Carrillo-Tudela, Carlos; Clymo, Alex; Coles, Melvyn

Working Paper

Equilibrium Job Turnover and the Business Cycle

IZA Discussion Papers, No. 14869

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Carrillo-Tudela, Carlos; Clymo, Alex; Coles, Melvyn (2021) : Equilibrium Job Turnover and the Business Cycle, IZA Discussion Papers, No. 14869, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/250530

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 14869

Equilibrium Job Turnover and the Business Cycle

Carlos Carrillo-Tudela
Alex Clymo
Melvyn Coles

NOVEMBER 2021

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 14869

Equilibrium Job Turnover and the Business Cycle

Carlos Carrillo-Tudela
University of Essex, CEPR, CESifo and IZA

Alex Clymo
University of Essex

Melvyn Coles
University of Essex

NOVEMBER 2021

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Equilibrium Job Turnover and the Business Cycle*

This paper develops and estimates a fully microfounded equilibrium business cycle model of the US labor market with aggregate productivity shocks. Those microfoundations are consistent with evidence regarding the underlying distribution of firm growth rates across firms [by age and size] and, when aggregated, are consistent with macro-evidence regarding gross job creation and job destruction flows over the cycle. By additionally incorporating on-the-job search, we systematically characterise the stochastic relationships between aggregate job creation and job destruction flows across firms, gross hire and quit flows [churning] by workers across firms, as well as the persistence and volatility of unemployment and worker job finding rates over the cycle.

JEL Classification: E24, E32, J62, J63

Keywords: job search, firm dynamics, business cycle

Corresponding author:
Carlos Carrillo-Tudela
Department of Economics
University of Essex
Wivenhoe Park
Colchester, CO4 3SQ
United Kingdom

E-mail: cocarr@essex.ac.uk

* We would like to thank participants in seminars at the universities of Cambridge, Oxford, LMU Munich, FAU Nuremberg, VCU Virginia and the Cleveland FED for their comments and suggestions as well as participants in the 3rd Dale T. Mortensen Centre Conference "Labour Markets and Search Frictions", Essex SaM workshop, NBER Summer Institute Macro Perspectives 2021, and EEA-ESEM 2021. We would also like to thank Jason Faberman for sharing with us his data. The usual disclaimer applies.

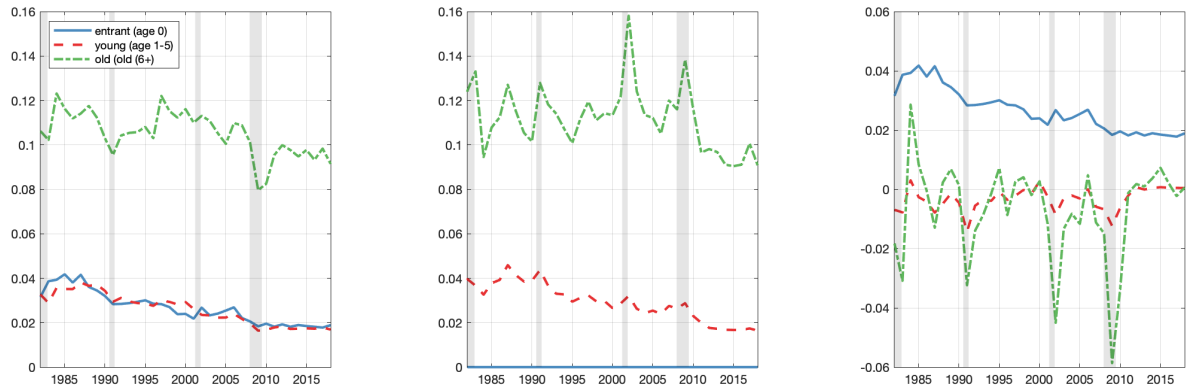
1 Introduction

This paper develops and structurally estimates a fully microfounded equilibrium business cycle model of the US labour market with aggregate shocks. Those microfoundations are consistent with evidence regarding the underlying distribution of firm growth rates across firms [by age and size] and, when aggregated, are consistent with macro-evidence regarding gross job creation and job destruction over the cycle (e.g. Davis and Haltiwanger, 1992, Davis, et al., 2013, Sedlacek and Sterk, 2017, Haltiwanger et al., 2018 and Elsby et al., 2021). By additionally incorporating on-the-job search, we systematically characterise and estimate the stochastic relationships between aggregate job creation and job destruction flows across firms, gross hire and quit flows [churning] by workers across firms, as well as the persistence and volatility of unemployment and of worker job finding rates over the cycle.

Figure 1 motivates our approach: using BDS data, it describes yearly gross job creation flows $[JC_t]$ and gross job destruction flows $[JD_t]$ by age of firm, normalised by total employment N_t. To reveal the underlying growth structure of firms, it groups the data into three firm types: mature firms [those at least 6 years old], young firms [aged between 1-5 years] and start-ups [those aged less than one year]. Because the majority of workers are employed in mature firms, gross job creation and gross destruction flows are largest for this group of firms. Figure 1(a) and (b) show that for both young and mature firms, gross job creation and job destruction flows are of the same order of magnitude. Figure 1(b), however, shows that job destruction flows are strongly countercyclical where every recession coincides with a large spike in gross job destruction flows. In contrast and to the same scale, Figure 1(a) finds the variation in job creation flows is typically smaller and less correlated with the business cycle, though the 2008 recession is a clear exception. Putting this information together, Figure 1(c) identifies the business cycle nature of the creative destruction process. On average, net job creation is negative in existing young and mature firms as new, more productive start-ups enter the economy. But the employment reallocation process is far from smooth with infrequent but large spikes of net job destruction. Which once again raises the two important questions as originally posed in Mortensen and Pissarides (1994): (i) why are net job destruction flows so volatile and (ii) why is the response of net job creation to rising unemployment so weak that recessions

generate long (persistent) phases of high unemployment?

Figure 1: Job destruction and job creation by firm age



(a) JC / total N (b) JD / total N (c) net JC / total N

Job creation and job destruction flows by firm age from the Business Dynamics Statistics database. The data give yearly JC and JD flows from 1978 to 2018, and all series are normalised by total economy-wide employment (specifically, the average of this and last year's employment, as is standard). Net JC is $JC - JD$ for each group.

The theoretical framework is based on Klette and Kortum (2004) extended to on-the-job search; e.g. Lentz and Mortensen (2008), Coles and Mortensen (2016). Different to the standard free entry approach of the canonical Diamond-Mortensen-Pissarides (DMP) framework, here instead it takes time for new start-ups and existing firms to discover and develop new product lines and so create new jobs (a process akin to Diamond, 1982). With on-the-job search, the terms of trade are not determined competitively but instead by an efficiency wage distortion: that firms which pay higher wages enjoy lower quit rates and so firms trade-off paying higher wages against the additional recruitment and training costs to replace workers who quit. Unlike Burdett and Mortensen (1998), however, wages are dynamically consistent in that there is no precommitment by firms on future wages paid. Instead there is an information asymmetry: firm productivity is private information. By suitably adapting a standard first price auction structure, equilibrium finds a higher posted wage signals higher firm productivity and, given that signal, employees then anticipate higher expected wages in the future (relative to other firms) and so quit rates fall.[1]

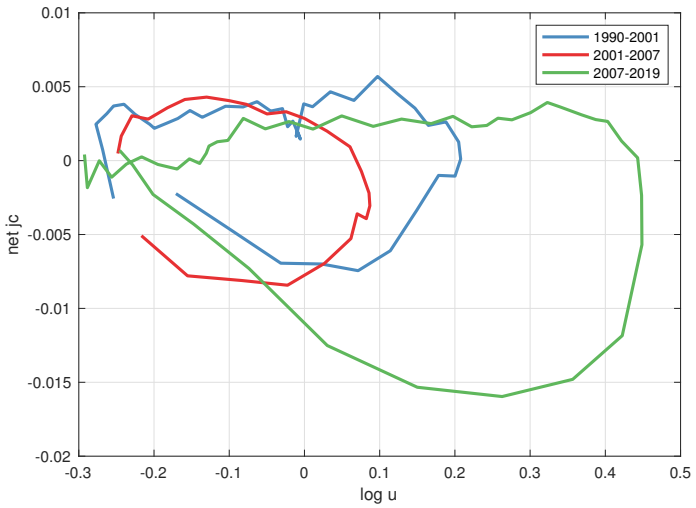
[1]The important sequential auctions approach, popularised by Postel-Vinay and Robin (2002), instead assumes second price auctions. The approaches are not payoff equivalent for rents lost to outside hiring firms are not the same.

The paper contributes to the rapidly growing literature on disperse firm growth rates and the aggregate dynamics of job creation, job destruction and employment; e.g. Schaal (2017), Bilal et al. (2021), Elsby and Gottfries (2021) and Elsby et al. (2021). A very useful property of our stochastic equilibrium, as is also the case in Schaal (2017), is that the aggregate state has finite dimension. Rather than structurally estimate the model based only on cross section evidence (and then use MIT shocks to consider out-of -steady state dynamics), here instead we use full estimation on both cross section and business cycle moments.[2] The broader approach yields several new and important insights. For example defining net job creation rate $njc = (JC - JD)/N$, Table 2 in the quantitative section reports the surprising business cycle fact:

- **net job creation (njc) is uncorrelated with unemployment.**

Although unemployment (U) has seemingly little impact on net job creation rates (njc), it does not imply there is no systematic business cycle relationship as Figure 2 demonstrates.

Figure 2: Joint cyclicality of U and njc



Cyclical net job creation rate is constructed from the quarterly data used by Davis, Faberman and Haltiwanger (2012), updated by these authors, and HP filtered with parameter 10^5. Cyclical unemployment is constructed using quarterly data from the Current Population Survey and also HP filtered with parameter 10^5.

The three US recessions [1990Q2-2018Q4] each generated a large anti-clockwise loop in (U_t, njc_t). For example the 2008 recession found unemployment increased quickly as

[2]Our approach also differs from the recent contribution by Audoly (2021). He also builds on the Coles and Mortensen (2016) framework but is required to use numerical methods to approximate the infinitely dimension state space of his stochastic equilibrium in order to analyse its cyclical properties.

njc fell far below trend value. Of course once njc began to recover the increase in unemployment slowed and unemployment ultimately recovered once njc rose above trend. Figure 2 makes transparent that the observed volatility and persistence of cyclical unemployment is directly related to the business cycle properties of net job creation, while njc is itself uncorrelated with unemployment. Explaining these features of the data raises an important challenge for any equilibrium theory of unemployment. For example our approach not only explains the aggregated JC and JD patterns by age of firm as described in Figure 1, but also the persistence and volatility of unemployment consistent with business cycle variations in aggregate job creation and destruction rates [$jc = JC/N$ and $jd = JD/N$] and so njc. In contrast the canonical free entry approach is typically criticised for generating counterfactual Beveridge curve correlations in the event of a large job destruction shock; i.e.vacancy creation rates react too strongly to increased unemployment.[3] The more direct criticism, however, is to ask whether the theory generates data consistent (jc, jd) dynamics.

The Klette and Kortum (2004) approach generates an equilibrium job creation/job destruction process at the firm level which is consistent with Gibrat's law, that individual firm growth rates depend on firm productivity but are otherwise independent of firm size. An important contribution is that we identify the firm start-up and firm specific productivity processes which generate distributions of firm size by age which are data consistent. Specifically we follow the Haltiwanger et al. (2017) insight that some new start-ups might be described as "gazelles": high productivity start-ups which exhibit high growth rates and mature into large firms. But there are also less fortunate start-ups which instead have low productivity, low growth and low survival rates. Because large mature firms are highly likely to have started life as gazelles, ex-ante start-up heterogeneity explains why "young" firms, on average, evolve differently to "mature" firms, though steady state still implies most firms are small while most workers are employed in large (mature) firms (which is an important feature of the data). Nevertheless the em-

[3]A well established literature argues that in a standard DMP matching model, large job destruction shocks are inconsistent with the data. Coles and Moghaddasi (2018), however, show the difficulties which arise in the seminal Mortensen and Pissarides (1994) paper, as detailed in the Shimer (2005) critique, are largely due to the free entry of vacancies assumption. Specifically the free entry approach is problematic because it implies net job creation flows which are far too elastic for the data (see Figure 3 below). With a less than infinitely elastic job creation process, for example an entry process akin to Diamond (1982) as considered here, Coles and Moghaddasi (2018) establish the key Shimer (2005) insight, that large job destruction shocks yield counterfactual Beveridge curve correlations, no longer applies.

pirical firm size distributions are not consistent with gazelles being gazelles forever. The example of Rubik's Cubes then motivates our approach. Rubik's Cube manufacturers initially had very high demand and production expanded quickly, but once everybody had bought their cube, sales subsequently slumped. Rather than assuming decreasing returns to scale (see Schaal, 2017, Bilal et al., 2021, Elsby and Gotffries, 2021, Elsby et al., 2021), Schumpeter product cycles instead suggest a particular time structure for firms: that an early gazelle phase with a high product price is potentially followed by a period of stabilization and possible decline. Assuming constant returns to scale but with sunk capital costs, our approach is to calibrate the firm productivity process so that the firm size distributions by age are consistent with the data and thus with the typical life cycle of firms.

