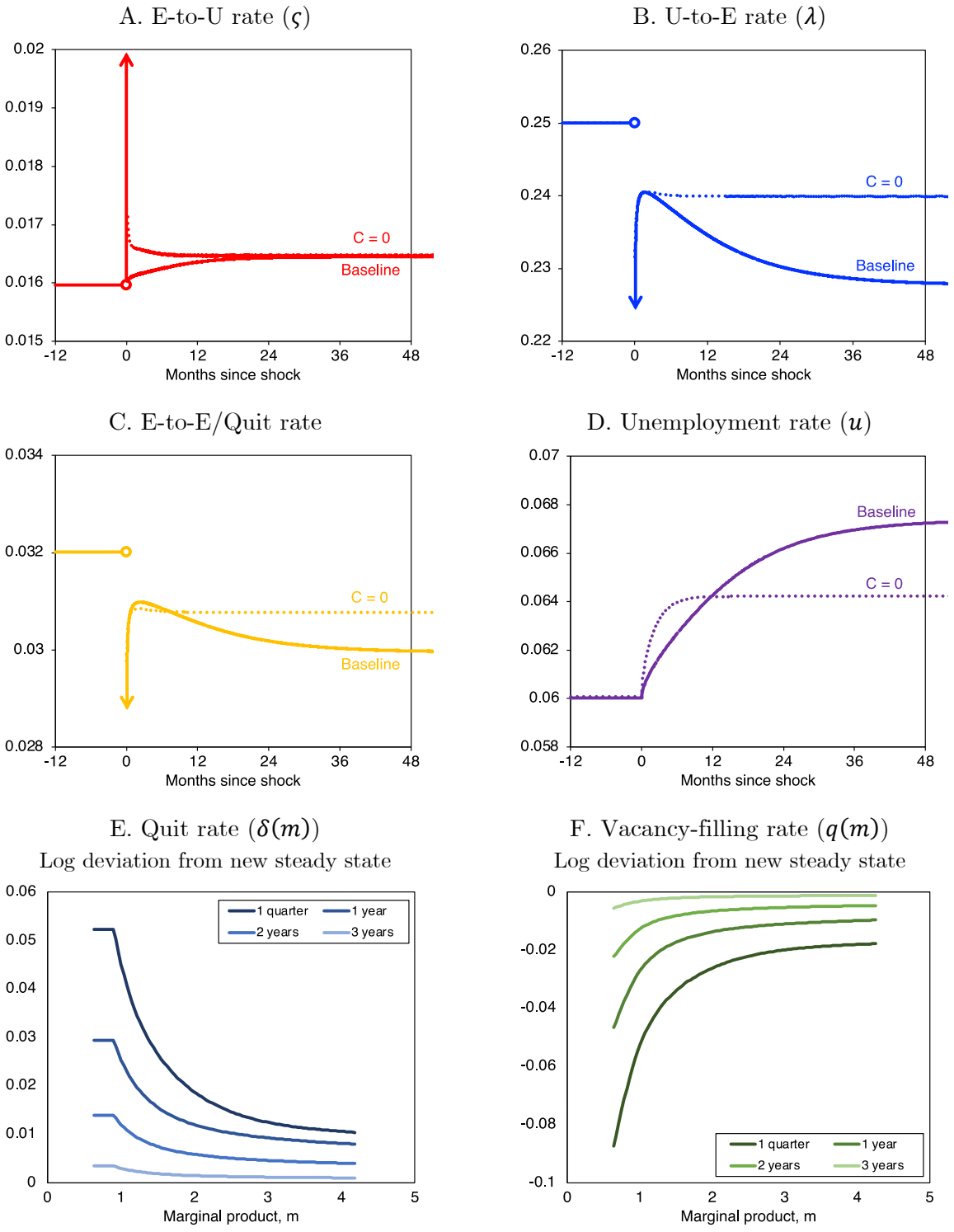
On impact of the shock, the adjustment boundaries jump. At the job destruction margin, a discrete mass of employees is laid off, as depicted by the upward arrow in the E-to-U rate in Panel A. At the job creation margin, firms previously in the expansion region no longer wish to create new positions, *ceteris paribus*. As a consequence, the offer arrival rate λ drops precipitously to restore incentives to expand, as annotated by the downward arrow in the U-to-E rate in Panel B. Shortly thereafter, idiosyncratic shocks replenish the expansion region, and λ recovers. These overshooting properties are related to the discontinuous nature of the aggregate shock, and are resolved in a matter of weeks.

More important for engaging with data at conventional frequencies are the subsequent responses. Crucially, the dynamics of job creation are slow-moving: After the first quarter, it takes approximately one year to close half the residual gap between λ_t and its new steady state. These persistent dynamics of job creation in turn carry over to the E-to-E rate in Panel C, the quit rate in Panel E and, most strikingly, the unemployment rate in Panel D: The dynamics of unemployment exhibit a half-life of 9.6 months. As we will confirm shortly, the model provides fertile ground for an account of the observed persistence of joblessness.

This result is intimately connected to the presence of vacancy chains in the model. We demonstrate this in two ways. First, we contrast the dynamics of the baseline calibrated model with its analogue without replacement hiring—namely the $C = 0$ counterpart studied in earlier sections. Second, we use the baseline model to decompose the aggregate hiring rate into components accounted for by the creation of new jobs, and by the length of the vacancy chains propagated by them. Both underscore the central role of vacancy chains in the persistence of job creation and unemployment in the model.

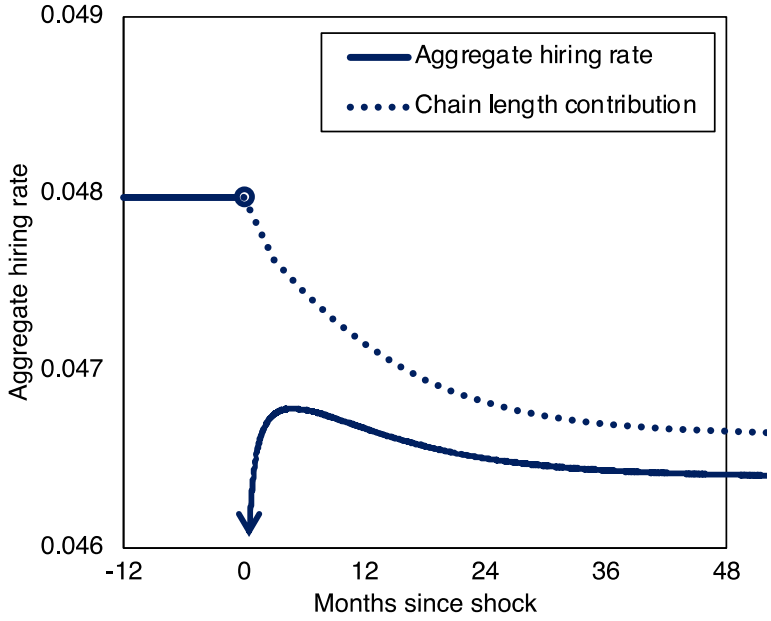
The first exercise is illustrated by the dotted lines in Panels A through D of Figure 9, which depict the analogous transition dynamics of the $C = 0$ model. This reiterates the message of Table 3 that the presence of replacement hiring amplifies the response of job creation, and thereby unemployment, to aggregate shocks. But, in addition, it further

Figure 9. Transition dynamics of calibrated model



Notes. Based on simulation of the model calibrated as described in Table 1. The figure illustrates the dynamic response to an unanticipated, permanent one-percent decline in aggregate labor productivity. The arrows depict points that are off-scale.

Figure 10. Chain length and the persistence of aggregate hires



Notes. Based on simulation of the model calibrated as described in Table 1. The figure illustrates the dynamic response to an unanticipated, permanent one-percent decline in aggregate labor productivity. The arrow depicts points that are off-scale.

reveals the role of vacancy chains in propagating labor market dynamics. Absent chains, the job-finding rate λ is effectively a jump variable. And, consequently, unemployment is far less persistent: its half-life is only 1.2 months (as opposed to over 9 months).

Figure 10 illustrates a second perspective. It asks, within the calibrated baseline model, how much of the variation in aggregate hires is accounted for by variation in the length of vacancy chains. To do so, it exploits the identity that the aggregate hiring rate, denoted H_t, is equal to the product of the new job creation rate, denoted v_t, and the average length of the hiring chains propagated by each new position. Using the fact that $m_{et} \approx m_{ut}$ for all t under the calibration, we have

$$\mathrm{H}_t \approx v_t \cdot \ell_{Ht}(m_{et}). \tag{38}$$

Figure 10 uses this identity to decompose the contribution of chain length to the dynamics of the aggregate hiring rate. Again, this reiterates the role of replacement hiring in the amplitude of labor market fluctuations in the model: The shortening of vacancy chains accounts for the majority of the decline in the hiring rate (following its initial

overshooting). But, in addition, Figure 10 confirms that the *entirety* of the persistence of the hiring rate in the model can be traced to the sluggish dynamics of chain length.

To understand why, the approximation that $m_{et} \approx m_{ut}$ is again useful. It implies that the expansion region is approximated by a reflecting upper barrier. The rate of new job creation ν therefore mirrors the rate of job destruction at the layoff boundary, ς in (14). A density $g(m_e)$ of workers receives shocks to their log marginal product of instantaneous variance σ^2. Positive innovations induce firms to create new positions until the marginal product is (approximately) reflected back to m_e, at a rate determined by the elasticity of labor demand $1/(1-\alpha)$,

$$\nu_t \approx \frac{\sigma^2/2}{1-\alpha} m_{et} g_t(m_{et}). \tag{39}$$

The immediate collapse and subsequent reversion of the hiring rate in Figure 10 is thus driven by jumps in the boundary m_{et}, which imply an overshooting of the new job creation rate ν that is the counterpart of the dynamics of the layoff rate ς in Figure 9A.

By contrast, chain length has naturally sluggish dynamics. The steady-state solution for chain length in Proposition 2 also holds out of steady state,

$$\ell_{Ht}(m) = 1 + \ln \frac{q_t(m)}{q_t(m_{ht})}, \text{ for all } m \in (m_{ht}, m_{ut}) \text{ and } t. \tag{40}$$

The dynamics of chain length are thus determined by the dynamics of the vacancy-filling rate $q_t(m)$ and, thereby, the worker distribution $G_t(m)$. As confirmed by Figure 9F, the latter is a slow-moving state variable, since its evolution is determined by the (slow) movement of employees across marginal products. Intuitively, chains end whenever a worker is hired from unemployment or the natural wastage region. The probability of the latter is shaped by the measure of searchers across those states, $u_t + s(1-u_t)G_t(m_{ht})$, a slow-moving state variable, reflecting the gradual evolution of the distribution of workers across marginal products and unemployment. Vacancy chains thus play a central role in the persistence of labor market stocks and flows in the model.

Quantitative assessment. We now assess the extent to which the model is able to account for the observed persistence of rates of unemployment and job finding in the data. We begin by documenting these properties using the data underlying Table 3 above.

A common barometer is to compare the dynamics of labor market outcomes to those of aggregate output per worker. We use an approach similar to Fujita and Ramey (2007).

We begin by estimating an autoregressive specification for output per worker, denoted z, conditional on lags of the unemployment rate u,

$$\ln z_t = a^z + \sum_{s=1}^{\mathcal{L}} b_s^z \ln z_{t-s} + \sum_{s=1}^{\mathcal{L}} c_s^z \ln u_{t-s} + e_t^z. \tag{41}$$

The estimated residuals $\hat{e}_t^z$ comprise innovations to z that are unforecastable given lags of u. In a second stage, we estimate distributed lag models of unemployment and the job-finding rate on these estimated residuals,

$$\begin{aligned} \ln u_t &= a^u + \sum_{s=0}^{\mathcal{L}-1} b_s^u \ln \hat{e}_{t-s}^z + \sum_{s=1}^{\mathcal{L}} c_s^u \ln u_{t-s} + \sum_{s=1}^{\mathcal{L}} d_s^u \ln \lambda_{t-s} + e_t^u, \text{ and} \\ \ln \lambda_t &= a^\lambda + \sum_{s=0}^{\mathcal{L}-1} b_s^\lambda \ln \hat{e}_{t-s}^z + \sum_{s=1}^{\mathcal{L}} c_s^\lambda \ln u_{t-s} + \sum_{s=1}^{\mathcal{L}} d_s^\lambda \ln \lambda_{t-s} + e_t^\lambda. \end{aligned} \tag{42}$$

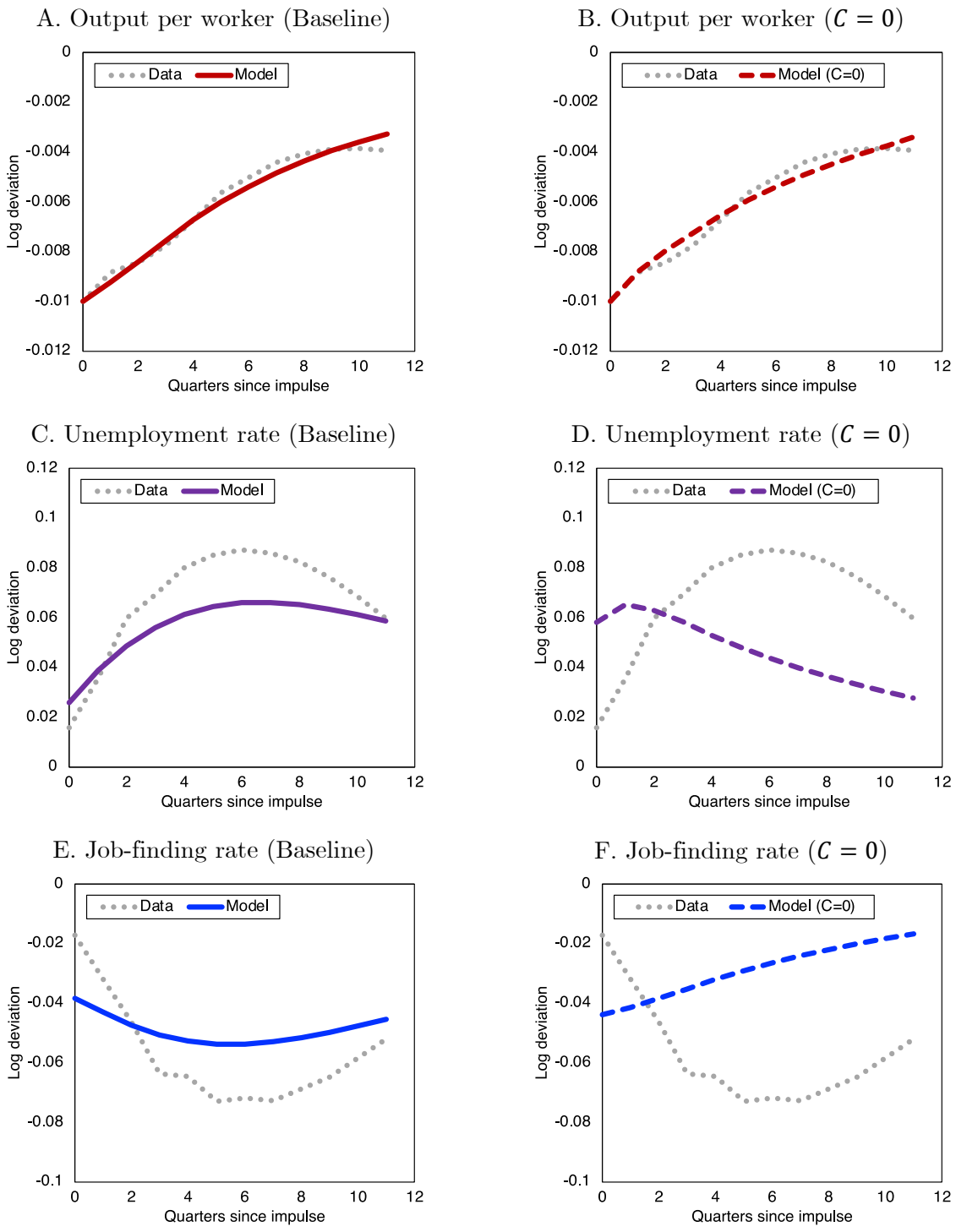
Note the timing in the lag structure of innovations to output per worker, which allows for a contemporaneous relationship between these innovations and rates of unemployment and job finding. We use a lag order of $\mathcal{L} = 4$ quarters in both stages.[17]

The dotted lines in Figure 11 plot the results for a one-log-point negative innovation to output per worker. This reveals that output per worker is very persistent in the data: it reverts only halfway to its pre-innovation level in eleven quarters. But Figure 11 also confirms the well-known persistence of unemployment and job-finding rates. These exhibit familiar hump-shaped dynamics with peaks as much as six quarters after the innovation. These persistent responses echo similar results found elsewhere in the literature (Blanchard and Diamond 1989; Hagedorn and Manovskii 2011).

How does the baseline calibrated model compare? The bold lines in Figure 11 depict the results of an exercise in the spirit of Boppart et al. (2018). Specifically, we use the impulse responses in Figure 9 as numerical derivatives to compute a first-order approximation of the model solution in a stochastic environment with recurrent innovations to aggregate productivity p. Given the environment that underlies Figure 9, the approximation is likely to be accurate when innovations are small and persistent.

[17] The impulse responses implied by (41) and (42) lie in between results from alternative specifications in which we (i) use a lag order of $\mathcal{L} = 6$; (ii) enter $\ln z$ in first differences instead of levels; and (iii) estimate (41) with lags of λ on the right-hand side. In each case, the responses of the unemployment and job-finding rates are similar to the baseline case reported in Figure 11 (see Appendix D). Interestingly, case (iii) is unstable when applied to data generated by the model with $C = 0$, since z and λ in that model are nearly collinear. For this reason, we omit lags of λ from (41) in our baseline case.

Figure 11. Descriptive impulse responses: Model versus data (quarterly frequency)



Notes. Impulse responses to a negative one percent innovation to output per worker implied by estimation of (41) and (42). Dotted lines represent the data; solid lines (left column) correspond to the baseline ($C > 0$) model; dashed lines (right column) correspond to the model recalibrated with $C = 0$.

To implement the algorithm, we assume log aggregate labor productivity in the model follows an AR(1), $\ln p_t = \rho_p \ln p_{t-1} + \epsilon_t$, with innovations $\epsilon_t \sim \mathcal{N}(0, \sigma_p^2)$, and simulate a quarterly time series of model outcomes implied by the sequence of innovations to p. We then estimate equations (41) and (42) on the model-generated data. To place model and data on an equal footing, we choose the persistence parameter ρ_p to minimize the (Euclidean) distance between the estimated impulse responses of output per worker in model and data, and σ_p to replicate the empirical standard deviation of output per worker. This yields a ρ_p of 0.923, and a σ_p of 0.00785.

The results for the baseline model in Figure 11 are encouraging. The AR(1) specification for $\ln p$ is able to replicate closely the empirical response of output per worker in Panel A. More importantly, this is associated with a large, prolonged increase in the unemployment rate in Panel C. In turn, mirroring the data, this can be traced in large part to a large, prolonged decrease in the rate of job finding in Panel E. Thus, as foreshadowed by the theoretical impulse responses in Figure 9 above, the baseline model has a propagation mechanism that can account for the empirical observation that unemployment dynamics are sluggish, and that this derives from sluggishness in job creation. Where model and data depart is in the timing of the response: Dynamics in the model—most notably, in the response of the job finding rate—are somewhat less hump-shaped than their empirical counterparts.

The remaining panels of Figure 11 apply the same methods to the recalibrated $C = 0$ model without replacement hiring.[18] These reinforce the relative success of the baseline case. Although the dynamics of output per worker are again largely replicated in Panel B, the responses of the unemployment and job-finding rates in Panels D and F are smaller, and much more short-lived, than their empirical analogues. Again, these reiterate the impression of the essentially-jump dynamics of job creation in the $C = 0$ case in Figure 9.

Summing up, we have demonstrated in this section that our model of vacancy chains provides a parsimonious reconciliation of the volatility and persistence of labor market stocks and flows: The addition of a single parameter—the sunk cost of new job creation, $C > 0$—amplifies and prolongs labor market dynamics. That these results are derived from a calibration that is simultaneously consistent with an array of plant-level facts—the incidence and decay of net action, and the dispersion of labor productivity—adds further credence to the model's outcomes.

[18] The resulting persistence and standard deviation of p in this case are $\rho_p = 0.912$, and $\sigma_p = 0.00948$.

5. Robustness and extensions

We now consider variations of the baseline model studied up to now. We focus on two. First, we explore the quantitative implications of varying the degree of replacement hiring for key outcomes of the model. Second, we show that the theory is amenable to an extension to incorporate offer matching in wage determination.