By incorporating on-the-job search into the stochastic equilibrium we additionally address two important issues. Firstly on-the-job search generates vacancy chains where if an already employed worker takes a new job created, the worker's previous employer may then hire a replacement worker. Because around one half of hires involves an already employed worker, vacancy chains and replacement hiring explain why gross quit and hire flows are more volatile than gross job creation flows over the cycle; e.g. Faberman and Nagypal (2008), Mercan and Schoefer (2020) and Elsby et al. (2021). Allowing on-the-job search also generates a second channel of job destruction not through layoff but instead by not replacing workers who quit. This job destruction channel has important implications for unemployed worker job finding rates. For example (and ignoring for the moment exogenous quits into unemployment) hires out of unemployment H_t^{UE} satisfy

$$JC_t = H_t^{UE} + JD_t^{Q} \tag{1}$$

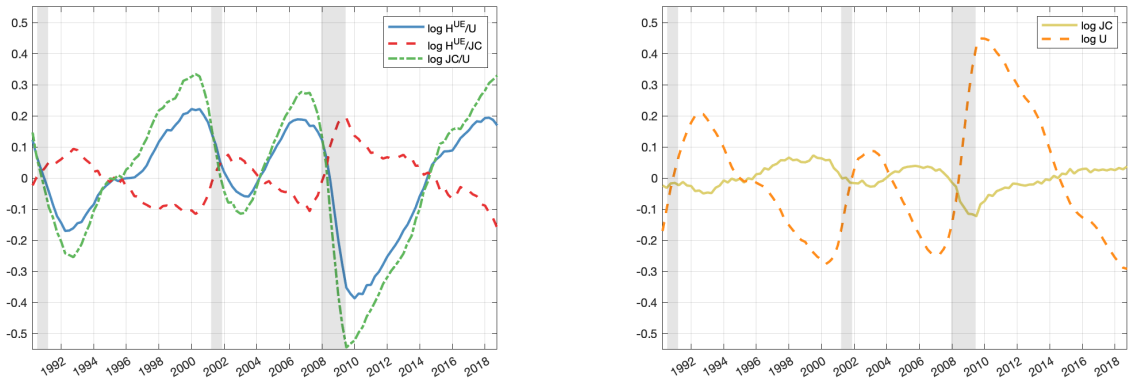
where JC_t describes the flow of new jobs created, re-interpreted instead as new vacancy chains created, while the right hand side describes the corresponding destruction of vacancy chains which occurs when either the job is taken by an unemployed worker H_t^{UE}, or by a worker who quits and the previous employer declines to hire a replacement JD_t^{Q}. By directly crowding out re-employment flows H_t^{UE} this job destruction channel has important business cycle implications. For example, unemployed worker job finding rates over the cycle are typically measured as H_t^{UE}/U_t. Taking on-the-job search into account,

consider its log decomposition:

$$\log \frac{H_t^{UE}}{U_t} = \log \frac{H_t^{UE}}{JC_t} + \log \frac{JC_t}{U_t}. \tag{2}$$

The first term H_t^{UE}/JC_t, describes *job creation yield*, the fraction of new jobs created which result in a worker being hired out of unemployment, where equation (1) establishes that $JD_t^Q \neq 0$ implies $H_t^{UE}/JC_t \neq 1$. The second term JC_t/U_t instead describes the gross flow of new jobs created per unemployed worker.

Figure 3: Job finding rate decomposition



(a) Decomposing UE rate

(b) JC vs. U

Left panel implements the decomposition in (2). JC flow is the quarterly flow from the Davis, Faberman, and Haltiwanger (2012) dataset, and the UE hiring flow is computed from the Current Population Survey following Shimer (2005). U gives the unemployment stock. Right panel plots the JC flow and U series separately. All series are logged and HP-filtered with parameter 10^5.

Figure 3(a) uses data from Davis et al. (2012) and plots these three items where, consistent with the standard view, job finding rates H_t^{UE}/U_t vary widely over the cycle and are highly persistent. Note, however, that job creation yield H_t^{UE}/JC_t actually increases in the recession: a new job created is more likely to result in the hiring of an unemployed worker. The model reproduces this feature of the data and explains it by crowding out, that higher unemployment crowds out on-the-job search and so both quits and JD^Q fall in the recession. The crucial insight, however, is that despite this increase in yield, there is a steep fall in job finding rates following a recession because of the even steeper fall in jobs created per unemployed worker JC_t/U_t. Furthermore Figure 3(b) demonstrates this occurs not because JC flows fall dramatically following the recession but because JC responds weakly to increasing unemployment; i.e. job finding rates fall steeply in the re-

cession because there are so many more unemployed workers chasing the relatively few new jobs being created per unemployed worker.

Finally our approach also resolves a seeming puzzle: although there is the well-known large firm wage effect, that firm size and wages paid are positively correlated (Brown and Medoff, 1989), and that job to job quits are typically to higher wage paying firms, Haltiwanger et al. (2018) find there is no systematic drift of workers from small to large firms. Our estimated model also exhibits these properties.

The paper is structured as follows. Section 2 describes the model, and Section 3 derives the equilibrium. Section 4 details the estimation of the quantitative model and steady-state results. Section 5 gives the business cycle results, and Section 6 concludes.

2 The Model

Time t is continuous, has an infinite horizon and we consider a stochastic equilibrium with aggregate shocks. There is a unit measure of equally productive workers who are risk neutral, infinitely lived and each has the same discount rate $r > 0$. At any point in time each worker is either employed (earning a wage w) or unemployed (with home production $b > 0$). Unemployed workers receive job offers according to a Poisson process with time varying parameter λ_{0t}, where a job offer is considered a random draw from the set of hiring firms. There is on-the-job search where employed workers instead receive job offers at rate $\lambda_{1t} = \phi\lambda_{0t}$ where $\phi \in (0, 1]$ is an exogenous parameter. In what follows $\lambda_{0t}, \lambda_{1t}$ are endogenous objects where a Markov equilibrium determines $\lambda_{0t} = \lambda_0(\Omega_t)$ and $\lambda_{1t} = \phi\lambda_0(\Omega_t)$ with Ω_t describing the aggregate state.

Firms are heterogeneous, risk neutral and have the same discount rate $r > 0$. For ease of exposition we initially assume a continuum of firm productivity states $i \in [0, 1]$ but in the application shall restrict to finite states. There are constant returns to production: given aggregate productivity $s \in \{1, 2, .., S\}$, firm $i \in [0, 1]$ with integer $n \in \mathbb{N}^+$ employees generates flow revenue $np^s(i)$ which is strictly increasing in i and s.

Unlike the matching function approach, here there are no "search frictions" as usually considered. Instead there is stock-flow matching where the stock of unemployed workers [on the long side of the market] matches immediately with the inflow of new jobs [on the short-side of the market]; e.g. Shapiro and Stiglitz (1984), Coles and Smith (1998)

among many others. Because matching is sequential and random, the equilibrium might be considered a limiting random matching equilibrium as frictions become small. The formal assumption is that by paying hiring cost c_0 the firm immediately and randomly fills a job from the set of workers who prefer this job to their current position (taking into account on-the-job search effectiveness $\phi < 1$). c_0 thus depends on the direct recruitment and training costs of hiring a new employee. On the other side of the market, however, a positive stock of unemployed workers and a finite flow of new jobs implies it takes time for unemployed workers to obtain work, where random matching yields crowding out effects in the sense that already employed workers (who seek better paid employment) adversely affect the matching opportunities of the unemployed.

Following Coles and Mortensen (2016) we assume firms cannot precommit to future wages and do not observe outside job offers (i.e. there is no job offer matching). Each firm thus pays a sequence of spot wages to each of its [equally productive] employees. The firm's productivity state i is private information and so, in a Bayesian equilibrium, the firm's posted wage is a signal of its productivity. Of course a signalling equilibrium implies the firm's wage strategy potentially depends on its entire wage posting history. Furthermore different employees are hired at different dates and so observe different parts of that wage history. In the stationary [Markov Bayesian] equilibrium considered here, the wage structure is analogous to that of a first price auction with independent private values, where hiring firms with higher [private] productivities bid strictly higher wages. Because there are repeated transactions, however, where each firm typically employs the same workers from day to day, the auction structure is not equivalent to a static one-shot first price auction. Instead with repeated trade, the current wage is a predictor of future wages. The key assumption for tractability is each firm's productivity follows a positively autocorrelated Markov process so that a high productivity firm is more likely to have high productivity in the future.

Because higher i firms have higher values, the auction structure finds higher i firms bid higher wages and the wage offer strategies of hiring firms fully reveal their state i. Should a firm cut its wage, its employees believe it has received an adverse i shock and, anticipating lower wages in the future, worker quit rates increase [to higher wage paying firms]. Because replacing a worker who quits is costly, each firm trades-off paying lower wages against a higher quit rate. We adopt the tie-breaking conventions that the worker

quits when indifferent and that the firm invests if indifferent. There is no recall of rejected job offers. Indeed if a worker rejects a job offer, equilibrium finds the job is anyway filled by someone else.

The equilibrium framework allows rich micro-firm dynamics. There is firm turnover where, on start-up [described below], a new firm has $n_0 \in \{1, 2, .., N_0 + 1\}$ employees and initial productivity i considered a random draw from the uniform distribution $U[0, 1]$. While the firm survives it is subject to a wide variety of shocks:

(i) **Aggregate productivity shocks:** given current state $s \in \{1, 2, .., S\}$, an aggregate productivity shock occurs at exogenous rate $\alpha_a \geq 0$ and transition matrix $\Upsilon_{ss'}$ describes the probability the new state is s';

(ii) **Firm specific productivity shocks:** at exogenous rate $\alpha_\gamma \geq 0$ a firm i has new productivity $i' \in [0, 1]$ considered a random draw from c.d.f. $\Gamma(i'|i)$. For the moment we shall assume no mass points in $\Gamma(.)$ but shall relax this for the application. The transition probabilities $\Gamma(.)$ satisfy first order stochastic dominance so that higher state i firms are more likely to be higher productivity firms in the entire future;

(iii) **Firm level job creation** is described by an idiosyncratic growth process: an expansion opportunity occurs at firm (i, n) according to a Poisson process with parameter $\mu_1 n$ where $\mu_1 > 0$ is exogenous. Associated with any expansion opportunity is an idiosyncratic capital investment cost $c^{JC} \geq 0$ considered a random draw from cdf $H^{JC}(.)$. If the firm invests, it pays c^{JC} and so creates an additional [unfilled] job. The post is then filled by paying an additional recruitment cost $c_0 \geq 0$ whereupon the firm's size increases to $n + 1$ and one job is created. If the firm declines the expansion opportunity, its firm size n remains unchanged and there is no recall;

(iv) **Firm level downsizing:** idiosyncratic capital destruction shocks occur at an exogenous rate δ_D. If a unit of capital is destroyed, the firm can re-invest at cost $c^{JD} \geq 0$ considered a random draw from $H^{JD}(.)$. If the firm re-invests the firm's size n is unchanged and $JD = 0$. If the firm does not re-invest, the corresponding employee is laid-off into unemployment and the firm downsizes to $n - 1$ with one job destroyed.

(v) **Quit shocks:** employees may receive a preferred outside offer and so quit. Whenever a quit occurs, the firm has the option of paying a recruitment cost $c_0 > 0$ to hire a replacement employee. If it does so, firm size n remains unchanged with $JD = 0$. If instead the

firm chooses not to hire a replacement employee, firm size falls to $n-1$ and one job is destroyed.

(vi) **Exogenous firm exit shocks:** at exogenous rate δ_F a firm experiences an exit shock and closes down with n jobs destroyed.

(vii) **Exogenous separations:** an employee separates into unemployment at exogenous rate λ_u and the firm then decides whether to hire a replacement. In the data it is ambiguous whether such a separation is a layoff or a quit. For ease of exposition, the theory section considers exogenous separations as quits. The quantitative section, however, calibrates λ_u to match the aggregate layoff rate of firms.