5.1 Varying the degree of replacement hiring

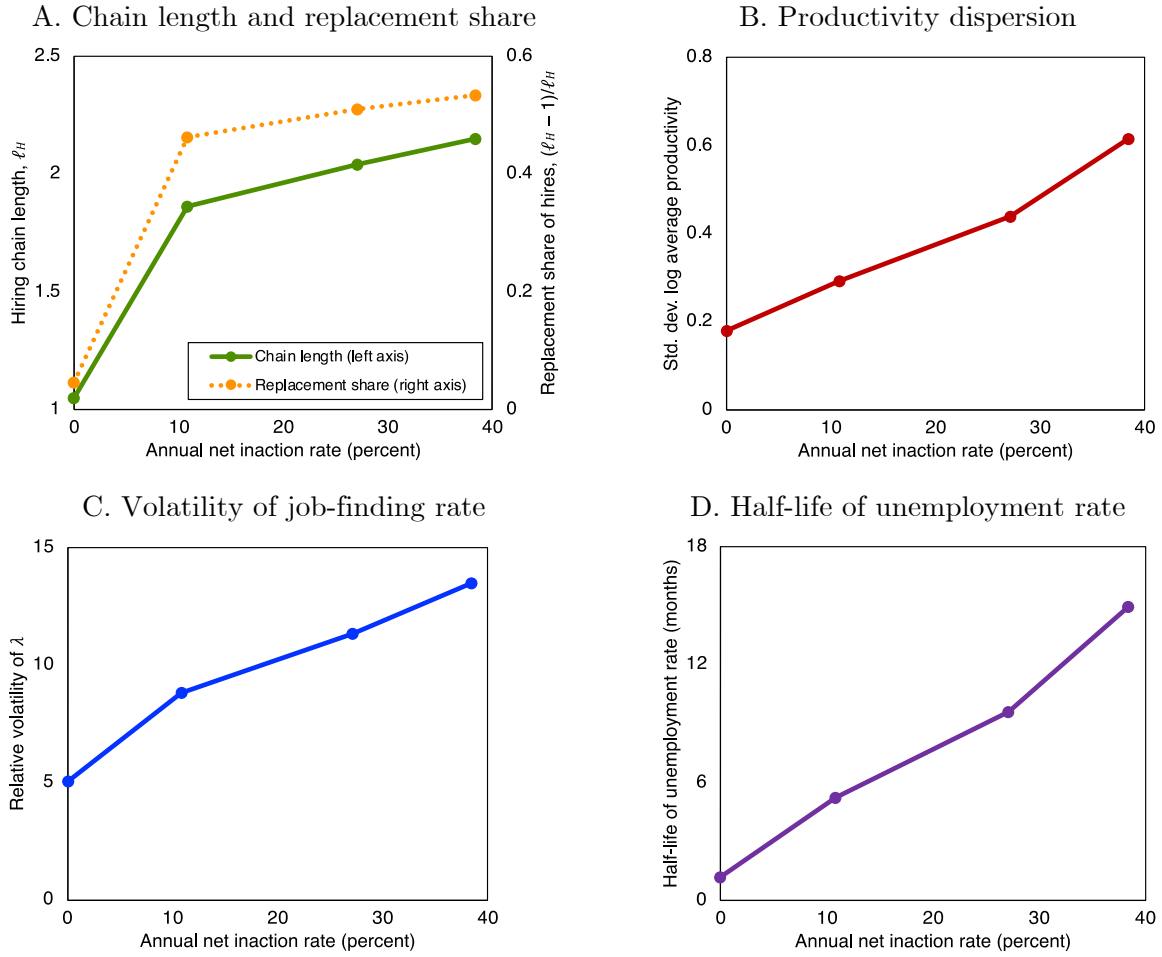
The key innovation of this paper has been to document empirical markers of replacement hiring and to devise a model that traces out their implications for labor market dynamics. In assessing the robustness of our findings, it is therefore natural to consider alternative targets for the degree of replacement hiring.

Recall that the calibration summarized in Table 1 targeted an employment-weighted annual rate of net inaction equal to 27.1 percent. This corresponds to one of the estimates reported in Figure 2 based on JOLTS microdata—specifically, that which defines inaction as being within one worker, or up to one percent, of the initial employment level. As we have seen, the calibration implied by this target dovetails with many other cross-sectional and times series dimensions of labor market data.

It is natural to ask, however, how outcomes would change if we were to target alternative indicators of the degree of replacement hiring. Accordingly, we explore the results of targeting the estimates of annual net inaction rates over the different windows of adjustment reported in Figure 2. Specifically, we consider employment-weighted annual net inaction rates of 10.8 percent (no window) and 38.4 percent (2 workers, or 2 percent window), in addition to the zero and 27.1 rates already considered. In each case, we simply insert the new target for the net inaction rate, and recalibrate as per Table 1.

Figure 12 presents the results of this exercise. This yields a few important takeaways. First, increases in C imply a higher incidence of full replacement (in which quits are offset by hires), which in turn stretches the length of vacancy chains (Panel A). Second, and relatedly, productivity dispersion (Panel B) as well as aggregate volatility and persistence (Panels C and D, respectively) also scale up with the size of the friction C. These results are in line with intuition—for example, a larger friction naturally implies more measured misallocation—as well as earlier results suggesting a close connection between chain length and the dynamic properties of the model (see Figure 10). Finally, it should be noted that,

Figure 12. Varying the degree of replacement hiring



Notes. Model outcomes implied by targeting annual net inaction rates of 0, 10.8, 27.1, and 38.4 percent, and recalibrating the parameters of the baseline model as per Table 1.

although we have targeted a net inaction rate of 27.1 percent, even a rate of 10.8 percent implies a substantial increase in volatility and persistence: The volatility of the job-finding rate nearly doubles when C is chosen to target the smaller rate of inaction, and the half-life of unemployment increases sixfold. In this sense, the results of our baseline analysis appear to be robust to reasonable variations in the degree of replacement hiring.

5.2 Offer matching

The baseline model explored in the preceding sections addressed the challenge of wage determination by invoking a simple model of wage bargaining in which firms cannot credibly match outside offers received by their workers. As we have seen, an advantage

of this approach is that it admits an analysis of wage outcomes that can be confronted with available data. But we now show that many of the insights of the baseline case are preserved in an extension of the model to accommodate offer matching using a generalization of the sequential auctions approach of Postel-Vinay and Robin (2002) developed in our companion paper (Elsby and Gottfries 2021).

Before we describe the solution, a few aspects of the environment should be noted. First, it is important in this case that gross hiring costs are sunk at the point of recruitment; otherwise, firms will prefer to hire unemployed workers to save on recruitment bonuses. We therefore assume that gross hiring costs take the form of a linear vacancy cost. Second, to simplify the contract structure, we assume that firms can commit to payments to workers only in the current dt period (as in Moscarini 2005). Consequently, recruitment and retention compensation must be delivered as instantaneous bonuses, and workers within a given firm are almost always paid the same flow wage.[19] Finally, we focus on an equilibrium in which workers with outside offers move to firms with higher marginal values in cases in which they are indifferent.[20]

Consider a worker employed in a firm with marginal value Π_n. Note that the firm's maximum willingness to pay for the worker is given by its turnover/replacement cost, $\min\{\Pi_n, c\}$. At rate $s\lambda$ the worker receives an outside offer from a firm with marginal value denoted $\widetilde{\Pi}_n$ and, thereby, willingness to pay $\min\{\widetilde{\Pi}_n, c\}$. If $\Pi_n < \widetilde{\Pi}_n$, she quits to the outside firm, and receives a recruitment bonus equal to the turnover/replacement cost of the present firm, $\min\{\Pi_n, c\}$. If $\Pi_n > \widetilde{\Pi}_n$, she remains at her current firm, and receives a retention bonus equal to the turnover/replacement cost of the outside firm, $\min\{\widetilde{\Pi}_n, c\}$. As in Postel-Vinay and Robin, in the absence of an outside offer, she receives a flow wage w that renders her indifferent to unemployment; equivalently, her worker surplus over unemployment, which we denote by W, is set equal to zero.

Retracing the steps that led to (8) in the baseline model, we can write the firm's value in this case as

$$r\Pi = xn^{\alpha} - wn - \delta n \min\{\Pi_n, c\} - (s\lambda - \delta)n\mathbb{E}\left[\min\{\widetilde{\Pi}_n, c\} \,|\widetilde{\Pi}_n < \Pi_n\right] + \mu x \Pi_x + \frac{1}{2}\sigma^2 x^2 \Pi_{xx}. \tag{43}$$

[19] It is well known that models with offer matching determine values, but not how these are delivered.

[20] We will see that cases of indifference arise in the replacement region where, in equilibrium, firms face the same marginal replacement costs. These ties can be remedied by, for example, the presence of an arbitrarily-small increase in the vacancy posting cost as the marginal product (or the marginal value of labor) rises.

Note that the firm now anticipates capital losses associated with payment of retention bonuses: At rate $(s\lambda - \delta)$, a contacted employee is retained, at the cost of an expected bonus equal to $\mathbb{E}\left[\min\left\{\widetilde{\Pi}_n, c\right\} | \widetilde{\Pi}_n < \Pi_n\right]$. In turn, the worker surplus W satisfies

$$rW = w - b + \delta \min\{\Pi_n, c\} + (s\lambda - \delta)\mathbb{E}\left[\min\left\{\widetilde{\Pi}_n, c\right\} | \widetilde{\Pi}_n < \Pi_n\right] - \delta n W_n + \mu x W_x + \frac{1}{2}\sigma^2 x^2 W_{xx}. \tag{44}$$

The worker receives the flow wage w net of the flow payoff to unemployment b. At rate δ, she quits, and receives a recruitment bonus equal to the replacement cost of her previous employer, $\min\{\Pi_n, c\}$. At rate $(s\lambda - \delta)$, she remains with her current firm, and receives in expectation a retention bonus equal to the replacement cost of the outside employer, $\mathbb{E}\left[\min\left\{\widetilde{\Pi}_n, c\right\} | \widetilde{\Pi}_n < \Pi_n\right]$. Finally, absent an outside offer, she faces capital gains associated with the evolution of her firm's employment n, and productivity x.

Recall that the flow wage paid in the absence of an outside offer solves $W = 0$, so that

$$w = b - \delta \min\{\Pi_n, c\} - (s\lambda - \delta)\mathbb{E}\left[\min\left\{\widetilde{\Pi}_n, c\right\} | \widetilde{\Pi}_n < \Pi_n\right]. \tag{45}$$

It follows that the firm's value takes the simpler form,

$$r\Pi = xn^{\alpha} - bn + \mu x \Pi_x + \frac{1}{2}\sigma^2 x^2 \Pi_{xx}. \tag{46}$$

Intuitively, in the presence of offer matching, workers anticipate recruitment and retention bonuses in the future, and thereby are willing to accept lower flow wages in the present. Firms thus effectively recoup the entirety of their turnover and retention costs.

Crucially, labor market equilibrium takes a similar form to that in the baseline model. The only qualitative difference is that the presence of offer matching eliminates the costs of turnover and replacement in (46) (in contrast to (8)).[21] Lemma 3 summarizes.