There is a unit measure of entrepreneurs who independently seek business ventures. At rate μ_0 an entrepreneur identifies a possible business venture whose investment cost $c^E \geq 0$ is considered an independent random draw from cost distribution $H^E(.)$. If the entrepreneur chooses not to invest, the venture is lost with no recall. If the entrepreneur invests, a start-up is created with a single employee drawn randomly from the pool of unemployed workers. Its productivity $i \sim U[0,1]$ is then revealed at which point we refer to the start-up as a new firm. The new firm thus has a single employee, is in state i and subsequently pays wages and expands/contracts like all other existing firms. For calibration purposes, however, we suppose each new firm also has N_0 immediate potential expansion jobs where $N_0 \in \mathbb{N}^+$ is exogenous. The job creation process is the same as for existing firms: associated with each potential expansion is an independent cost draw $c^{JC} \sim H^{JC}$ and recruitment cost c_0 to hire a worker. If the new firm invests its initial size n_0 increases by 1. If the new firm does not invest, the expansion opportunity is lost with no recall. Hence each new firm begins life with initial employment $n_0 = 1 + \widetilde{n}_i$ where hires $\widetilde{n}_i$ are a binomially distributed random variable with N_0 independent trials and an [endogenous] probability of investment which depends on (i, Ω). Consistent with the data, this extension allows that many new firms have starting employment $n_0 > 1$ while allowing that start-up size is potentially sensitive to the aggregate state Ω_t. This completes the description of the model.

Some Preliminary Comments: Events (iii)-(iv) describe hold-up problems at the firm's job creation and job destruction margins. Although hold-up often leads to inefficient outcomes, this is not necessarily the case (e.g. Hosios, 1991). Nevertheless outside

of a competitive equilibrium, an optimal contract might require employees to contribute to [firm specific] re-investment costs. We rule this out. Rather than introduce a complicated negotiating problem where such investments may not be observable/verifiable by workers, and given the firm's productivity i is already private information, the framework simply adopts a standard hold-up structure: the firm either immediately invests or the opportunity is lost. That is not to say that wages are unaffected by such costs, for the re-investment process generates a positive user cost of capital which reduces match surplus. The hold-up problem essentially implies wages paid reflect the [expected] user cost of capital rather than specific realisations.

Because a start-up (exogenously) recruits one unemployed worker with no crowding out by employed workers, we consider that recruitment channel separately from the analysis that follows. Of course we take all recruitments into account when describing gross job creation flows. Throughout we distinguish between rates and flows by using lower case to describe rates; e.g. $jc(i,\Omega)$ will denote the job creation rate per employee at incumbent firms, and upper case $JC(i,\Omega)$ will denote the total job creation flow across all incumbent and entrant firms.

3 Equilibrium

Because a dynamic signalling equilibrium with repeated trade, aggregate productivity shocks and (privately observed) firm specific productivity shocks is complex, we only consider stationary [Markov Bayesian] equilibria with the following properties. Let U_t denote the measure of workers who are unemployed at date t and let $G_t(i)$ denote the fraction of employed workers at firms no greater than $i \in [0,1]$. For ease of exposition we assume $G_t(.)$ has a connected support and that its density exists. If $s_t \in \{1,2,..,S\}$ is aggregate productivity at date t, the aggregate state is $\Omega_t = (s_t, U_t, G_t(.))$.[4] The stationary [Markov Bayesian] equilibrium implies each firm is fully described by (i,n,Ω). Reflecting the constant returns to scale structure, optimality also implies the wage and investment strategies of firms are size independent. At first sight this seems inconsistent with the large firm wage effect, that wages paid are positively correlated with firm size.

[4]Note this aggregate state is infinitely dimensional. Importantly we will show that it reduces to a finite vector when we instead consider a finite set of productivity states $i \in \{1,2,..,I\}$.

The approach also implies a version of Gibrat's Law which some argue is not consistent with data on firm growth outcomes. The quantitative section, however, finds there is no inconsistency once we allow for ex-ante new firm heterogeneity, where "gazelles" evolve differently to other new firms.

Definition 1 (Stationary equilibrium). *For $\Omega = (s, U, G(.))$, a stationary [Markov Bayesian] equilibrium is the following set of functions:*
Existing firms (i, n, Ω):

1. *$w(i, \Omega)$ is the profit maximising wage strategy of each firm (i, n, Ω);*
2. *$jc(i, \Omega) \geq 0$ is the profit maximising job creation rate per employee, and so $n[jc(i, \Omega)]$ describes its expected gross job creation flow;*
3. *$jd(i, \Omega) \geq 0$ is the profit maximising job destruction rate per employee, and so $n[jd(i, \Omega)]$ describes its expected gross job destruction flow;*
4. *$h(i, \Omega)$ is the optimal hiring rate per employee, and so $n[h(i, \Omega)]$ describes its expected gross hire flow;*

New firms (i, n_0, Ω):

5. *$P^E(\Omega)$ is the probability an entrepreneur invests in a start-up in state Ω and, given realised i, its starting size $n_0 = 1 + \widetilde{n}(i, \Omega)$ maximises expected profit;*

Worker search:

6. *$F(w, \Omega)$ is the distribution of wage offers across (new and existing) firms;*
7. *$\lambda_0(\Omega)$ and $\lambda_1(\Omega)$ are the corresponding job offer arrival rates for unemployed and employed workers respectively;*
8. *$\widehat{q}(w, \Omega)$ is the optimal quit rate of a worker employed at a firm paying wage w;*
9. *employed and unemployed workers use job search strategies to maximise expected lifetime value where given any wage paid w, the worker's belief on the firm's underlying state i is consistent with the set of equilibrium wage strategies and Bayes' rule;*

Markov restriction:

10. *Ω follows a first order Markov process consistent with the equilibrium strategies of firms and workers, where $\mu_0 P^E(\Omega)$ describes the additional inflow of unemployed workers into new start-ups and δ_F is the exogenous closure of firms through bankruptcy;*

11. $q(i,\Omega) = \widehat{q}(w(i,\Omega),\Omega)$ *is the equilibrium quit rate per employee at productivity i firms.*

A stationary equilibrium is a complex object, especially because a Bayesian equilibrium requires a complete description of worker beliefs for all possible wage announcements. Coles and Mortensen (2016) show that without restrictions on out-of-equilibrium beliefs, it is possible to support a plethora of equilibria. Such possibilities, however, are ruled out by the following restriction.

Assumption 1 (Monotone Beliefs). *For any Ω, worker beliefs on the firm's state i is first order stochastically increasing in the posted wage.*

The restriction to monotone beliefs rules out punishment strategies. For example consider a posted wage w' where there does not exist any firm $i' \in [0,1]$ with $w' = w(i',\Omega)$. If firm i increases its wage paid to $w' > w(i,\Omega)$, it is consistent with a Bayesian equilibrium that its employees then believe the firm's state $i = 0$ [the least productive state] and the higher wage paid is then punished by an increased higher quit rate. Such punishment strategies can always be used to deter firms from offering higher wages. Monotone beliefs rules this out: firms which pay higher wages induce more favourable worker beliefs on its productivity i. The arguments in Coles and Mortensen (2016) establish the following Claim which we state here without proof.

Claim 1. *A stationary equilibrium with monotone beliefs implies the job offer wage distribution $F(w,.)$ is continuous and has a connected support.*

The Claim describes a standard property in the equilibrium wage dispersion literature – essentially the monotonicity restriction ensures that standard arguments apply. An important corollary is that wage strategies $w(i,\Omega)$ must then be continuous and strictly increasing in i across hiring firms. Consider now $\underline{w}_t$ defined as the lowest wage paid in the market. Coles and Mortensen (2016) assumes $\phi = 1$; i.e. $\lambda_0(\Omega) = \lambda_1(\Omega)$, from which it followed that the reservation wage of workers $R(\Omega_t) = b$ [worker home productivity] and the lowest wage paid $\underline{w}_t = R(\Omega_t) = b$. Assuming $\phi < 1$, however, is the more empirically reasonable case. For this case it is straightforward to show workers still adopt a reservation wage strategy $R_t = R(\Omega_t)$ and equilibrium wage posting implies the lowest wage offered $\underline{w}_t = R_t$. The conditions which determine $R(\Omega_t)$, however, are complex, and to focus on firm growth and turnover, we simplify by assuming there is a binding

minimum wage policy: the government imposes a minimum wage $w_{\min}$ below which firms cannot pay. Thus although unemployed workers are willing to accept a lower starting wage $w_0 = R(\Omega) < w_{\min}$, the minimum wage policy constrains the lowest wage paid $\underline{w} = w_{\min}$.[5]

To fix ideas we quickly describe the key features of the stationary equilibrium, which we will derive formally in the following sections. In the following, firm $i = i^c(\Omega)$ is the marginal surviving firm [the firm closure margin] while firm $i = i^h(\Omega)$ is indifferent to replacing a worker who quits [the hiring margin]. Equilibrium finds firms with $i < i^c$ immediately close down, firms with $i \in [i^c, i^h)$ survive but do not recruit, while firms $i \geq i^h$ have a strictly positive hire flow for they (at least) replace workers who quit. The equilibrium wage strategies will be found to satisfy:

$$\begin{array}{ll} 1. & w(i,\Omega) = w_{\min} \text{ for } i \in [i^c, i^h]; \\ 2. & w(i,\Omega) \text{ is continuous and strictly increasing in } i \text{ for } i \geq i^h. \end{array} \tag{3}$$

Although the distribution of wage offers across hiring firms contains no mass points, that does not imply there is no mass point in the distribution of wages paid. Low productivity firms with $i \in [i^c, i^h)$ are in decline – they survive but do not invest in new job creation and do not replace workers who quit. Equilibrium finds all such firms pay the minimum wage. Bayes rule, (3), and monotone beliefs then imply the following equilibrium worker beliefs:

Belief 1: if a firm posts wage $w' \in (w_{\min}, \overline{w}]$ where $\overline{w} = w(1,\Omega)$, the worker believes the firm's productivity $i = \widehat{i}(w',\Omega)$ where $\widehat{i}$ is the unique solution to $w(\widehat{i},\Omega) = w'$; i.e. beliefs $\widehat{i}$ are the inverse of the equilibrium wage function;

Belief 2: if a firm posts wage $w' = w_{\min}$ and it is an outside job offer the worker believes the firm's productivity $\widehat{i} = i^h$ for it is a hiring firm. At a non-hiring firm an employee instead believes $\widehat{i} \in [i^c, i^h]$ where the specific choice plays no important role;

Belief 3: if a firm posts wage $w' > \overline{w} = w(1,\Omega)$, monotonicity implies the worker believes firm productivity $\widehat{i} = 1$.

These beliefs imply an employee holding an outside offer will quit if and only if the wage offered by the outside firm is (weakly) higher than the worker's current wage w'.

[5]Formally we assume b sufficiently small that $R(\Omega_t) < w_{\min}$ for all realised Ω_t.

This follows because a higher outside offer and Beliefs 1-3 imply the outside firm is believed to have higher productivity, and first order stochastic dominance in $\Gamma(.)$ and (3) then imply the outside firm is more likely to post higher wages in the entire future. Hence the quit rate of each worker at a continuing firm (i, n, Ω) which posts wage w' is

$$\widehat{q}(w', \Omega) = \lambda_1(\Omega)[1 - F(w', \Omega)] + \lambda_u. \tag{4}$$

Since all firms below the hiring margin ($i < i^h(\Omega)$) pay wage $w_{\min}$, equation (4) implies that they all have common quit rate: $q(i, \Omega) = \widehat{q}(w_{\min}, \Omega) = \lambda_1(\Omega) + \lambda_u$ for all $i < i^h(\Omega)$.

Given this characterisation of worker quit behaviour, we now consider the optimal choices of (existing) firms (i, n, Ω) and (new) firms (i, n_0, Ω).

3.1 Firm Optimality [Existing Firms]

Consider any existing firm (i, n, Ω) in a stationary equilibrium. Standard arguments imply the following Bellman equation for continuing firms with value $\Pi(i, n, \Omega) > 0$:

$$\begin{aligned} r\Pi(i,n,\Omega) = \max_{w' \geq w_{\min}} \quad & n[p^s(i) - w'] + n\widehat{q}(w', \Omega) \max[\Pi(i, n-1, \Omega) - \Pi(i, n, \Omega), -c_0] \\ & + \mu_1 n \mathrm{E} \max[\Pi(i, n+1, \Omega) - \Pi(i, n, \Omega) - [c_0 + c^{JC}], 0] \\ & + \delta_D n E \max[\Pi(i, n-1, \Omega) - \Pi(i, n, \Omega), -c^{JD}] + \delta_F[-\Pi(i, n, \Omega)] \\ & + \alpha_\gamma \int_0^1 [\max[\Pi(j, n, \Omega), 0] - \Pi(i, n, \Omega)] \, d\Gamma(j|i) \\ & + \alpha_a \sum_{s'} \Upsilon_{ss'} [\Pi(i, n, \Omega(s')) - \Pi(i, n, \Omega(s))] + \frac{\partial \Pi(i, n, \Omega)}{\partial t}. \end{aligned} \tag{5}$$

Given posted wage $w' \geq w_{\min}$, the firm's flow return equals its flow profit, plus the capital gains which arise when (i) a quit occurs (where the firm has the option of paying c_0 to hire a replacement), (ii) an expansion opportunity occurs with cost $c^{JC} \sim H^{JC}$, (iii) a downsizing shock occurs with cost $c^{JD} \sim H^{JD}$, (iv) a bankruptcy shock occurs, (v) a firm specific productivity shock occurs, (vi) an aggregate productivity shock occurs and the final term is shorthand for describing the change in $\Pi(.)$ as the state variables $(U_t, G_t(.))$ evolve endogenously over time.