Lemma 3 *Suppose there is offer matching, and that firms are subject to a linear vacancy cost c_v. Then, (i) the marginal value in the natural wastage and replacement regions is as reported in Lemma 1 and Proposition 1 with $\omega_0 = b$, $\omega_1 = 0$, $s\lambda = 0$, and $c = \left[c_v + \int_{m_l}^{m_h} J(\tilde{m}) dq(\tilde{m})\right] / q(m_h)$. (The marginal value takes the same form in both regions.) (ii) The quit, hiring and vacancy-filling rates in the natural wastage and replacement regions are as in Lemma 1 and Proposition 1. (iii) The expansion region is degenerate.*

[21] Importantly, turnover continues to flow from low-m to high-m firms, and firms continue to face an effective gross hiring cost c, and net expansion cost C. The sole difference is that offer matching eliminates the *costs* of turnover and replacement faced by firms in (46).

Table 4. Amplitude of labor market stocks and flows: Model with offer matching

Moment	Model (with offer matching)	
	$C > 0$	$C = 0$
Response relative to output per worker		
Unemployment rate	10.3	4.5
Vacancy rate	10.3	4.5
U-to-E rate	7.4	1.2
E-to-U rate	3.6	3.6
E-to-E/Quit rate	5.8	1.2

Notes. Model outcomes are (absolute values of) steady-state elasticities with respect to output per worker.

As in the baseline case, the model can thus be solved analytically. Indeed, the structure of equilibrium is simpler. The absence of turnover and replacement costs in (46) implies not only that the marginal value takes the same form in both natural wastage and replacement regions, but also that the expansion region is degenerate.

It is natural to question the quantitative implications of the offer matching model. Lemma 3 provides two important perspectives on this. First, it implies that there exists a (re)calibration of the offer matching model such that its steady-state equilibrium is quantitatively indistinguishable from that implied by the baseline calibration in Table 1. Holding fixed the baseline parameters (*modulo* the restrictions implied by Lemma 3), one can choose the vacancy cost c_v, the flow payoff from unemployment b, and the expansion cost C to replicate the boundaries m_l, m_h, and m_e implied by the baseline calibration. Moreover, since the expansion region under the latter is small, $m_e \approx m_u$, remaining equilibrium outcomes—the layoff rate ς, hiring rate $\eta(m)$, quit rate $\delta(m)$, and vacancy-filling rate $q(m)$—will quantitatively be almost identical to those in the baseline calibration. (There is a qualitative difference, however, since the parameters ω_0, c, and C required to generate the same boundaries will differ.) In this precise sense, the cross-sectional content of both models of wage determination is nearly equivalent.

A second implication of Lemma 3 pertains to the source of labor market equilibration. In the baseline model, this is achieved through the response of *turnover costs* to labor market tightness, as summarized by the slope of the job creation condition in Figure 8. Offer matching, by contrast, eliminates turnover costs in (46); instead, the labor market equilibrates through the response of *recruitment costs* to labor market tightness, captured by the expression for c in Lemma 3. It follows that the response of the offer matching

model to changes in aggregate labor productivity will differ. Table 4 reports the results of calibrations of the offer matching model, with and without an expansion cost, analogous to Table 3 for the baseline case.[22] While offer matching moderates responses in general, the presence of the expansion cost approximately doubles the volatility of unemployment and amplifies the volatility of the job-finding rate to an even greater extent, with orders of magnitude broadly in the neighborhood of the data in Table 3.

Taken together, the results of the present paper are thus both qualitatively and quantitatively robust to the structure of wage determination.

6. Summary and discussion

This paper offers three contributions to our understanding of labor markets. First, it documents a set of stylized facts on the empirical prevalence of replacement hiring. Employers often hire to replace workers who quit. Replacement hiring leaves a clear imprint on establishment dynamics, epitomized by long spells of inaction in net employment adjustment, despite substantial intervening turnover.

Second, the paper proposes a model in which these establishment-level facts can be interpreted. The model captures the interaction of a novel structure of frictions, blending firm dynamics with on-the-job search, and a sunk cost of job creation. It admits an analytical characterization of steady-state labor market equilibrium, surmounting the associated technical challenge of finding a fixed point of the distributions of job values. The key novel implication of the model is that, as in the data, many firms choose to hire solely to replace their quits. This behavior propagates vacancy chains. The poaching of one worker triggers a cascade of further hires among the many firms that seek to maintain their employment.

Third, the paper traces out the quantitative implications of vacancy chains for the aggregate labor market. When calibrated to replicate the stylized facts of replacement hiring that we document, the implied sunk costs of job creation are substantial. In turn, the calibrated model further captures many salient cross-sectional and business cycle facts. The substantial job creation friction that underlies the vacancy chain naturally gives rise to considerable dispersion in productivity across employers, mirroring the data. Furthermore, the presence of vacancy chains both amplifies and propagates labor market

[22] Since worker bargaining power is set equal to zero in the sequential auctions model, we no longer target the average wage gains from on-the-job search in this case.

dynamics. Intuitively, vacancy chains grow progressively shorter in recessions, as more workers accumulate in shrinking firms or in unemployment. The implied quantitative magnitudes of both the volatility and persistence of job creation and unemployment bear a close resemblance to their empirical counterparts.

Returning to the themes that motivated this paper, an interpretation of our results is that they evoke the presence of a form of *capital*, one that is embodied in the creation of new jobs. Its existence follows from the lengths to which firms will go to replace workers who quit. When filtered through the model, the data imply investments into this form of capital that are substantial, corresponding to as much as three years' pay. This observation prompts several questions for future work. First, returning to our original motivation, what form does this capital take? Are its origins in physical capital—e.g. an unused machine? Or organizational capital—e.g. the blueprint of tasks in the firm? Second, and relatedly, how does this capital depreciate? Our model has linked the latter to firm shrinkage, such that the capital is implicitly specific to the scale of the firm. We believe further empirical work can elucidate these questions. For example, comprehensive matched worker-firm microdata would allow more direct *measurement* of vacancy chains, by tracing the movement of workers across firms. Moreover, combining these data with further information—on the occupational structure within firms, or their investment in physical capital, for example—would in turn provide insights into the origins of this form of capital. We hope the present paper will stimulate further research along these lines.

Appendix

The Appendix is organized as follows. Section A establishes the results presented in the main text. We work through these sequentially, region by region. Thus, we first present Lemma 1 on the *natural wastage region*; then our main results, Propositions 1 and 2, on the *replacement region*; and finally, Lemma 2 on the *expansion region*. As noted in the main text, it is possible for a fourth, *partial replacement* region to exist. Although our quantitative analyses have found the latter to be degenerate for the wide range of empirically-relevant parameter values we have explored, the proofs of Lemma 1 and Proposition 1 nonetheless address how the structure of the natural wastage and full replacement regions depends on the presence of a partial replacement region. A complete characterization of the latter is then presented later in section B. Details of the computational methods used for our quantitative analyses are provided in section C. We conclude in section D with additional quantitative results.

A. Proofs of Lemmas and Propositions

Proof of Lemma 1. Since $\delta(m) = s\lambda$, and therefore $\delta'(m) = 0$, in the natural wastage region, the Bellman equation for the firm's marginal value takes the form in (11). We seek a solution for $J(m)$ that satisfies the latter and two pairs of boundary conditions. First, the value-matching conditions,

$$J(m_l) = 0, \text{ and, } J(m_h) = c; \tag{47}$$

and, second, the smooth-pasting conditions,

$$J'(m_l) = 0, \text{ and, } J'(m_h^-) = J'(m_h^+) = \kappa. \tag{48}$$

We shall see that κ is determined by whether a partial replacement region exists. If it does, $\kappa = 0$; if not, κ will be determined after characterizing the firm's marginal value in the replacement region in Proposition 1.

It can be verified that the stated solution in (12) satisfies (11). It remains to infer the coefficients J_1 and J_2, and the boundaries m_l and m_h, that satisfy the boundary conditions in (47) and (48). In what follows, we verify that these can be recovered using an extension of the method devised by Abel and Eberly (1996).

The smooth-pasting conditions in (48) imply the coefficients

$$J_1 = -\frac{(1-\omega_1)\vartheta(\mathscr{g};\kappa)m_l^{1-\gamma_1}}{\gamma_1\rho(1)}, \text{ and, } J_2 = -\frac{(1-\omega_1)[1-\vartheta(\mathscr{g};\kappa)]m_l^{1-\gamma_2}}{\gamma_2\rho(1)}, \tag{49}$$

where $\mathscr{g} \equiv m_h/m_l$, and

$$\vartheta(\mathscr{g};\kappa) \equiv \frac{\mathscr{g}^{\gamma_2} - \left[1 - \frac{\rho(1)\kappa}{1-\omega_1}\right]\mathscr{g}}{\mathscr{g}^{\gamma_2} - \mathscr{g}^{\gamma_1}}. \tag{50}$$

Together with the value-matching conditions in (47), these yield the following implicit solutions for the boundaries m_l and m_h,

$$\frac{(1-\omega_1)m_l}{\rho(1)}\left\{1 - \frac{\vartheta(\mathscr{g};\kappa)}{\gamma_1} - \frac{1-\vartheta(\mathscr{g};\kappa)}{\gamma_2}\right\} = \frac{\omega_0}{\rho(0)}, \tag{51}$$

and

$$\frac{(1-\omega_1)m_h}{\rho(1)}\left\{1 - \mathscr{g}^{\gamma_1-1}\frac{\vartheta(\mathscr{g};\kappa)}{\gamma_1} - \mathscr{g}^{\gamma_2-1}\frac{1-\vartheta(\mathscr{g};\kappa)}{\gamma_2}\right\} = c + \frac{\omega_0}{\rho(0)}. \tag{52}$$

The latter provides a solution for the coefficients J_1 and J_2, and the boundaries m_l and m_h, for a given $\kappa \equiv J'(m_h)$. Recall that if a partial replacement region does exist, $\kappa = 0$, and thus the latter equations complete the solution for the boundaries and the marginal value function under natural wastage. We address the case in which a partial replacement region does not exist below.

The remaining results follow from the fact that the marginal product m evolves according to a geometric Brownian motion in the natural wastage region, as noted in (16). The solution for the separation rate into unemployment ς applies standard results on geometric Brownian motion at a reflecting boundary, in this case the lower boundary m_l. The solution for the worker distribution $G(m)$ follows from a canonical implication of geometric Brownian motion that implied stationary distributions follow a power law. See Proposition 3 in Elsby and Gottfries (2021) for example.