The constant returns structure implies that value is linear in employment, giving

$\Pi(i,n,\Omega) \equiv nv(i,\Omega)$, and (5) simplifies to

$$(r+\alpha_\gamma+\alpha_a+\delta_F)v(i,\Omega) = \max_{w' \geq w_{\min}} \quad p^s(i) - w' - \widehat{q}(w',\Omega)\min[v(i,\Omega),c_0] + \mu_1 E \max[v(i,\Omega) - [c_0 + c^{JC}],0] - \delta_D E_c \min[v(i,\Omega),c^{JD}] + \alpha_\gamma \int_0^1 \max[v(j,\Omega),0]d\Gamma(j|i) + \alpha_a \sum_{s'} \Upsilon_{ss'}[v(i,\Omega(s'))] + \frac{\partial v(i,\Omega)}{\partial t}. \tag{6}$$

$v(i,\Omega)$ is the key element in what follows and describes the firm's expected value per employee. Bellman equation (6) implies the following optimal investment policies:

1. if an employee quits, the firm hires a replacement if and only if $v(i,\Omega) \geq c_0$ [otherwise $JD = 1$];

2. if an expansion opportunity arises with cost c^{JC}, the firm invests and expands if and only if $v(i,\Omega) \geq c^{JC} + c_0$ [whereupon $JC = 1$];

3. if a capital destruction shock occurs with cost c^{JD}, the firm re-invests if and only if $v(i,\Omega) \geq c^{JD}$ [otherwise $JD = 1$].

Because first order stochastic dominance in $\Gamma(.|i)$ implies equilibrium firm values are increasing in i, the firm closure margin i^c is identified where $v(i^c,\Omega) = 0$. The hiring margin i^h is instead given by $v(i^h,\Omega) = c_0$ where firms with $i < i^h$ have $v(i^h,\Omega) < c_0$ and so do not hire. Finally note that a firm only expands when an expansion opportunity arises with investment cost $c^{JC} < v(i,\Omega) - c_0$. Hence the job creation rate $jc(i,\Omega) = \mu_1 H^{JC}(v(i,\Omega) - c_0)$ for $i \geq i^h$ and is zero otherwise.

3.2 Firm Optimality [New Firms]

Suppose in state Ω, an entrepreneur creates a new firm with revealed productivity $i \sim U[0,1]$. If $i < i^c$ the entrepreneur closes the firm [because $v(i,\Omega) < 0$]. If $i \in [i^c, i^h)$ the firm survives but the entrepreneur does not invest in new jobs [because $v(i,\Omega) < c_0$] and so initial firm size $n_0 = 1$. For $i \geq i^h$, the entrepreneur invests in each expansion opportunity if and only if realised $c^{JC} \leq v(i,\Omega) - c_0$. Thus start-up employment $n_0 = 1 + \widetilde{n}(i,\Omega)$ where $\widetilde{n}(i,\Omega)$ is a binomially distributed random variable with expected value

$N_0 H^{JC}(v(i,\Omega) - c_0)$. The expected value of a start-up is therefore

$$\Pi^{SU}(\Omega) = \int_{i^c}^{1} \left\{ v(i,\Omega) + N_0 \int_0^{v(i,\Omega)-c_0} [v(i,\Omega) - c_0 - c'] dH^{JC}(c') \right\} di.$$

Hence given the investment opportunity, the entrepreneur proceeds with a new start-up when $c^E \leq \Pi^{SU}(\Omega)$ and so $P^E(\Omega) = H^E(\Pi^{SU}(\Omega))$. Initial firm size [in expectation] is then $En_0 = 1 + N_0 H^{JC}(v(i,\Omega) - c_0)$, noting that $H^{JC}(v(i,\Omega) - c_0) = 0$ for all $i < i^h$.

The above establishes the characterisation of firm level job creation, job destruction and hire strategies, which we summarise in the following proposition.

Proposition 1 (Optimal job creation, job destruction, and hiring policies). *A stationary equilibrium with monotone beliefs implies:*
(i) for existing firms (i,n,Ω) with $i \geq i^h$:

$$\begin{aligned} jc(i,\Omega) &= \mu_1 H^{JC}(v(i,\Omega) - c_0) \\ jd(i,\Omega) &= \delta_D[1 - H^{JD}(v(i,\Omega))] \end{aligned}$$

and because these firms replace workers who quit, their hiring rate is

$$h(i,\Omega) = q(i,\Omega) + jc(i,\Omega); \tag{7}$$

(ii) for existing firms (i,n,Ω) with $i \in [i^c, i^h)$, $jc(i,\Omega) = 0$, and their quit rate is $q(i,\Omega) = \lambda_1 + \lambda_u$. As these firms do not replace workers who quit, their hiring rate is $h(i,\Omega) = 0$, and

$$jd(i,\Omega) = \delta_D[1 - H^{JD}(v(i,\Omega))] + \lambda_1 + \lambda_u.$$

(iii) for new firms (i,Ω), optimal investment implies expected starting firm size

$$En_0 = 1 + \frac{N_0}{\mu_1} jc(i,\Omega).$$

For (iii), we used that $H^{JC}(v(i,\Omega) - c_0) = jc(i,\Omega)/\mu_1$.

To calculate the total job creation flow, by the definition of G, $[1-U]G'(i)jc(i,\Omega)$ describes the total job creation flow from all existing $i \geq i^h$ firms. Similarly the uniform distribution implies gross job creation flows (excluding the initial unemployed worker) at new $i \geq i^h$ firms is $\mu_0 P^E(\Omega) N_0 H^{JC}(v(i,\Omega) - c_0)$. Adding both flows together yields the total job creation flow by firm productivity

$$JC(i,\Omega) = \left\{ [1-U]G'(i) + \frac{\mu_0}{\mu_1} N_0 P^E(\Omega) \right\} jc(i,\Omega) \tag{8}$$

Doing the same for job destruction and adding job destruction due to the firm exit shock similarly yields the total job destruction flow by firm productivity

$$JD(i,\Omega) = [1-U]G'(i)\left\{\delta_D[1-H^{JD}(v(i,\Omega))] + \delta_F + \mathbf{1}(i<i^h)(\lambda_u+\lambda_1)\right\} \tag{9}$$

Given this characterisation of optimal hiring strategies and job flows, the next step is to determine the equilibrium wage strategies.

3.3 Equilibrium Wages $w(i,\Omega)$

For any firm (i,n,Ω) with $v(i,\Omega) > 0$, combining (4) with the first line of (6) implies that the optimal wage satisfies:

$$w(i,\Omega) = \text{argmin}_{w' \geq w_{\min}} \left[w' + \lambda_1[1-F(w',\Omega)]\min[v(i,\Omega),c_0]\right]. \tag{10}$$

Equation (10) describes the equilibrium trade-off between paying lower wages and triggering a higher [costly] quit rate. Identifying the equilibrium wage strategies is a well known fixed point problem where the individually optimal wage strategies $w(i,\Omega)$ must, when aggregated, yield the assumed distribution of offered wages $F(w',\Omega)$. This fixed point problem is additionally complicated by being outside of steady state with an endogenously evolving distribution of employment across heterogeneous firms.

The first step to characterising the equilibrium wage outcome is to construct the arrival rate of job offers $\lambda_1[1-F(w,.)]$. Following Proposition 1, the equilibrium flow of hires at each existing firm (i,n,Ω) is $nh(i,\Omega) = n[q(i,\Omega)+jc(i,\Omega)]$ if $i \geq i^h$, and 0 otherwise. Hence by definition of G, $[1-U]G'(i)[q(i,\Omega)+jc(i,\Omega)]$ describes the total gross hire flow from all existing $i \geq i^h$ firms. Similarly the uniform distribution implies gross hire flows at new $i \geq i^h$ firms is $\mu_0 P^E(\Omega)N_0 H^{JC}(v(i,\Omega)-c_0)$. Adding both flows together yields gross hire flows

$$H(i,\Omega) = \begin{cases} 0 & \text{if } i < i^h \\ [1-U]G'(i)q(i,\Omega) + JC(i,\Omega) & \text{if } i \geq i^h, \end{cases} \tag{11}$$

where for $i \geq i^h$ the first term describes hire flows due to replacing workers who quit at existing firms and the second term is hire flows due to endogenous job creation by both existing and new firms. Finally, $H(i,\Omega) = 0$ for $i < i^h$, since neither incumbents nor entrants hire.

Although job offers are randomly made, $\phi < 1$ implies the unemployed receive relatively more offers. Let $\lambda = U\lambda_0 + (1-U)\lambda_1$ denote the total flow of job offers and so fraction $\alpha = [U\lambda_0]/\lambda = U/(U + \phi(1-U))$ of job offers go to the unemployed, the remaining $1-\alpha$ go to employed workers. Note further that an equilibrium job offer by firm $i \geq i^h$ is only accepted by unemployed workers and those employed at firms $i' \leq i$. Hence random contacts implies a job offer by hiring firm i is only accepted with probability $\alpha + (1-\alpha)G(i)$. Thus if state i firms have gross hiring flows $H(i,\Omega)$, it necessarily follows that the gross flow of job offers by such firms is $H(i,\Omega)/[\alpha + (1-\alpha)G(i)]$ where the denominator takes into account that not all job offers are accepted. Aggregating across i now identifies the total flow of job offers:

$$\lambda(\Omega) = \int_0^1 \frac{H(i,\Omega)}{\alpha + (1-\alpha)G(i)} di. \tag{12}$$

Furthermore monotonicity of the wage strategies implies the fraction of wage offers greater than w' is:

$$1 - F(w',\Omega) = \frac{\int_{i'}^1 \frac{H(i,\Omega)}{\alpha+(1-\alpha)G(i)} di}{\int_0^1 \frac{H(i,\Omega)}{\alpha+(1-\alpha)G(i)} di} = \frac{1}{\lambda}\int_{i'}^1 \frac{H(i,\Omega)}{\alpha + (1-\alpha)G(i)} di,$$

where i' solves $w(i',\Omega) = w'$. Hence the wage offer arrival rate for employed workers with current wage w' is:

$$\lambda_1[1 - F(w',.)] = \frac{\lambda_1}{\lambda}\int_{\widehat{i}(w',\Omega)}^1 \frac{H(i,\Omega)}{\alpha + (1-\alpha)G(i)} di. \tag{13}$$

Using (13) in (10) now implies Lemma 1 below.

Lemma 1. *A stationary equilibrium with monotone beliefs implies $w(i,\Omega)$ solves:*

$$w(i,\Omega) = \text{argmin}_{w' \geq w_{\min}} \left[w' + \left[\frac{\lambda_1}{\lambda}\int_{\widehat{i}(w',\Omega)}^1 \frac{H(i,\Omega)}{\alpha + (1-\alpha)G(i)} di \right] \min[v(i,\Omega), c_0] \right]. \tag{14}$$

We can now describe equilibrium wages. Consider any hiring firm $i > i^h$ and suppose it posts optimal wage $w' > w_{\min}$. The necessary condition for optimal w' implies

$$1 - \frac{\lambda_1}{\lambda}\frac{H(\widehat{i},\Omega)}{\alpha + (1-\alpha)G(\widehat{i})} c_0 \frac{d\widehat{i}}{dw'} = 0,$$

where paying a marginally higher wage signals a marginally higher productivity $\widehat{i}$ and so yields a corresponding marginal fall in the quit rate. But a stationary equilibrium requires

$w(i,\Omega)$ describes the solution to this first order condition. Because $\widehat{i}$ is the inverse wage function, the equilibrium wage equation must therefore satisfy the differential equation:

$$\frac{dw}{di} = \frac{\lambda_1}{\lambda}\frac{H(i,\Omega)}{\alpha+(1-\alpha)G(i)}c_0 \text{ for } i > i^h,$$

and integration now yields Proposition 2.[6]

Proposition 2 (**Optimal wages**). *A stationary equilibrium with monotone beliefs implies*

$$w(i,\Omega) = \begin{cases} w_{\min} & \text{if } i \in [i^c, i^h) \\ w_{\min} + \frac{c_0\lambda_1}{\lambda}\int_{i^h}^{i}\frac{H(j,\Omega)}{\alpha+(1-\alpha)G(j)}dj & \text{if } i \geq i^h. \end{cases} \tag{15}$$

Given these wage strategies, it is easy to verify that firms $i < i^h$ (who have $v < c_0$) strictly prefer to post wage $w = w_{\min}$ [paying a higher wage reduces the employee quit rate but the firm's gain by doing so $v < c_0$ and this wage deviation is profit reducing]. Conversely each hiring firm is indifferent to posting any wage $w \in [w_{\min}, \overline{w}]$. This property arises because all hiring firms face the same trade-off: paying a higher wage marginally reduces the quit rate but the replacement recruitment cost c_0 is the same for all.[7,8]

Because all (surviving) firms $i \in [i^c, i^h]$ strictly prefer to post wage $w = w_{\min}$, while all

[6]Proposition 2 implies equilibrium wage dispersion depends directly on c_0. If hiring costs $c_0 \to 0$, so that it is near costless to hiring replacement workers, then $\overline{w} \to w_{\min}$ and equilibrium converges to the Diamond paradox.

[7]The wage equation can also be usefully expressed in terms of the firm's equilibrium quit rate. Combining (15) with (17) (given below) yields $w(i,\Omega) = w_{\min} + c_0\left(q_0(\Omega) - q(i,\Omega)\right)$, where $q_0(\Omega) = \lambda_1(\Omega) + \lambda_u$ denotes the quit rate at non-hiring firms at the bottom of the ladder. This makes explicit the retention motive, by showing that firms achieve lower quit rates by paying higher wages.

[8]An extended approach might instead assume it takes time $\varepsilon > 0$ to fill a vacancy and so, taking foregone profit into account during the recruitment phase, the cost of a quit $c(i)$ is then strictly increasing in i. The arguments above would yield equilibrium wage equation $\frac{dw}{di} = \frac{\lambda_1}{\lambda}\frac{H(i)}{\alpha+(1-\alpha)G(i)}c(i)$. Although each hiring firm $i > i^h$ remains marginally indifferent to raising its wage to reduce its quit rate, equilibrium implies higher productivity firms strictly prefer to bid higher equilibrium wages for their replacement costs are greater. A useful interpretation of Proposition 2 is that it describes (the limiting) equilibrium when the time to fill a vacancy becomes small.

firms $i \geq i^h$ are indifferent to doing so, Proposition 1 implies (6) simplifies to:

$$\begin{aligned}(r+\alpha_\gamma+\alpha_a+\delta_F)v(i,\Omega) = \; & p^s(i) - w_{\min} - [\lambda_1(\Omega)+\lambda_u]\min[v(i,\Omega),c_0] \\ & + \mu_1 E \max[v(i,\Omega) - [c_0 + c^{JC}],0] - \delta_D E_c \min[v(i,\Omega), c^{JD}] \\ & + \alpha_\gamma \int_0^1 \max[v(j,\Omega),0] d\Gamma(j|i) + \alpha_a \sum_{s'} \Upsilon_{ss'}[v(i,\Omega(s'))] + \frac{\partial v(i,\Omega)}{\partial t},\end{aligned} \tag{16}$$

recalling that $\widehat{q}(w_{\min},\Omega) = \lambda_1(\Omega)+\lambda_u$, as implied by (4). The first line of (16) reveals the efficiency wage structure of this approach: that given firms set wages optimally, it is the outside job offer rate of employees, $\lambda_1(\Omega)$, which drives firm value and thus firm investment choices. Equation (19) below now closes the model by identifying a closed form solution for $\lambda_1(\Omega)$. Doing this reveals the important role played by vacancy chains.

3.4 Vacancy Chains

The equilibrium job offer rates λ_0, λ_1 depend on the aggregated hire flows of firms but the interaction between hires and quits is complicated: Equation (11) shows that hiring flows depend positively on quit rates due to firms which hire to replace quit workers. Additionally, standard job ladder logic means that a firm's quits are determined by the hiring rates of firms higher up the job ladder. Specifically, combining $q(i,\Omega) = \widehat{q}(w(i,\Omega),\Omega)$ with equations (4) and (13) yields

$$q(i,\Omega) = \lambda_u + \frac{\lambda_1}{\lambda}\int_i^1 \frac{H(j,\Omega)}{\alpha+(1-\alpha)G(j)}dj. \tag{17}$$

The interaction between (11) and (17) generates a multiplier effect where quits and hires positively affect each other. This has the potential to explain both their volatility and why they track each other over the cycle [as shown in Figure 3(b)]. The proof of Lemma 2 below identifies their closed form solution.

That solution, however, has a simple intuition which reflects the underlying vacancy chain process, and which we illustrate first. Suppose an existing firm $i > i^h$ creates a new job, either through investment in job creation or because an existing employee is separated into unemployment and the firm hires a replacement. This creates a vacancy chain. If the vacancy is filled by an already employed worker but at a firm $i' \geq i^h$, the "vacancy chain" survives in the sense that firm i' immediately creates a replacement post.

Conversely if the job is filled by an unemployed worker or a worker employed at $i < i^h$, the "vacancy chain" is destroyed for there is no replacement job. With stock-flow matching this process implies

$$\lambda_0 U + \lambda_1[1-U]G(i^h) = \int_{i^h}^1 \left\{ JC(i,.) + \lambda_u[1-U]G'(i) \right\} di \tag{18}$$

because the left hand side describes the flow destruction of vacancy chains which must equal the flow creation of vacancy chains, shown on the right hand side. Combining (18) with $\lambda_1(\Omega) = \phi\lambda_0(\Omega)$ yields the equilibrium job offer arrival rate for employed workers:

$$\lambda_1(\Omega) = \frac{\phi \int_{i^h}^1 \left\{ JC(i,.) + \lambda_u[1-U]G'(i) \right\} di}{U + \phi[1-U]G(i^h)}. \tag{19}$$

As $\lambda_0(\Omega) = \frac{1}{\phi}\lambda_1(\Omega)$, equation (19) also controls the re-employment rate of unemployed workers. There are three crowding out effects. First there is direct crowding out by other unemployed workers: ceteris paribus a higher U implies lower λ_0 and λ_1. Second if a job offer goes to an employed worker in a firm $i < i^h$, the vacancy chain is destroyed and the employment opportunity is lost. Thus $[1-U]G(i^h)$ also crowds out worker re-employment rates. But there is also wage crowding out. Consider for example the (most preferred) unfilled job $i = 1$. Because all want the best paid job, random offers implies the probability it is filled by an unemployed worker is $U/[U + \phi(1-U)]$ which is 37% according to the estimates in the quantitative section. In contrast only employed workers at firms $i \leq i^h$ will quit to a hiring firm paying minimum wage and so the probability a minimum wage job is filled by an unemployed worker is then $U/[U + \phi(1-U)G(i^h)]$ which, in steady state, is estimated at 83%. This behaviour generates the job ladder: on-the-job search especially crowds out unemployed worker job finding prospects at the highest wage paying firms.

Lemma 2 now identifies the solution for quits and hires across the firm distribution.

Lemma 2. *A stationary equilibrium with monotone beliefs implies equilibrium quit rate*

$$q(i,\Omega) = \begin{cases} \lambda_u + \lambda_1(\Omega) & \text{if } i \in [i^c, i^h) \\ \lambda_u + \frac{\phi \int_i^1 \{JC(j,.) + \lambda_u[1-U]G'(j)\} dj}{U + \phi[1-U]G(i)} & \text{if } i \geq i^h, \end{cases} \tag{20}$$

and corresponding hire rate

$$h(i,\Omega) = \begin{cases} 0 & \text{if } i \in [i^c, i^h) \\ jc(i,\Omega) + \lambda_u + \frac{\phi \int_i^1 \{JC(j,.) + \lambda_u[1-U]G'(j)\}dj}{U+\phi[1-U]G(i)} & \text{if } i \geq i^h. \end{cases} \tag{21}$$

Finally, the employed worker offer arrival rate is given by $\lambda_1(\Omega) = q(i^h, \Omega) - \lambda_u$.

Proof of Lemma 2 is in the Online Appendix. Equation (20) reflects the underlying vacancy chain process but from the perspective of a worker employed at a firm i. The integral in (20) describes the rate at which new jobs are created at higher wage firms $j > i$. For employee i, however, the relevant vacancy chain is destroyed once the job is either filled by an unemployed worker or by an employed worker at firm $i' < i$, for worker i is not interested in a replacement job $i' < i$. Thus $U + \phi[1-U]G(i)$ describes the death rate of employee i's vacancy chain and (20) then describes worker i's equilibrium quit rate.

The hiring policies described in Proposition 1 then yield (21) and note the multiplier effect. A favourable aggregate productivity shock implies job creation rates increase at all levels $i \geq i^h$. Hire flows increase at each $i \geq i^h$ not only because $jc(i,\Omega)$ increases but because greater job creation rates at firms $\widetilde{i} > i$ imply more quits and so hire flows additionally increase via quit replacement. A second important insight for what follows is that higher unemployment U not only crowds out unemployed worker job finding rates, it also reduces worker quit rates. The mechanism reflects the vacancy chain effect: higher unemployment directly shortens the average length of a vacancy chain and so reduces (replacement) hires. This explains why, in the data, worker quit rates are positively correlated with firm job creation rates but much more strongly [negatively] correlated with unemployment. We return to this issue in the quantitative section.

Given closed form solution (19) for $\lambda_1(\Omega)$ and the resulting turnover dynamics for (U, G), (16) determines $v(.)$ and so closes the model. For this case, however, the aggregate state includes the function $G(.)$ which is infinitely dimensional. The next section quickly specialises the model to finite firm productivity states and formally establishes that the aggregate state then reduces to a finite vector. This is important for it is then possible to estimate the model precisely using standard simulated method of moments.

3.5 Finite Productivities

For estimation purposes we specialise the model to finite I productivity states, so that firm productivity $p = p^{is}$ with $i \in \{1, 2, ..., I\}$. A very important property of the model is that the aggregate state now reduces to a finite vector, a result which holds even in the extended case with endogenous worker reservation wages $R = R(\Omega)$ and $\phi < 1$. This property does not arise in the standard Burdett and Mortensen (1998) framework, e.g. Coles (2001), Moscarini and Postel Vinay (2013) and Audoly (2020), and only holds in Coles and Mortensen (2016) for the special case $\phi = 1$.

The reason for the finite state space result is simple but subtle and reflects the replacement hiring process. Define N_i as the measure of workers employed in firms in state i, the employment vector $\underline{N} = (N_1, ..., N_I)$ where adding up implies unemployment $U = 1 - \sum_{i=1}^{I} N_i$. We now show the aggregate state reduces to vector $\Omega = (s, \underline{N})$ with corresponding vector of firm values $\underline{v}(\Omega) = \{v_i(\Omega)\}_{i=1}^{I}$, job creation rates $\{jc_i(\Omega)\}_{i=1}^{I}$ and so on as previously determined.

The first step is to extend the notation because firms in the same state i post different wages [i.e. there is equilibrium wage dispersion within each state i]. The cleanest approach is to assume firms select wage strategies as follows: i) On start-up, a firm is allocated a wage rank $\chi \sim U[0, 1)$. In the stationary equilibrium, firm (i, χ, n, Ω) posts wage with rank χ in the firm i wage distribution. ii) On receiving a firm specific productivity shock with updated productivity i' the firm also updates to a new wage rank $\chi' \sim U[0, 1)$. Because all χ-wage strategies yield equal value, such wage selection is consistent with equilibrium. We choose this wage selection process because it guarantees first order stochastic dominance in wages, and so a worker will always quit to a higher wage offer.

To match to the previous notation, consider the following partition of line $[0, 1]$ into a grid $\{x_0, x_1, .., x_I\}$ where $x_0 = 0$, $x_i = x_{i-1} + \gamma_{0i}$ and $x_I = 1$. A firm in state $i \in \{1, 2, ..., I\}$ with wage rank $\chi \in [0, 1)$ is correspondingly defined as being in state $x \in [0, 1]$ where $x = x_{i-1} + \chi[x_i - x_{i-1}]$. Each start-up is then equivalently defined as having initial state $x \sim U[0, 1]$, where $p^s(x) = p^{is}$ for $x \in [x_{i-1}, x_i) \subset [0, 1]$. The only material difference to the previous section is that firm productivity $p^s(x)$ is increasing in $x \in [0, 1]$ but not strictly increasing. The underlying wage structure (3), however, continues to apply and (16) describes the equilibrium values $v_i(\Omega)$.