Proof of Proposition 1. The firm's marginal value satisfies (23) in the replacement region. We seek a solution for $J(m)$ that satisfies the latter and two pairs of boundary conditions. First, the value-matching conditions,

$$J(m_p) = c, \text{ and, } J(m_e) = c + C; \tag{53}$$

and, second, the smooth-pasting conditions,

$$J'(m_p^+) = J'(m_p^-) = \kappa, \text{ and, } J'(m_e) = 0. \tag{54}$$

If a partial replacement region does exist, $m_h < m_p$ and $\kappa = 0$. If a partial replacement region does not exist, $m_h = m_p$ and κ will be recovered using $J'(m_h^-)$, evaluated in the natural wastage region, and $J'(m_h^+)$, evaluated in the full replacement region (see below).

For ease of reference, we restate here the solution in (24),

$$J(m) = \frac{(1-\omega_1)m}{\varrho(1)} - \frac{\omega_0}{\varrho(0)} - \mathcal{J}_0(m) + \mathcal{J}_1 m^{\tilde{\gamma}_1} + \mathcal{J}_2 m^{\tilde{\gamma}_2}. \tag{55}$$

It can be verified that the latter is the general solution of (23), where $\mathcal{J}_0(m)$ is a particular solution that satisfies

$$r\mathcal{J}_0(m) = D(m) + \mu m \mathcal{J}_0'(m) + \frac{1}{2}\sigma^2 m^2 \mathcal{J}_0''(m), \tag{56}$$

and

$$D(m) \equiv c[\delta(m) - (1-\alpha)m\delta'(m)]. \tag{57}$$

Applying the method of variation of parameters yields

$$\mathcal{J}_0(m) = A(m)m^{\tilde{\gamma}_1} + B(m)m^{\tilde{\gamma}_2}, \tag{58}$$

where

$$A(m) = \frac{1}{\sigma^2/2}\int \frac{1}{\mathcal{W}} m^{\tilde{\gamma}_2 - 2} D(m)\mathrm{d}m\,, \text{ and, } B(m) = -\frac{1}{\sigma^2/2}\int \frac{1}{\mathcal{W}} m^{\tilde{\gamma}_1 - 2} D(m)\mathrm{d}m, \tag{59}$$

and $\mathcal{W}$ is the Wronskian,

$$\mathcal{W} \equiv \begin{vmatrix} m^{\tilde{\gamma}_1} & m^{\tilde{\gamma}_2} \\ \tilde{\gamma}_1 m^{\tilde{\gamma}_1 - 1} & \tilde{\gamma}_2 m^{\tilde{\gamma}_2 - 1} \end{vmatrix} = (\tilde{\gamma}_2 - \tilde{\gamma}_1) m^{\tilde{\gamma}_1 + \tilde{\gamma}_2 - 1}. \tag{60}$$

Substituting, integrating by parts, and defining $\mathcal{J}_0(m_p) \equiv 0$, yields the stated solution,

$$\mathcal{J}_0(m; m_p) = c\frac{1-\alpha}{\sigma^2/2}\int_{m_p}^{m} \left[\varpi \left(\frac{m}{\tilde{m}}\right)^{\psi_1} + (1-\varpi)\left(\frac{m}{\tilde{m}}\right)^{\psi_2}\right] \frac{\delta(\tilde{m})}{\tilde{m}}\mathrm{d}\tilde{m}, \tag{61}$$

where we have emphasized its dependence on m_p, and $\varpi \equiv [1/(1-\alpha) - \psi_1]/(\psi_2 - \psi_1)$. To confirm that the latter weight is in the unit interval, note that $\varrho(0) = \varrho(1/(1-\alpha)) = r > 0$. Thus $\psi_1 < 0$, and $\psi_2 > 1/(1-\alpha)$.

To infer the coefficients $\mathcal{J}_1$ and $\mathcal{J}_2$, and the relevant boundaries, that satisfy the boundary conditions in (53) and (54) we again extend the method of Abel and Eberly (1996). The smooth-pasting conditions in (54) imply the coefficients

$$\mathcal{J}_1 = -\frac{(1-\omega_1)\Theta_1(\mathcal{G};\kappa,m_p)m_p^{1-\tilde{\gamma}_1}}{\tilde{\gamma}_1\varrho(1)}, \text{ and, } \mathcal{J}_2 = -\frac{(1-\omega_1)\Theta_2(\mathcal{G};\kappa,m_p)m_p^{1-\tilde{\gamma}_2}}{\tilde{\gamma}_2\varrho(1)}, \quad (62)$$

where $\mathcal{G} \equiv m_e/m_p$,

$$\Theta_1(\mathcal{G};\kappa,m_p) \equiv \frac{\left[1-\varrho(1)\frac{\kappa-\mathcal{J}_0'(m_p;m_p)}{1-\omega_1}\right]\mathcal{G}^{\tilde{\gamma}_2} - \left[1+\varrho(1)\frac{\mathcal{J}_0'(\mathcal{G}m_p;m_p)}{1-\omega_1}\right]\mathcal{G}}{\mathcal{G}^{\tilde{\gamma}_1}-\mathcal{G}^{\tilde{\gamma}_1}}, \quad (63)$$

and

$$\Theta_2(\mathcal{G};\kappa,m_p) \equiv \frac{\left[1+\varrho(1)\frac{\mathcal{J}_0'(\mathcal{G}m_p;m_p)}{1-\omega_1}\right]\mathcal{G} - \left[1-\varrho(1)\frac{\kappa-\mathcal{J}_0'(m_p;m_p)}{1-\omega_1}\right]\mathcal{G}^{\tilde{\gamma}_1}}{\mathcal{G}^{\tilde{\gamma}_2}-\mathcal{G}^{\tilde{\gamma}_1}}. \quad (64)$$

Together with the value-matching conditions in (53), these yield the following implicit solutions for the boundaries m_p and m_e,

$$\frac{(1-\omega_1)m_p}{\varrho(1)}\left\{1-\frac{\Theta_1(\mathcal{G};\kappa,m_p)}{\tilde{\gamma}_1}-\frac{\Theta_2(\mathcal{G};\kappa,m_p)}{\tilde{\gamma}_2}\right\} = c+\frac{\omega_0}{\varrho(0)}, \quad (65)$$

and

$$\begin{aligned}\frac{(1-\omega_1)m_e}{\varrho(1)}&\left\{1-\mathcal{G}^{\tilde{\gamma}_1-1}\frac{\Theta_1(\mathcal{G};\kappa,m_e/\mathcal{G})}{\tilde{\gamma}_1}-\mathcal{G}^{\tilde{\gamma}_2-1}\frac{\Theta_2(\mathcal{G};\kappa,m_e/\mathcal{G})}{\tilde{\gamma}_2}\right\} \\ &= c+C+\frac{\omega_0}{\varrho(0)}-\mathcal{J}_0(m_e;m_e/\mathcal{G}).\end{aligned} \quad (66)$$

In the case in which a partial replacement region does not exist, $m_p = m_h$, (65) and (66) implicitly determine m_h and m_e as functions of $\mathcal{G} \equiv m_e/m_h$ for a given κ. The ratio of these implicit solutions then yields a fixed point in $\mathcal{G}$. Solution of that fixed point then recovers m_h and m_e, which in turn determine the coefficients $\mathcal{J}_1$ and $\mathcal{J}_2$ in (62). It then remains to determine κ. This follows from the smooth-pasting condition, $J'(m_h^-) = J'(m_h^+)$. Using (49) and (62) above, this can be written as

$$\begin{aligned}\frac{1-\omega_1}{\rho(1)}&\{1-\mathcal{g}^{\gamma_1-1}\vartheta(\mathcal{g};\kappa)-\mathcal{g}^{\gamma_2-1}[1-\vartheta(\mathcal{g};\kappa)]\} \\ &= \frac{1-\omega_1}{\varrho(1)}\{1-\Theta_1(\mathcal{G};\kappa,m_h)-\Theta_2(\mathcal{G};\kappa,m_h)\}+\mathcal{J}_0'(m_h;m_h).\end{aligned} \quad (67)$$

This implicitly determines a solution for κ, and thereby a solution for the full system of coefficients $\{J_1, J_2, \mathcal{J}_1, \mathcal{J}_2\}$ and boundaries $\{m_l, m_h, m_e\}$.

In the case in which a partial replacement region does exist, $m_p > m_h$, (65) and (66) implicitly determine m_p and m_e as functions of $\mathcal{G} \equiv m_e/m_p$. The ratio of these implicit solutions then yields a fixed point in $\mathcal{G}$. Solution of that fixed point then recovers m_p and m_e, which in turn determine the coefficients $\mathcal{J}_1$ and $\mathcal{J}_2$ in (62).

The solution for the gross hiring rate $\eta(m)$ and quit rate $\delta(m)$ in (27) follows from the fact that they are equal in the replacement region. Applying Proposition 3 from Elsby and Gottfries (2021), this requires that

$$\eta(m) = -\frac{\sigma^2/2}{1-\alpha}\frac{m\delta'(m)}{\delta(m)} = \delta(m), \tag{68}$$

subject to the boundary condition $\delta(m_p)$. It is straightforward to verify that the stated solution (27) satisfies these. Note that, if a partial replacement region does not exist, $m_p = m_h$, $\delta(m_p) = \delta(m_h) = s\lambda$. If a partial replacement region does exist, $m_p > m_h$, and $\delta(m_p)$ is given by (80) in Lemma 4 (see below).

The solution for the vacancy-filling rate in (28) follows from the solution for the quit rate $\delta(m)$ in (27), and application of Proposition 3 in Elsby and Gottfries (2021),

$$q(m) \propto \exp\left[\frac{1-\alpha}{\sigma^2/2}\int^m \frac{\delta(\tilde{m})}{\tilde{m}}\mathrm{d}\tilde{m}\right] = 1 + \frac{1-\alpha}{\sigma^2/2}\delta(m_p)\ln\left(\frac{m}{m_p}\right) \propto \frac{1}{\delta(m)}. \tag{69}$$

Proof of Proposition 2. The expected remaining hiring chain length at m, $\ell_H(m)$, satisfies the recursion

$$\ell_H(m) = 1 + \frac{s(1-u)}{u + s(1-u)G(m)}\int_{m_h}^m \ell_H(\tilde{m})g(\tilde{m})d\tilde{m}. \tag{70}$$

Each hire at m poaches a measure $s(1-u)G(m)/[u + s(1-u)G(m)]$ of employed workers, each of which induces a further hire, with a chain of length $\ell_H(\tilde{m})$, for all $\tilde{m} < m$ in the replacement region. Recalling that $q(m) = \chi[\psi + (1-\psi)G(m)]$, and that $\psi = u/[u + s(1-u)]$,

$$\ell_H(m) = 1 + \frac{1}{q(m)}\int_{m_h}^m \ell_H(\tilde{m})q'(\tilde{m})d\tilde{m}. \tag{71}$$

Rearranging and differentiating,

$$q'(m)[\ell_H(m) - 1] + q(m)\ell'_H(m) = \ell_H(m)q'(m). \tag{72}$$

Cancelling and rearranging implies that $\ell'_H(m) = q'(m)/q(m)$. Given the boundary condition $\ell_H(m_h) = 1$, the solution is as stated.