So why is the state space finite? The critical property is that despite there being wage and quit rate dispersion across firms $x \in [x_{i-1}, x_i)$ within a productivity level, all such firms have identical expected employment dynamics. Why? Because all firms with $i \geq i^h$ immediately replace any worker who quits and so their expected employment dynamics are independent of χ. Additionally, all firms with $i < i^h$ post the same wage, $w_{\min}$ and so their expected employment dynamics are also independent of χ. Now recall that $G(x)$ describes the distribution of employment across firms $x \in [0, 1]$. By definition of the partition above, firm $x = x_i$ has productivity $i + 1$ and rank 0 and so $G(x_i) = \sum_{j=1}^{i} N_i/(1-U)$. Because firm size is orthogonal to rank we then have

$$G(x) = \frac{\sum_{j=1}^{i-1} N_j + \frac{x - x_{i-1}}{x_i - x_{i-1}} N_i}{1 - U} \text{ for all } x \in [x_{i-1}, x_i).$$

and so $\underline{N}$ is indeed a sufficient statistic for $G(.)$. Thus with finite productivity states the previous analysis all goes through with the added simplification that the aggregate space is a finite vector $\Omega = (s, \underline{N})$.

4 Quantitative Analysis

We estimate the model using simulated method of moments and targeting a wide range of worker and job flows as well as firm dynamics moments for the US economy. We use data from the Business Dynamics Statistics (BDS), Job Opening and Labor Turnover Survey (JOLTS), Compustat and the Current Population Survey (CPS) for the period 1990Q2 to 2018Q4. In Sections 4.1 and 4.2 we describe the estimation, with further details in the Online Appendix. Steady state results are presented in Sections 4.3 and 4.4, and business cycle results in Section 5.

4.1 Parameterization and pre-set parameters

We set a period to be equal to a month and the time preference parameter to $r = 0.0043$ to match a yearly discount rate of 5%. We assume $S = 3$ aggregate productivity states indexed by $s = 1, 2, 3$. Let a_s denote the aggregate productivity shifter in state s such that it follows a discretised AR(1) process, where a new value is drawn at rate $\alpha_a = 1/3$ from the transition matrix $\Upsilon_{ss'}$. The latter is obtained using a Rouwenhurst approximation

for a given autocorrelation parameter ρ_a and variance parameter σ_a. These parameters are chosen to match the persistence and variance of a_s to that of aggregate output per worker in the data.[9] We further suppose $I = 5$ firm productivity states to keep the state space reasonably tractable while allowing us to match well the microdata, indexed by $i = 1, 2, ..., I$. A firm in state (i, s) has productivity $p^{is} = a_s p_i$ and equilibrium implies the aggregate state is $\Omega = (s, \underline{N})$ where $\underline{N} = \{N_1, N_2, .., N_5\}$ is the vector of employment across states i. Throughout we restrict parameter values so that $v_i(\Omega) > 0$ for all Ω, otherwise entire productivity bins of firms instantaneously close down in the event of an adverse aggregate shock which generates discrete pulses of job destruction which are much too volatile for the data.

We now turn to describe firms' microturnover structure. Define mature states $I^m = \{2, 3, 4\}$. While mature firms may transition across these states, we assume any firm in a mature state $i \in I^m$ cannot transit to states $i = 1, 5$; i.e. only entrant firms may have the extreme productivities $i = 5$ [gazelles] and $i = 1$ [struggling entrants]. Allowing more extreme productivity states for entrant firms is important to account for the difference in growth outcomes between new start ups and existing (mature) firms. The reason for 3 mature states $i \in I^m$ is to capture disperse job creation and job destruction outcomes across mature firms. Specifically, the parameters of the model are chosen to impose the following properties of each bin in the non-stochastic steady state. Mature firms $i = 4$ have positive expected growth: they occasionally create new jobs and do not destroy jobs. Firms in state $i = 3$ are instead in expected decline, with positive job destruction and zero job creation. They are, however, still profitable enough that they replace workers who quit, and thus shrink only due to layoffs. Firms in state $i = 2$ are the least productive, so much so that they do not replace workers who quit, and additionally have the highest job destruction rate. Thus, we impose in the estimation that $i^h = 3$ in steady state, meaning that only firms with $i = 1, 2$ do not replace worker quits. The mature states $i \in I^m$ are a convenient shorthand for describing different types of (mature) firm behaviors: some are growing, others are declining, with firm decline being driven by a mix of layoffs and unreplaced quits.

[9] Output per worker in the data is constructed using the ratio of real GDP over total employment. In the model, output per worker and the productivity shifter a_s are not identical, due to endogenous composition effects. However, the differences are small (see Table 2) so we calibrate the process of a_s directly in order to remove ρ_a and σ_a from the estimation routine and save on computation.

Conditional on survival, all firms receive a firm specific productivity shock at rate $\alpha_f = \frac{1}{3}$ [i.e. roughly once a quarter], which is within the range of estimates from the data (see Online Appendix). We assume firm $i \in I^m$ transits to state $j \in I^m$ with probability γ_j and so is independent of i. For parsimony we simply set $\gamma_3 = \gamma_4 = \frac{1}{2}[1 - \gamma_2]$. For firms $i = 5$ then γ_{55} describes the probability the firm remains a gazelle and so determines gazelle persistency. With probability $1 - \gamma_{55}$, the firm otherwise becomes a mature firm $j \in I^m$ in proportions γ_j. Similarly if instead in state $i = 1$, γ_{11} is the probability the firm remains as a struggling entrant. With probability $1 - \gamma_{11}$, the firm instead becomes mature $j \in I^m$ in the proportions γ_j. In this way the transition matrix $\{\gamma_{ij}\}$ is fully described by the choice of just three parameters $(\gamma_2, \gamma_{55}, \gamma_{11})$. Consistent with this structure, we assume an entrant firm is a gazelle with probability γ_{05}, a struggling entrant with probability γ_{01} and with complementary probability $(1 - \gamma_{05} - \gamma_{01})$ is a mature firm in states $j \in I^m$ with proportions γ_j. Our quantitative results show that this very simple turnover structure is sufficiently rich to capture the firm age and size dynamics described in the data.

The specification of distributions H^{JC}, H^{JD} is central to determining the response of job creation and job destruction gross flows to aggregate shocks. Following Coles and Moghaddasi (2018), we specify distributions $H^{JC}(.), 1 - H^{JD}(.)$ which are isoelastic with respective elasticities ξ_{JC}, ξ_{JD}. In particular, for firms with $i \geq i^h$ the firm level job creation and job destruction rates are considered as $jc_i(\Omega) = \mu_1[v_i(\Omega) - c_0]^{\xi_{JC}}$ and $jd_i(\Omega) = \delta_{JD}[v_i(\Omega))]^{-\xi_{JD}}$. A useful motivation is that the typical free entry approach assumes the job creation margin is infinitely elastic; i.e. entry is infinite when $c_0 < v(\Omega)$ and so free entry implies $v(\Omega) = c_0$. Our estimation instead identifies the job creation elasticity ξ_{JC} so that implied gross job creation flows are indeed consistent with the data, and similarly with ξ_{JD} and job destruction flows. Heterogeneity in firm types also allows a cleansing effect of recessions: a negative aggregate productivity shock will trigger relatively high job destruction rates in the lowest surplus states $i = 1, 2$, though such flows may be relatively short-lived if employment $N_1 + N_2$ quickly falls.

An important challenge is to make the notions of job creation and destruction in our model equivalent to how they are measured in the data. To do this, we make assumptions so that each productivity bin only performs job creation *or* job destruction, but not both. This is achieved by assuming a lower support for H^{JC} of $\underline{c}^{JC}$ and an upper support for

H^{JD} of $\bar{c}^{JD}$. We place these support parameters to ensure that, in steady state, only firms in states 1, 2 and 3 destroy jobs after a δ_D job destruction shock, and only firms in states 4 and 5 create new jobs after a μ_1 job creation shock.

To estimate the entry process we parametrise $H^E(\Pi^{SU}(\Omega)) = (\Pi^{SU}(\Omega))^{\xi_E}$ to also generate a constant entry elasticity with respect to the expected value of entry. For quantitative reasons we additionally modify the model so that new firms enter with two unemployed workers, rather than one, which raises the minimum size of entrants to two workers. In our model, if a firm ever reaches zero employees, the constant returns to scale structure implies that this is an absorbing state, where the firm never produces or has positive employment again. We thus treat these firms as having exited, and so we measure exit in our model as both the bankruptcy shock and firms who drop from one to zero employees. We set $\delta_F = 0.0004$ to give a 0.1% yearly exit rate from this shock, to roughly match the very low exit rate among firms with more than 500 employees in the BDS data. The remainder of exit in the model is driven by job destruction.

4.2 Estimation Strategy

The above functional forms and the remaining non-pre-set structural parameters of the model lead to a set of 23 parameters to estimate. Table 1 present all these parameters and their corresponding targeted moments. Here we briefly discuss the targeted moments, noting that parameters are jointly estimated.

Firm productivity process To recover the entrant's productivity process and their maximum number of unfilled positions, N_0, we target the age and employment-age distributions of firms in 2005 obtained from the BDS. In particular, we target the fraction of firms at age 0, 1, 2, 3-5, 6-10 and 11-15, the fraction of aggregate employment at firms in the same age groups. The productivities of the entrant-specific productivity states p_1 and p_5, control how large is JC and JD among new entrants relative to older firms. N_0 controls the average initial size of entrants, particularly measured size at age 0. The transition probabilities γ_{11} and γ_{55} control for how long entrant JC and JD rates remain elevated. The entrant probabilities γ_{01} and γ_{05} control whether this is experienced by a small or large share of entrants. In the data, the shares of employment by firm age imply the net job creation rates of firms of different ages; while the shares of firms by age imply the

Table 1: Parameter values and target moments

	Interpretation	Value	Source
	Pre-set parameters		
r	Discount rate	0.0043	5% annual interest rate
α_a	Arrival rate of agg shocks	0.3333	Normalisation
α_γ	Arrival rate of firm prod shocks	0.3333	Autocorr. of idiosyncratic prod. (see text)
ρ_a	Persistence of aggregate productivity shock	0.7800	Autocorr. of aggregate labour prod.
σ_a	Std. of aggregate productivity shock	0.0120	Std. of aggregate labour prod.
δ_F	Arrival rate of bankruptcy shock	$8.3E-05$	Exit rate of firms with size > 500
	Firm productivity process		
p_1	Prod. in state 1	0.9265	Firm age distribution
p_2	Prod. in state 2	0.6747	Firm age distribution
p_3	Prod. in state 3	0.7133	Std. of idiosyncratic labour prod. = 30%
p_4	Prod. in state 4	1.3076	Normalise $Y/N = 1$
p_5	Prod. in state 5	1.1379	Firm age distribution
γ_{11}	Prob. remain in state 1	0.9000	Firm age distribution
γ_{55}	Prob. remain in state 5	0.4885	Firm age distribution
γ_{01}	Share born with prod. 1	0.5191	Firm age distribution
γ_{05}	Share born with prod. 5	0.1001	Firm age distribution
γ_2	Prob. mature draw state 2	0.158	80% of quits replaced
N_0	Potential unfilled positions of entrants	51.652	Firm age distribution
	Cost structure of JC and JD		
ξ_e	Elasticity of entry with respect to firm value	3.5139	Std deviation of cyclical firm entry
ξ_{JC}	Elasticity of JC with respect to firm value	2.2113	Std deviation of cyclical JC
ξ_{JD}	Elasticity of JD with respect to firm value	3.7975	Std deviation of cyclical JD
$\underline{c}^{JC}$	Lower bound of $H^{JC(.)}$	0.9504	JC only in states 4 and 5
$\bar{c}^{JD}$	Upper bound of $H^{JD(.)}$	1.7301	JD only in states 1, 2 and 3
μ_0	Firm entry flow	0.0004	Average firm size 22.4 employees
μ_1	Arrival rate of capital investment shock	0.0114	Average JC rate = 2.17% per month
δ_D	Arrival rate of capital destruction shock	0.0144	Average JD rate = 2.43% per month
c_0	Worker replacement cost	0.7797	Autocorr. of JC and JD
	Worker turnover		
λ_u	Quit rate to unemployment	0.0087	Worker EU rate 7.78% per quarter
ϕ	Employed fixed search intensity	0.1027	EE rate of 6.81% per quarter
$w_{\min}$	Minimum wage	0.6515	Labour share = 2/3

Calibrated parameter values and source moments. See text and Online Appendix for further details.

different exit rates across age groups, thus informing these parameters.[10]

In the case of mature firms we use several moments to identify their productivities and stochastic process. We use p_4 to normalise aggregate labour productivity in the model to one: $Y/N = 1$. We use p_3, particularly its value relative to p_4, to target the standard deviation of within-firm labour productivity obtained from Compustat, which we estimate

[10]The estimation finds that the productivity grid is non-monotone, as $p_1 > p_2$ and $p_5 < p_4$. Nonetheless, firm *values* remain monotone, with $v_i < v_{i+1}$ for all i, which is sufficient for the job ladder to be directed monotonically by i, and hence for our notion of equilibrium to remain well defined. The disconnect between the ordering of productivities and values occurs simply because the entrant states $i = 1, 5$ are more persistent than the mature states.

to be 30% (see Online Appendix). Further, noting that p_2 controls the overall JD and exit rates of state 2 firms we target the average exit rate of firms consistent with on the overall distribution of firms by age. The high JD rate in this state means that it drives 37% of total exit, despite only containing 14% of firms. γ_2 controls the equilibrium mass of firms in state 2, who choose not to replace workers who quit, and so the higher is γ_2, the larger the fraction of firms who do not replace quit workers. Using data from Davis et al. (2012), we estimate that firms replace around 80% of workers who quit (see Online Appendix) and target this fraction.[11]

Cost structure of JC and JD To inform the JC and JD elasticity parameters ξ_{JC} and ξ_{JD} we target the volatility of the JC and JD rates series obtained from Davis et al. (2012) using JOLTS data. Both in the model and data, we use an HP-filter with parameter 10^5 to obtain the cyclical component these series and compute its standard deviation. The arrival rates μ_1 and δ_D control the average JC and JD rates, respectively. We target a monthly JD rate of 2.43%, the average documented in Davis et al. (2012) based on JOLTS data. We also target a 5.73% unemployment rate, the average from the CPS in our sample, which requires a 2.17% monthly JC rate (excluding JC from firm entry) to balance employment flows in steady state. We are careful to account for other sources of job creation (e.g. firm entry) and job destruction (e.g. unreplaced quits) when computing these series in the model.