The expected remaining vacancy chain length at m, $\ell_V(m)$, satisfies the recursion

$$\ell_V(m) = 1 + \frac{s(1-u)}{u + s(1-u)G(m)} \int_{m_h}^{m} \frac{q(m)}{q(\widetilde{m})} \ell_V(\widetilde{m}) g(\widetilde{m}) d\widetilde{m}. \tag{73}$$

Each vacancy at m poaches a measure $q(m)\, s(1-u)G(m)/[u + s(1-u)G(m)]$ of employed workers, each of which induces $1/q(\widetilde{m})$ further vacancies, with chains of length $\ell_V(\widetilde{m})$, for all $\widetilde{m} < m$ in the replacement region. Recalling that $q(m) = \chi[\psi + (1-\psi)G(m)]$, and that $\psi = u/[u + s(1-u)]$,

$$\ell_V(m) = 1 + \int_{m_h}^{m} \ell_V(\widetilde{m}) \frac{q'(\widetilde{m})}{q(\widetilde{m})} d\widetilde{m}. \tag{74}$$

Differentiating, $\ell_V'(m) = \ell_V(m)\, q'(m)/q(m)$. Given the boundary condition $\ell_V(m_h) = 1$, the solution is as stated.

Proof of Lemma 2. In the expansion region, the quit rate satisfies the differential equation stated in (17). Its solution is given by

$$\delta(m) = \frac{(1-\omega_1)m}{\alpha c} - \frac{\omega_0 + r(c+C)}{c} + \delta_1 m^{\frac{1}{1-\alpha}}. \tag{75}$$

The coefficient δ_1, and the upper boundary m_u, are determined by the boundary conditions,

$$\delta(m_e) = \delta(m_p)\left[1 + \frac{1-\alpha}{\sigma^2/2}\delta(m_p)\ln\left(\frac{m_e}{m_p}\right)\right]^{-1}, \text{ and, } \delta(m_u) = 0. \tag{76}$$

It can be verified that the stated solution in (18) satisfies these.

The solutions for the gross hiring rate $\eta(m)$, and the vacancy-filling rate $q(m)$, follow directly from Proposition 3 in Elsby and Gottfries (2021).

Proof of Lemma 3. (i) Given the firm value in (46), the marginal value of labor to the firm $J \equiv \Pi_n$ can be written

$$J(m) = m - b + \mu m J'(m) + \frac{1}{2}\sigma^2 m^2 J''(m) \tag{77}$$

in both the natural wastage and replacement regions. It follows that both Lemma 1 and Proposition 1 hold *mutatis mutandis* with ω_0, ω_1, and $s\lambda$ exchanged respectively with b, 0, and 0. A corollary is that there is a common solution for the marginal value $J(m)$ that holds in both the natural wastage and replacement regions.

Now consider the effective cost per hire. This is equal to the sum of the vacancy cost and the expected recruitment bonus as a ratio of the vacancy-filling rate. We will show that this is a constant, equal to c, for hiring firms. For now, allow it to be a function of the marginal product, $c(m)$. Then we can write,

$$c(m) = \begin{cases} \dfrac{c_v + \int_{m_l}^{m} J(\tilde{m})\mathrm{d}q(\tilde{m})}{q(m)} & \text{if} \quad m < m_h, \\ \dfrac{c_v + \int_{m_l}^{m_h} J(\tilde{m})\mathrm{d}q(\tilde{m}) + \int_{m_h}^{m} c(\tilde{m})\mathrm{d}q(\tilde{m})}{q(m)} & \text{if} \quad m \geq m_h. \end{cases} \tag{78}$$

Since, by definition, $J(m) < c(m)$ for all $m < m_h$, and $J(m_h) = c(m_h)$, it can be verified that $c'(m) < 0$ for all $m < m_h$, and $c'(m_h) = 0$, confirming that no firm with $m < m_h$ will wish to hire. Furthermore, the effective hiring cost for hiring firms is a constant, $c(m) = c$ for all $m \geq m_h$, as otherwise a hiring firm with a higher cost of replacement could hire using the same strategy as a firm with a low cost of replacement. This confirms the stated result,

$$c = \frac{c_v + \int_{m_l}^{m_h} J(\tilde{m})\mathrm{d}q(\tilde{m})}{q(m_h)}. \tag{79}$$

(ii) follows from the fact that the proofs for the quit, hiring, and vacancy-filling rates apply *mutatis mutandis.*

(iii) follows from the absence of turnover/replacement costs in the firm's value (46).

B. Partial replacement region

We noted that it is possible, though quantitatively implausible given our targeted moments, for a *partial replacement* region to exist. For completeness, we provide a characterization of that region here.

Lemma 4 *In any nondegenerate partial replacement region,* $m \in (m_h, m_p)$*, (i) the firm's marginal value* $J(m) = c$*. (ii) The quit rate is given by*

$$\begin{aligned} \delta(m) = s\lambda + \frac{1}{c}\Bigg\{ & \frac{(1-\omega_1)(m-m_h)}{\alpha} \\ & - \left[\frac{(1-\omega_1)m_h}{\alpha} - \omega_0 - (r+s\lambda)c\right]\left[\left(\frac{m}{m_h}\right)^{\frac{1}{1-\alpha}} - 1\right]\Bigg\}, \end{aligned} \tag{80}$$

is strictly decreasing and concave. (iii) The gross hiring rate is given by

$$\eta(m) = -\frac{\sigma^2/2}{1-\alpha}\frac{m\delta'(m)}{\delta(m)}. \tag{81}$$

(iv) The vacancy-filling rate is given by

$$q(m) = \chi\psi \exp\left[\frac{1-\alpha}{\sigma^2/2}\int_{m_h}^{m}\frac{\delta(\tilde{m})}{\tilde{m}}\mathrm{d}\tilde{m}\right]. \tag{82}$$

Proof of Lemma 4. In any nondegenerate partial replacement region, the Bellman equation for the firm's marginal value takes the form

$$\begin{aligned} rJ(m) = (1-\omega_1)m - \omega_0 - [\delta(m) - (1-\alpha)m\delta'(m)]J(m) \\ + [\mu + (1-\alpha)\delta(m)]mJ'(m) + \frac{1}{2}\sigma^2 m^2 J''(m). \end{aligned} \tag{83}$$

Since in any partial replacement region $J(m) = c$, and therefore $J'(m) = J''(m) = 0$, the latter becomes a differential equation in the quit rate,

$$rc = (1-\omega_1)m - \omega_0 - [\delta(m) - (1-\alpha)m\delta'(m)]c. \tag{84}$$

Its solution is given by

$$\delta(m) = \frac{(1-\omega_1)m}{\alpha c} - \frac{\omega_0 + rc}{c} + \delta_1 m^{\frac{1}{1-\alpha}}. \tag{85}$$

The coefficient δ_1 is determined by the boundary condition

$$\delta(m_h) = s\lambda. \tag{86}$$

This yields the solution for the quit rate in (80).

The solutions for the gross hiring rate $\eta(m)$ in (81), and the vacancy-filling rate $q(m)$ in (82), follow directly from Proposition 3 in Elsby and Gottfries (2021).

C. Computational Appendix

The model admits closed form solutions for most of the equilibrium objects. These are used for much of the steady-state analysis. To compute the net inaction rate, however, we use the binominal approximation of a Brownian motion. In addition, the cross-sectional results presented in Table 2 are computed by simulating sample paths of firms.

Where we must depart further from the analytical solutions is in the solution for the transition dynamics following an MIT shock. To solve for these, we rely on finite difference methods, similar to the approach taken in Elsby and Gottfries (2021) in a related

application. Their solution method has to be extended here because agents need to forecast their expected future replacement cost when choosing their optimal hires.

We start with the two partial differential equations for the out-of-steady-state Hamilton-Jacobi-Bellman (HJB) and Fokker-Planck (FPE) equations. The former is as reported in (36), which we restate here for convenience,

$$\begin{aligned} rJ_t(m) = (1-\omega_1)m - \omega_0 - [\delta_t(m) - (1-\alpha)m\delta_t'(m)]\min\{J_t(m), c\} \\ + \left[\mu + (1-\alpha)\delta_t(m)\mathbf{1}_{\{\mathrm{d}n^*<0\}}\right]mJ_t'(m) + \frac{1}{2}\sigma^2 m^2 J_t''(m) + \frac{\partial J_t(m)}{\partial t}. \end{aligned} \tag{87}$$

For the out-of-steady-state FPE, this turns out to be simplest if we define $\mathbb{G}_t(m) \equiv [u_t/s] + (1-u_t)G_t(m)$. Then, following steps analogous to those in Elsby and Gottfries (2021), the FPE can be written recursively in $\mathbb{G}_t(m)$ as

$$\begin{aligned} \frac{\partial \mathbb{G}_t(m)}{\partial t} = -\delta_t(m)\mathbb{G}_t(m) - \mu m \mathbb{g}_t(m) - (1-\alpha)m\frac{\partial}{\partial m}[\delta_t(m)\mathbb{G}_t(m)] \\ + \frac{1}{2}\sigma^2\frac{\partial}{\partial m}[m^2\mathbb{g}_t(m)] + \left(\frac{1}{s}-1\right)\left[\frac{\sigma^2/2}{1-\alpha}m_{lt}\mathbb{g}_t(m_{lt}) - s\lambda_t\mathbb{G}_t(m_{lt})\right]. \end{aligned} \tag{88}$$