To inform the firm entry elasticity ξ_e we target the volatility of the firm entry series obtained from the BDS. The firm entry flow shifter, μ_0, is chosen to control the number of firms in equilibrium. For given targeted total employment N, μ_0 controls the equilibrium number of firms, and hence the average employment per firm. We choose μ_0 so that average firm size (total employment / total number of firms) is equal to 22.4, as in the BDS data in 2005.

To inform the worker hiring cost, c_0, we target the autocorrelation of aggregate JC and JD rates in the data. Intuitively, the larger is c_0 the larger is the general equilibrium effect that rising unemployment in a recession raises firm value, since firms must pay the replacement cost c_0 less often in recessions. The larger is this offsetting effect, the less

[11] We additionally validate this estimate by comparing other measures of replacement hiring in our model to the estimates of Elsby et al. (2021) in the data.

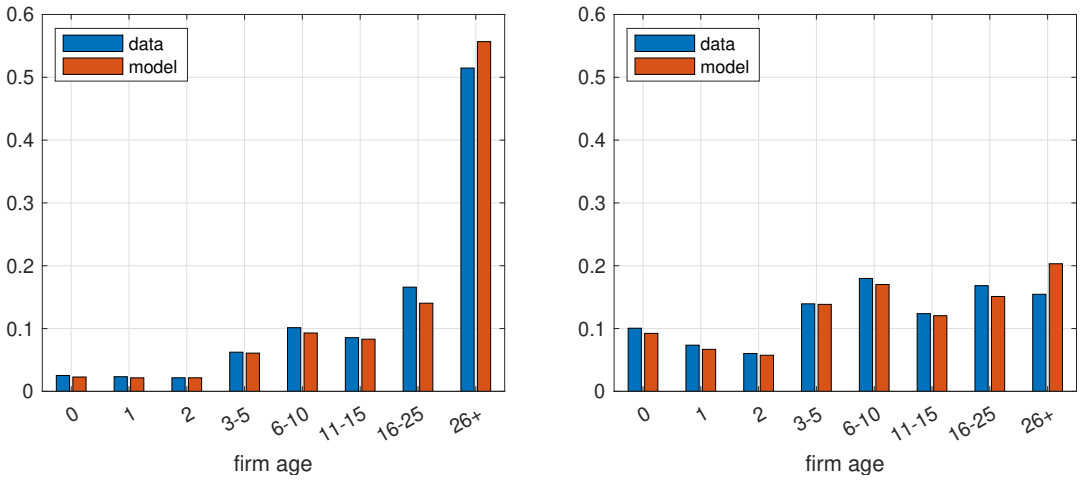
persistent are JC and JD rates, as they recover faster in recessions.

Worker turnover To estimate the worker turnover parameters, we require the model to be consistent with the average worker transition rates from Davis et al. (2012). In particular, we target a quarterly 7.78% layoff rate, which we equate with the EU rate in our model. We match this rate using layoffs due to job destruction from firms in states $1, 2, 3$ and add exogenous separations, λ_u, to capture that in growing firms in states $4, 5$ some workers transit into unemployment. The relative search intensity of employed workers, ϕ, is then chosen to match the quarterly quit rate of 6.81%. Finally, the exogenous minimum wage, $w_{\min}$, is used to target a labour share of $2/3$.

4.3 Steady state results: Micro firm growth rates

The data find that 9.8% of existing firms close per year which, in a steady state, are replaced by an equivalent inflow of new start-up firms. Figure 4 shows that the model fits very well the firm age distribution. Figure 4(a) describes the fraction of workers employed in firms within a particular age range, while Figure 4(b) describes the fraction of firms in that age range. Taking into account the range of each age bin, Figure 4(b) implies the average death rate of firms falls steeply with age. Allowing ex-ante start-up heterogeneity is central to capturing this age structure: some start-ups are born struggling entrants $[i = 1]$ which quickly die, others are longer-lived gazelles $[i = 5]$, while the rest enter the mature states I^m directly. Matching to the firm survival data, the estimated model finds 52% of new start-ups are born struggling entrants with a high associated firm death rate. Conversely firms which survive to age 25 are on average very large: more than 50% of all workers are employed in firms over 25 years old, yet there are relatively few such firms. Although the survival rate of gazelles is high, the relative scarcity of old but very large firms implies that the gazelle state cannot be highly persistent. Matching data moments finds around 10% of new start-ups are born as gazelles but, conditional on survival, the expected duration of a gazelle is short, being only 2 months ($\gamma_{55} = 0.49$). In this way few start-ups remain gazelles for long and relatively few are fortunate to grow into very large firms. Conversely the struggling entrant state $i = 1$ is estimated to be much more persistent with an expected duration of 10 months ($\gamma_{11} = 0.9$). Their high death rate then implies a struggling entrant is more likely to close than reach a mature state.

Figure 4: Firm age distributions in the model and data



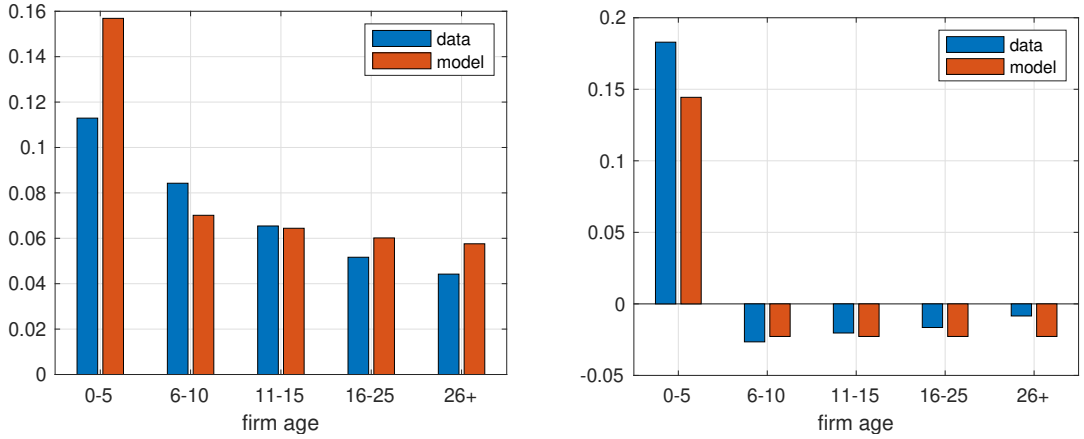
(a) Employment (b) Number of firms

Left (right) panel plots the fraction of total employment (firms) contained in each firm age bin. Data corresponds to the BDS data for firms in 2005. Model corresponds to the steady state of the model.

Given this firm heterogeneity structure, the estimated parameters imply that most employment N_i is in the mature states $i \in I^m$ and so job creation flows by mature firms (specifically, JC_4) are responsible for the larger part of gross job creation. But mature firms are also responsible for the larger part of gross job destruction flows through $JD_2 + JD_3$. Combining these two effects, Figure 5(a) shows that the implied *net* job creation flows of mature firms are negative as in the data. It is only young firms who have positive net job creation, due to the extra job creation of entrants, and in particular gazelles. In this way a typical gazelle life cycle is to burn brightly for a relatively short while, during which it becomes a net job creator, but as it matures it ultimately enters an (ergodic) phase of general decline. At the same time, the model remains consistent with the higher exit rates of young firms due to non-gazelle entrants, as shown in Figure 5(b).

An important insight is that despite firm closure rates being high the amount of job destruction due to such closures is surprisingly small for most firm closures are small firms. For example, according to the 2005 BDS data, firm exit rates in the 1-4 employee size category is 12.3% per annum, while the exit rate for all larger size categories is much smaller, for example it is only 2.9% in the next size bin of firms with 5-9 employees. The large 9.8% closure rate thus reflects that 88% of firm closures involve the very smallest of firms. Although average firm closure rates are high, the model confirms that the larger

Figure 5: Net job creation and exit in the model and data

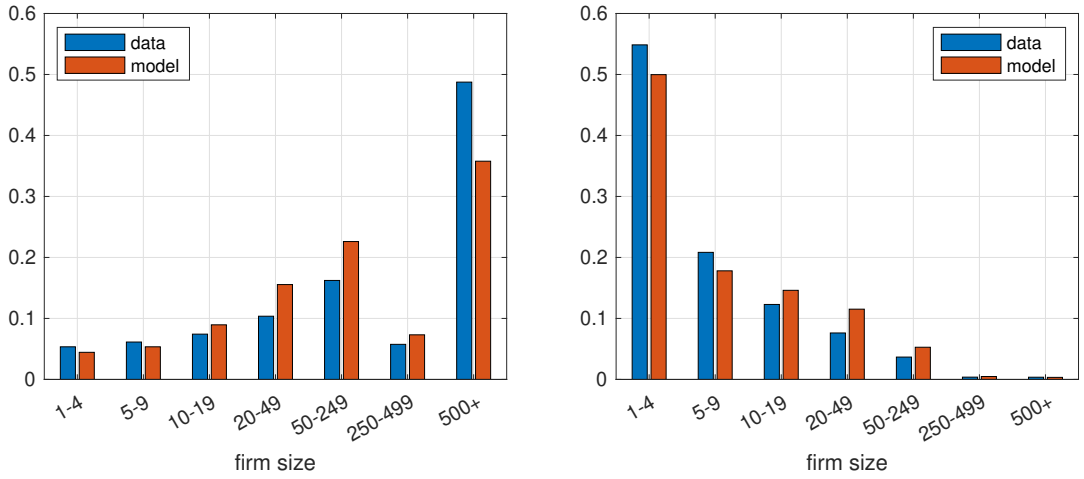


(a) Net JC rate (b) Exit rate

Left panel plots the yearly net job creation rate in firm age bin, computed as the net JC flow divided by the employment denominator. Data corresponds to the BDS data for firms in 2005. Model corresponds to the steady state of the model. Right panel plots the fraction of firms who exit per year. Model yearly rates computed as $1 - e^{-12r}$, where r are the theoretical monthly rates.

component of job destruction is the gradual downsizing of employment by larger state $i = 2, 3$ mature firms.

Figure 6: Firm size distributions in the model and data



(a) Employment (b) Number of firms

Left (right) panel plots the fraction of total employment (firms) contained in each firm size bin. Data corresponds to the BDS data for firms in 2005. Model corresponds to the steady state of the model.