These are both discretized and solved using the finite difference method. The FPE is solved forward using the fully implicit method. The marginal value function J is solved backwards iterating on the HJB equation. We impose the super contact condition via a penalty method (similar to Elsby and Gottfries 2021). This method results in a system of nonlinear equations for each time step for the HJB equation. In particular, to illustrate the penalty method, (89) presents the equation corresponding to the implicit scheme on an equi-spaced grid

$$\begin{aligned} rJ_t(m_i) = (1-\omega_1)m_i - \omega_0 - [\delta_t(m_i) - (1-\alpha)m_i\delta_t'(m_i)]\min\{J_t(m_i), c\} \\ + \left[\mu + (1-\alpha)\delta_t(m_i)\mathbf{1}_{\{\mathrm{d}n^*<0\}}\right]m_i\frac{J_t(m_{i+1}) - J_t(m_{i-1})}{2\Delta_i} \\ + \frac{1}{2}\sigma^2 m_i^2\frac{J_t(m_{i+1}) + J_t(m_{i-1}) - 2J_t(m_i)}{\Delta_i^2} + \frac{J_{t+\Delta_t}(m_i) - J_t(m_i)}{\Delta_t} \\ + \mathbf{1}_{\{J_t(m_i)<0\}}[0 - J_t(m_i)]P + \mathbf{1}_{\{J_t(m_i)>c+C\}}[c + C - J_t(m_i)]P, \end{aligned} \tag{89}$$

where P represents the penalty (a large positive number). Given a conjectured $\delta_t(m_i)$ and $J_{t+\Delta_t}(m_i)$, this is a system of nonlinear equations in $J_t(m_i)$ that can be solved at each time step. However, given our approach, the function $\delta_t(m_i)$ is not known, but instead only λ_t and $\mathbb{G}_t(m)$. We can calculate the quit function $\delta_t(m_i)$ using $\mathbb{G}_t(m)$, λ_t, and a guess for the hiring boundary m_{ht}. The full iteration over the HJB equation therefore involves a

guess of m_{ht}, after which we calculate $\delta_t(m_i)$. Thereafter we solve the system of nonlinear equations for $J_t(m_i)$ given $J_{t+\Delta_t}(m_i)$ and $\delta_t(m_i)$. We then calculate $\widehat{m}_{ht}$ such that $J_t(\widehat{m}_{ht}) = c$. If m_{ht} and $\widehat{m}_{ht}$ are sufficiently close, we stop; otherwise, we update our initial guess of m_{ht} and return to the HJB iteration. Note also that, to improve accuracy, computations of (89) are in fact based on the half-implicit (Crank Nicolson) scheme on a grid for the logarithm of the marginal product (rather than marginal product) that has more grid points around the boundaries (where the solution is more nonlinear).

We solve for the response of model outcomes to an aggregate shock by iterating over the path for the job finding rate λ until excess demand is sufficiently small. In particular, the algorithm then works using the following steps:

1. We solve for the job offer arrival rate λ in each steady state, as well as the marginal value function J, worker distribution $\mathbb{G}$, and the unemployment rate $u \equiv U/L$.

2. We make an initial guess for the transition path for the job offer arrival rate λ and the time path of the boundaries m_l, m_h and m_e.

3. Solving the fixed point of boundaries (given a path for λ).

    a. Impose MIT shock. Shift the distribution of workers across marginal products according to the sign of the aggregate shock (e.g., the distribution shifts left if p falls) and impose firing consistent with the conjectured lower reflecting boundary m_{lt}.

    b. Iterate the FPE forward. At each t, compute the quit rate $\delta_t(m)$ in the full replacement region, which can be done using the conjectured hiring boundary m_{ht}, and given that $\mathbb{G}_t(m)$ and λ_t are known. Evaluating at the conjectured expansion boundary m_{et} gives $\delta_t(m_{et})$. Using the latter, we can calculate m_{ut} and the full quit function $\delta_t(m)$. With this information, we can then solve forward for the worker distribution using the integrated FPE.

    c. We solve the marginal value function (HJB) equation backwards within the natural wastage and full replacement regions (given the path for the distribution $\mathbb{G}_t(m)$ that we solved for (in 3b) and the job finding rate λ_t). We then calculate updated boundaries m_{lt}, m_{ht}, and m_{et} by iterating on the HJB equation, as described above.

d. We calculate the difference between the updated boundaries from the HJB iteration and those conjectured. If the difference is small, we stop (and move to 4); otherwise, we update our conjecture for boundaries and return to 3a.

4. Lastly, we calculate excess demand. If excess demand is sufficiently small, we stop. Otherwise, we update the time path of λ based on each period's excess demand (and make a new conjecture for the time path for the boundaries), and return to step 3. We find that a sluggish updating rule, with relatively more updating in earlier periods, helps with stability of the solution.

We examine the accuracy of our numerical scheme by comparing its steady-state outcomes with our steady-state analytical results for the marginal value J and worker distribution G. In all cases, errors induced by the numerical scheme are very small.

D. Additional descriptive empirical impulse responses

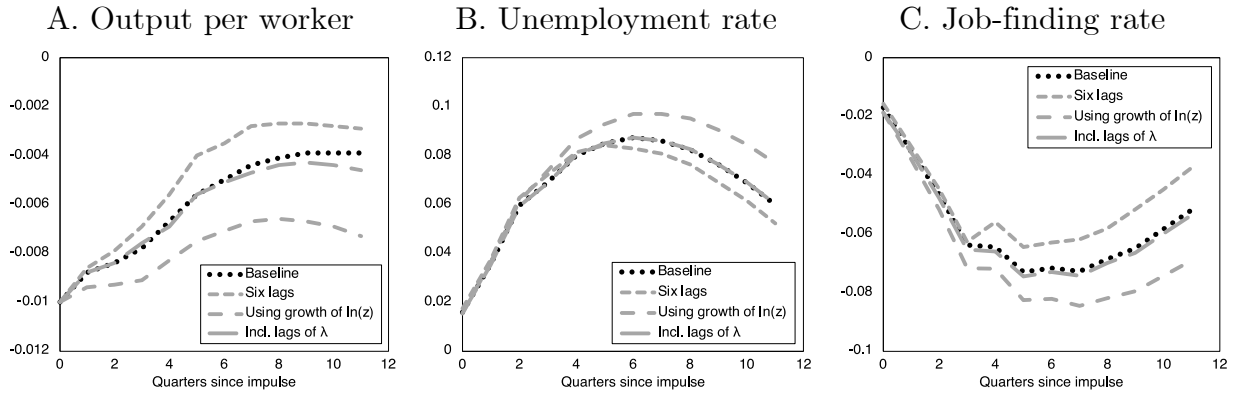
This appendix reports empirical impulse response functions for several variants of our baseline specification in section 4.2. The latter proceeded in two stages. First, in (41), we project log output per worker, $\ln z_t$, on its own lags, and lags of the log unemployment rate, $\ln u_{t-s}$. Then, in (42), we regress the log unemployment rate, $\ln u_t$, and log job-finding rate, $\ln \lambda_t$, on lags of each, as well as contemporaneous and lagged values of the residuals from the first stage, $\hat{e}_t^z$. In our baseline analysis, we set a lag length of $\mathcal{L} = 4$.

We consider several variations on these regressions. First, we re-estimate using six lags ($\mathcal{L} = 6$) in both stages, rather than four. Conventional lag length selection criteria (e.g., AIC, SIC) do not favor more than four lags, but these criteria may lead one to under-fit models in small samples (Nickelsburg 1985).[23] Second, we re-estimate using the growth of log output per worker, $\Delta \ln z_t$, in the place of its log level, $\ln z_t$, in the first stage (41). This specification is motivated by the observation that output per worker appears to have a unit root, and so the first difference renders it stationary. Finally, we examine the inclusion of lags of the job-finding rate, $\ln \lambda_{t-s}$, on the right-hand side of the first stage. This has the appealing quality of treating the first and second stages symmetrically, in that the same covariates are used in both. We did not pursue this latter specification

[23] However, Ivanov and Kilian (2005) find that the Schwarz Information Criterion (SIC) leads to the most accurate prediction of impulse response functions for sample sizes on the order of those we use.

in the main text solely because the regression is unstable when applied to data generated by the model with $C = 0$: The near-jump dynamics of output per worker z and the job-finding rate λ are nearly collinear in this case.

Figure D.1. Descriptive impulse responses in the data: Robustness



Notes. Impulse responses to a negative one percent innovation to output per worker implied by estimation of alternative specifications of (41) and (42) detailed in the text.

The results from these alternative specifications are displayed in Figure D.1. These bracket our baseline estimates. Furthermore, the responses of the unemployment and job-finding rates are very similar. If we consider six lags instead of four, the impulse responses of $\ln z$, $\ln u$, and $\ln \lambda$ are all somewhat *less* persistent. (A corollary is that fewer lags lead to more persistent responses.) If we instead enter the growth of output per worker into the regression (but use four lags), the impulse responses across the board are somewhat *more* persistent. Finally, the addition of lags of the job finding rate in the first stage has a negligible effect on the baseline estimates.

References

Abel, Andrew B. and Janice C. Eberly. 1996. "Optimal investment with costly reversibility." *Review of Economic Studies* 63(4): 581-593.

Acemoglu, Daron, and William Hawkins. 2014. "Search with multi-worker firms." *Theoretical Economics* 9(3): 583-628.

Akerlof, George A., Andrew Rose, and Janet Yellen. 1988. "Job switching and job satisfaction in the US labor market." *Brookings Papers on Economic Activity* 1988.2: 495-594.

Audoly, Richard. 2019. "Firm Dynamics and Random Search over the Business Cycle." Mimeo, University College London.

Barlevy, Gadi. 2008. "Identification of Search Models using Record Statistics." *Review of Economic Studies* 75(1): 29-64.

Barnichon, Regis. 2010. "Building a composite Help-Wanted Index." *Economics Letters* 109(3): 175-178.

Bartlesman, Eric, John Haltiwanger and Stefano Scarpetta. 2013. "Cross-Country Differences in Productivity: The Role of Allocation and Selection." *American Economic Review* 103(1): 305-334.

Bentolila, Samuel and Giuseppe Bertola. 1990. "Firing costs and labour demand: how bad is Eurosclerosis?" *Review of Economic Studies* 57(3): 381-402.

Bewley, Truman. 1999. *Why Wages Don't Fall During a Recession*, Harvard University Press.

Bilal, Adrien, Niklas Engbom, Simon Mongey and Giovanni L. Violante. 2019. "Firm and Worker Dynamics in a Frictional Labor Market." Mimeo, Princeton University.

Binmore, Ken, Ariel Rubinstein, and Asher Wolinsky. 1986. "The Nash bargaining solution in economic modelling." *RAND Journal of Economics* 17: 176-188.

Blanchard, Olivier, and Peter Diamond. 1989. "The Beveridge curve." *Brookings papers on Economic Activity* 1989(1): 1-76.

Bloom, Nicholas. 2009. "The Impact of Uncertainty Shocks." *Econometrica* 77(3): 623-685.

Boppart, Timo, Per Krusell, and Kurt 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.

Brown, Charles, and James L. Medoff. 1996. "Employer Characteristics and Work Environment." *Annales d'Économie et de Statistique* 41/42: 275-298.

Bruegemann, Bjoern, Pieter Gautier, and Guido Menzio. 2018. "Intra Firm Bargaining and Shapley Values." *Review of Economic Studies* 86: 564-592.

Burgess, Simon, Julia Lane and David Stevens. 2001. "Churning dynamics: an analysis of hires and separations at the employer level." *Labour Economics* 8(1): 1-14.

Burdett, Kenneth and Dale T. Mortensen. 1998. "Wage differentials, employer size, and unemployment." *International Economic Review* 39: 257-273.

Card, David, Ana Rute Cardoso, Joerg Heining, and Patrick Kline. 2018. "Firms and labor market inequality: Evidence and some theory." *Journal of Labor Economics* 36(S1): S13-S70.

Carrillo-Tudela, Carlos, Alex Clymo, and Melvyn G. Coles. 2021. "Equilibrium Job Turnover and the Business Cycle." Mimeo, University of Essex.

Chase, Ivan D. 1991. "Vacancy Chains." *Annual Review of Sociology* 17: 133-54.

Coles, Melvyn G. and Dale T. Mortensen. 2016. "Equilibrium labor turnover, firm growth, and unemployment." *Econometrica* 84(1): 347-363.

Contini, Bruno and Riccardo Revelli. 1997. "Gross flows vs. net flows in the labor market: What is there to be learned?" *Labour Economics* 4: 245-263.

Cooper, Russell, John Haltiwanger, Jonathan L. Willis. 2007. "Search frictions: Matching aggregate and establishment observations." *Journal of Monetary Economics* 54(1): 56-78.

Cooper, Russell, John Haltiwanger, Jonathan L. Willis. 2015. "Dynamics of labor demand: Evidence from plant-level observations and aggregate implications." *Research in Economics* 6(1): 37-50.

David, Joel M. and Venky Venkateswaran. 2019. "The Sources of Capital Misallocation." *American Economic Review* 109(7): 2531-2567.

Davis, Steven J., and John Haltiwanger. 1992. "Gross Job Creation, Gross Job Destruction, and Employment Reallocation." *Quarterly Journal of Economics* 107(3): 819-864.

Davis, Steven J., R. Jason Faberman and John C. Haltiwanger. 2012. "Labor market flows in the cross section and over time." *Journal of Monetary Economics* 59(1): 1-18.

Davis, Steven J., R. Jason Faberman and John C. Haltiwanger. 2013. "The establishment-level behavior of vacancies and hiring." *Quarterly Journal of Economics* 128(2): 581-622.

Di Addario, Sabrina, Patrick Kline, Raffaele Saggio, and Mikkel Solvsten. 2020. "It ain't where you're from, it's where you're at: hiring origins, firm heterogeneity, and wages." IRLE Working Paper No. 104-20.

Elsby, Michael W. L. and Axel Gottfries. 2021. "Firm Dynamics, On-the-Job Search and Labor Market Fluctuations." *Review of Economic Studies*, Forthcoming.

Elsby, Michael W. L. and Ryan Michaels. 2013. "Marginal Jobs, Heterogeneous Firms, and Unemployment Flows." *American Economic Journal: Macroeconomics* 5(1): 1-48.

Elsby, Michael W. L., Ryan Michaels and David Ratner. 2015. "The Beveridge Curve: A Survey." *Journal of Economic Literature* 53(3): 571-630.

Elsby, Michael W. L., Ryan Michaels and David Ratner. 2019. "Vacancy Chains." Mimeo, University of Edinburgh.

Elsby, Michael W. L., Donggyun Shin, and Gary Solon. 2016. "Wage Adjustment in the Great Recession and other Downturns: Evidence from the United States and Great Britain." *Journal of Labor Economics* 34(S1): S249-S291.

Faberman, R. Jason and Eva Nagypal. 2008. "Quits, Worker Recruitment, and Firm Growth: Theory and Evidence." Federal Reserve Bank of Philadelphia Working Paper No. 08-13.

Fallick, Bruce, and Charles A. Fleischman. 2004. "Employer-to-employer flows in the U.S. labor market: the complete picture of gross worker flows." Finance and Economics Discussion Series 2004-34, Board of Governors of the Federal Reserve System.

Foote, Chris. 1998. "Trend Employment Growth and the Bunching of Job Creation and Destruction." *Quarterly Journal of Economics* 113(3): 809-834.

Fujita, Shigeru, and Giuseppe Moscarini. 2017. "Recall and Unemployment." *American Economic Review* 107(12): 3875-3916.

Fujita, Shigeru, and Makoto Nakajima. 2016. "Worker flows and job flows: A quantitative investigation." *Review of Economic Dynamics* 22: 1-20.

Fujita, Shigeru, and Garey Ramey. 2007. "Job matching and propagation." *Journal of Economic Dynamics and Control* 31(11): 3671-3698.

Gavazza, Alessandro, Simon Mongey and Giovanni L. Violante. 2018. "Aggregate recruiting intensity." *American Economic Review* 108(8): 2088-2127.

Gottfries, Axel. 2019. "Bargaining with Renegotiation in Models with On-the-Job Search." Mimeo, University of Edinburgh.

Gouin-Bonenfant, Émilien (2022). "Productivity Dispersion, Between-Firm Competition, and the Labor Share," mimeo, Columbia University.

Hagedorn, Marcus and Iourii Manovskii. 2011. "Productivity and The Labor Market: Comovement over the Business Cycle." *International Economic Review* 52(3): 603-619.

Hall, Robert E. and Paul R. Milgrom. 2008. "The limited influence of unemployment on the wage bargain." *American Economic Review* 98(4): 1653-1674.

Haltiwanger, John, Ron S. Jarmin, and Javier Miranda. 2013. "Who Creates Jobs? Small versus Large versus Young." *Review of Economics and Statistics* 95(2): 347-361.

Hamermesh, Daniel. 1989. "Labor Demand and the Structure of Adjustment Costs." *American Economic Review* 79(4): 674-689.

Harrison, J. Michael, and Michael I. Taksar. 1983. "Instantaneous Control of Brownian Motion." *Mathematics of Operations Research* 8(3): 439-453.

Hopenhayn, Hugo. 2014. "Firms, Misallocation, and Aggregate Productivity: A Review." *Annual Review of Economics* 6: 735-770.

Hopenhayn, Hugo and Richard Rogerson. 1993. "Job turnover and policy evaluation: A general equilibrium analysis." *Journal of Political Economy* 101(5): 915-938.

Ivanov, Ventzislav and Lutz Kilian. 2005. "A Practitioner's Guide to Lag Order Selection for VAR Impulse Response Analysis." *Studies in Nonlinear Dynamics and Econometrics* 9(1): 1-36.

Kaas, Leo and Philipp Kircher. 2015. "Efficient firm dynamics in a frictional labor market." *American Economic Review* 105(10): 3030-3060.

Kline, Patrick, Neviana Petkova, Heidi Williams, and Owen Zidar. 2019. "Who Profits from Patents? Rent-Sharing at Innovative Firms." *Quarterly Journal of Economics* 134(3): 1343–1404.

Krause, Michael U. and Thomas A. Lubik. 2006. "The cyclical upgrading of labor and on-the-job search." *Labour Economics* 13(4): 459-477.

Lazear, Edward P. and James R. Spletzer. 2012. "Hiring, Churn and the Business Cycle." *American Economic Review Papers and Proceedings* 102(3): 575-579.

Lentz, Rasmus, and Dale T. Mortensen. 2012. "Labor market friction, firm heterogeneity, and aggregate employment and productivity." Mimeo, University of Wisconsin.

Lise, Jeremy, and Jean-Marc Robin. 2017. "The macrodynamics of sorting between workers and firms." *American Economic Review* 107(4): 1104-35.

Manning, Alan. 2011. "Imperfect Competition in the Labor Market." *Handbook of Labor Economics.* Elsevier.

Menzio, Guido, and Shouyong Shi. 2011. "Efficient Search on the Job and the Business Cycle." *Journal of Political Economy* 119 (3): 468-510.

Mercan, Yusuf and Benjamin Schoefer. 2020. "Jobs and Matches: Quits, Replacement Hiring, and Vacancy Chains." *American Economic Review: Insights* 2(1): 101-124.

Mortensen, Dale T. 2003. *Wage Dispersion: Why Are Similar Workers Paid Differently?* MIT Press.

Mortensen, Dale T., and Christopher A. Pissarides. 1994. "Job creation and job destruction in the theory of unemployment." *Review of Economic Studies* 61(3): 397-415.

Moscarini, Giuseppe. 2005. "Job Matching and the Wage Distribution," *Econometrica* 73(2): 481-516.

Moscarini, Giuseppe and Fabien Postel-Vinay. 2013. "Stochastic search equilibrium." *Review of Economic Studies* 80(4): 1545-1581.

Moscarini, Giuseppe and Kaj Thomsson. 2007. "Occupational and Job Mobility in the US." *Scandinavian Journal of Economics* 109(4): 807-836.

Mukoyama, Toshihiko, Christina Patterson and Ayşegül Şahin. 2018. "Job Search Behavior over the Business Cycle." *American Economic Journal: Macroeconomics* 10 (1): 190-215.

Nagypal, Eva. 2007. "Labor Market Fluctuations and On-the-Job Search." Mimeo, Northwestern University.

Nickell, Stephen. 1978. "Fixed Costs, Employment and Labour Demand Over the Cycle." *Economica* 45(180): 329-345.

Nickelsburg, Gerald. 1985. "Small-Sample Properties of Dimensionality Statistics for Fitting VAR Models to Aggregate Economic Data. A Monte Carlo Study." *Journal of Econometrics* 28(2):183-192.

Oi, Walter Y. 1962. "Labor as a Quasi-Fixed Factor." *Journal of Political Economy* 70(6): 538-555.

Postel-Vinay, Fabien and Jean-Marc Robin. 2002. "Equilibrium wage dispersion with worker and employer heterogeneity." *Econometrica* 70(6): 2295-2350.

Schaal, Edouard. 2017. "Uncertainty and unemployment." *Econometrica* 85(6): 1675-1721.

Shimer, Robert. 2005. "The cyclical behavior of equilibrium unemployment and vacancies." *American Economic Review* 95(1): 25-49.

Shimer, Robert. 2006. "On-the-job search and strategic bargaining." *European Economic Review* 4(50): 811-830.

Solon, Gary, Robert Barsky, and Jonathan Parker. 1994. "Measuring the Cyclicality of Real Wages: How Important is Composition Bias?" *Quarterly Journal of Economics* 109(1): 1-25.

Stole, Lars A. and Jeffrey Zwiebel. 1996. "Intrafirm Bargaining Under Non-binding Contracts." *Review of Economic Studies* 63(3): 375-410.

White, Harrison C. 1970. *Chains of Opportunity: System Models of Mobility in Organizations*. Cambridge, Mass: Harvard University Press.