Although the model is matched to employment by firm age and survival rates, it is not matched to the distribution of firm size nor employment by firm size. The Gibrat's Law structure automatically implies substantial size dispersion across firms of the same

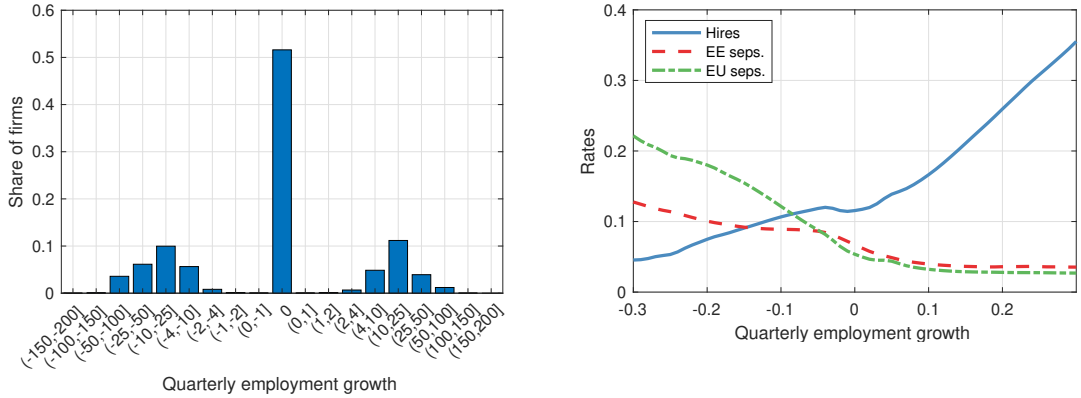
age, because realised firm size depends on the firm's history of productivity outcomes and growth rates. If the model implied firm growth processes were a poor descriptor of actual firm size evolution, the simulated firm size distributions would not likely match the actual distributions of employment across firms. The match turns out to be very good indeed, as shown in Figure 6. The figure reveals that in the data the majority (over 55%) of firms are very small, in the 1-4 employee bin and constitute only 5% of employment. In contrast, nearly 50% of total employment is in firms which employ more than 500 workers while the number of such firms is very small. This property of the labour market is well known. The important point, however, is that the growth structure here also replicates the (un-targeted) distributions of firm size and of employment by firm size. This provides (at least indirect) support for the Gibrat's Law approach taken here.

Finally Figure 7(a) shows the (micro) firm growth structure is also consistent with the (untargeted) empirical employment growth distribution documented in Davis et al. (2012) and Elsby et al. (2021).[12] A key feature of the data is that around 55% of firms report zero growth (see Elsby et al., 2021), while a somewhat equal share of the remaining firms report either positive growth or negative growth. The model yields precisely this outcome: employment at most firms does not change over the year with an approximately even break of firms showing positive and negative growth. Of course for firms where employment does not change, hiring is not zero for those firms actively replace workers who quit. Nevertheless the important point, as also argued in Bertola and Caballero (1994) and Cooper and Halitwanger (2006), is that many firms do not change employment and the smooth evolution of aggregate unemployment arises because of the aggregation of disperse employment decisions made by heterogeneous firms. A representative firm approach with smooth, convex adjustment costs is inconsistent with microeconomic behaviour.

We now turn to describing the job ladder structure of the model and how we match data to information on quit and hire outcomes at the firm level.

[12]We refer the reader to these papers to view the empirical employment growth distributions based on JOLTS and the Quarterly Census of Employment and Wages.

Figure 7: Employment growth distribution and relation with worker flows



(a) Employment growth (b) Flows

Left panel plots the distribution of quarterly employment growth rates across firms, excluding entry and exit, as in Elsby et al. (2021). Right panel plots the average hiring, EE and EU separation rates of firms with each employment rate (computed for bins of width 0.01 and smoothed with a 10 bin moving average). Both figures calculated from a simulation of a panel of firms in the steady state of the model.

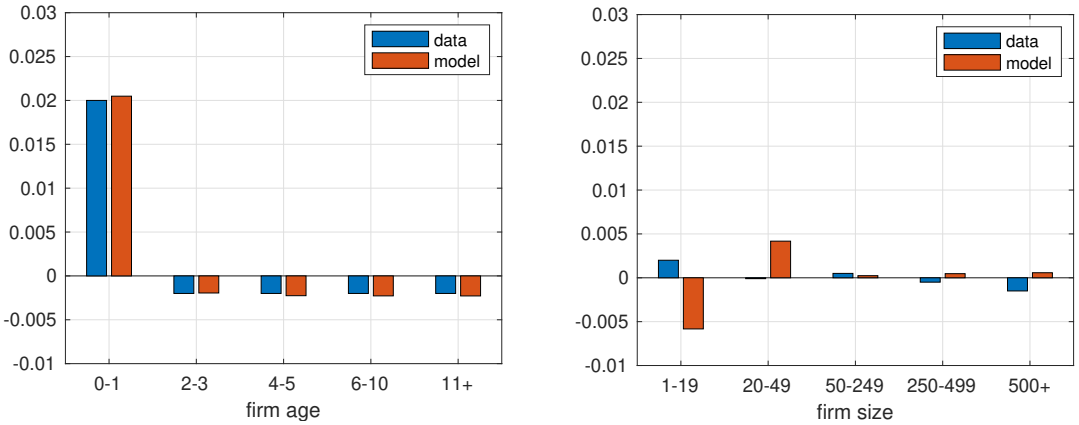
4.4 Steady state results: Quit Turnover and Wage Competition

Quits are costly: a firm must either pay a cost c_0 to hire a replacement worker, or choose not to replace the worker and downsize. Although job creation rates jc_i and job destruction rates jd_i are the same for all firms i, wage competition with on the job search implies firms in the same state i post different wages $w \in [w_i, w_{i+1})$, where paying a marginally higher wage marginally reduces worker quit rates. An important difference to the Burdett and Mortensen (1998) framework, however, is that wage offers here are not fundamentally related to firm size. Instead higher productivity firms post higher wages which fundamentally changes the structure of quit turnover dynamics over the cycle. Specifically it allows that small but fast growing gazelles poach employees from lower productivity but possibly larger firms. Haltiwanger et al. (2018) show this is an important property of the data: there is clear evidence that workers typically quit to better wages, but there is no evidence of a systematic drift of workers from small to large firms.

Our estimation targets the average turnover rates in the economy, but is also consistent with un-targeted turnover properties across the firm growth rate distribution. Davis et al. (2012) document "hockey sticks" relationships between hires, layoffs and quits and firms' growth rates. Figure 7(b) shows that the model also generates such hockey sticks. Job-to-job quit rates decline as we move from the [low wage] negative growth firms to the

[high wage] positive growth firms; i.e. the highest growth firms pay the highest wages, have the lowest quit rates and expand with a high hiring rate. Figure 7(b) also shows that the firms with the largest negative growth rates shrink more by layoffs than quits. Thus although unreplaced quits describe a significant channel for job destruction, high separation rates at the faster declining firms instead depend more on high layoff rates, consistent with the hockey stick relationships in the data.

Figure 8: Poaching flows by age and size



(a) Age (b) Size

Left (right) panel plots the quarterly net poaching rate by firm age (size) bin. This is computed as total hires from poaching less total separations from poaching as a fraction of employment in each bin. Data are for 2005, from the Census job-to-job database. Model corresponds to the steady state of the model, with quarterly rates computed $1 - e^{-3r}$, where r are the theoretical monthly rates.

Figure 8 shows that the model also exhibits a job ladder by firm age. Following Haltiwanger et al. (2018), in this case we measure the job ladder through a firm's net poaching rate, defined as the difference between the rate at which it hires employed workers h_p less the number of workers it loses to other firms s_p and so describes net quit drift. These data were not targeted. By firm age, the model generates the observed poaching structure: young firms are large net poachers [from older firms] while older firms are net losers to young firms. This large net poaching figure reflects that gazelles [10% of entrants] will hire many new workers while struggling entrants [52% of entrants] have few workers who can be poached.

Because firm specific productivity, and so firm growth, is positively autocorrelated, the model implies a positive, but weak, correlation between firm size and productivity. The wage setting process, in turn, then generates a positive correlation between firm size

and wages, where a one standard deviation rise in firm size leads to a 14% standard deviation rise in wages (see also Brown and Medoff, 1989).[13] But job-to-job quit turnover is confounded by entrant gazelles who are currently small but responsible for most of the net poaching in the data. As described in Haltiwanger et al. (2018), there is no simple relationship between the job ladder and firm size.

5 Business Cycle Facts and Insights

This section considers the business cycle properties of this approach, its match to the data and corresponding new insights into business cycle frequencies [further details are provided in the Online Appendix]. Table 2 describes the business cycle facts, where moments marked with † symbols are targeted, the rest are untargeted. The first (data) rows record the estimated volatility and persistence of key aggregate variables. Consistent with Figure 1 in the Introduction, the job destruction rate $(jd = JD/N)$ is more volatile than the job creation rate $(jc = JC/N)$ and, perhaps surprisingly given the insights of Mortensen and Pissarides (1994), the job destruction rate is also the more persistent. Importantly for what follows, notice that jc is the least persistent time series of all, while unemployment U is both the most persistent and has the greatest volatility.

In the model aggregate productivity [the sole driving variable] evolves exogenously with a persistence $(\rho = 0.85)$ and volatility $(\sigma = 0.01)$ matched to that of measured aggregate productivity Y/N. Note these productivity shocks are small and less persistent than unemployment, though more persistent than jc, jd [both as measured in the data and the simulated model]. The only targeted moments in Table 2 are the volatility and persistence of job creation, job destruction and layoff rates.[14] Yet despite only targeting those

[13] We regress wage on firm size (measured as number of employees) in the ergodic distribution. Specifically, we construct an average wage for every firm size on our firm size grid, by integrating over the within-size productivity and rank distribution. We then regress this average wage on firm size using weighted OLS, with one observation per size node, with the weight for each size node given by its density in the ergodic distribution.

[14] Note that while the simulated persistence (ρ_{t-1}) of job creation and job destruction are of similar magnitudes, in the data the persistence of job creation is about half of that of job destruction. A key reason why the model does not produce a better fit in this dimension is because we only use the parameter c_o to capture the cyclical behaviour of both jc and jd. The estimation procedure tries to resolve this tension by choosing a c_o that places the values of (ρ_{t-1}) for jc and jd somewhere in the middle of their empirical targets. Further, as job destruction flows in the model are mostly made up of workers laid-off after a δ_D shock (70% of all job destruction), the cyclical behaviour the layoff rate then follows closely that of jd.

Table 2: Logged and HP-filtered Business Cycle Statistics. Data and Model

	jc	jd	quits	hires	layoffs	UE	U	N	net jc
Volatility and Persistence									
Data									
σ	0.042†	0.055†	0.116	0.058	0.048†	0.169	0.204	0.018	0.006
ρ_{t-1}	0.433†	0.755†	0.945	0.904	0.761†	0.959	0.977	0.973	0.722
Model									
σ	0.041	0.051	0.177	0.077	0.046	0.218	0.221	0.014	0.005
ρ_{t-1}	0.674	0.595	0.934	0.899	0.546	0.936	0.934	0.937	0.468
Cyclical Correlation									
Data									
U	-0.305	-0.254	-0.923	-0.760	0.174	-0.972	1.000	-0.922	0.012
N	0.061	0.386	0.780	0.539	-0.015	0.867	-0.922	1.000	-0.230
net JC	0.751	-0.870	0.245	0.511	-0.760	0.120	0.012	-0.230	1.000
Model									
U	-0.616	-0.571	-0.989	-0.940	0.305	-0.994	1.000	-0.988	0.054
N	0.602	0.559	0.983	0.925	-0.294	0.985	-0.988	1.000	-0.053
net JC	0.751	-0.847	0.086	0.287	-0.933	0.054	0.054	-0.053	1.000

Time series in the model are obtained by simulating the model for 1,000 years and then aggregating to quarterly frequency as in the data. The cyclical components of the (log) of these time series are obtained using an HP filter with parameter 10^5. Net jc ($jc - jd$) is not logged as it takes negative values. jc, jd, quits, hires, and layoffs are rates relative to employment, and UE gives the job finding rate of unemployed, H^{UE}/U. At any time t the flows refer to flows between t and $t+1$, and the stocks (U and N) are measured at t. Moments targeted in the estimation are marked with a † symbol.

moments, Table 2 reveals the model's remarkable success at qualitatively matching the untargeted persistence, volatility and cyclical properties of the remaining key aggregate variables. For example the framework matches the fact that hire and quit rates are much more persistent and more volatile than job creation jc and job destruction jd rates. Similarly the framework also generates the wide and persistent variation in unemployment U and job finding UE rates measured as H^{UE}/U.[15]

Most importantly Table 2 identifies a new data property which is central for understanding the macroeconomic properties of unemployment:

- **net job creation is strongly positively autocorrelated and uncorrelated with unemployment**.

Given the dynamic specification $\Delta U_t = JD_t - JC_t$, it is this property of the data which

[15]Not shown in the table are the dynamics of firm entry and exit, which are instead calculated at an annual frequency as in the BDS data. The model closely replicates the standard deviation of entry, which is 0.0688 in the model and 0.0720 in the data. The firm exit flow is untargeted, and the model generates a standard deviation of 0.0193, which is somewhat lower than the 0.0656 in the data. However, the model successfully matches that firm exit is less volatile than firm entry, as in our data and discussed by Lee and Mukoyama (2015).

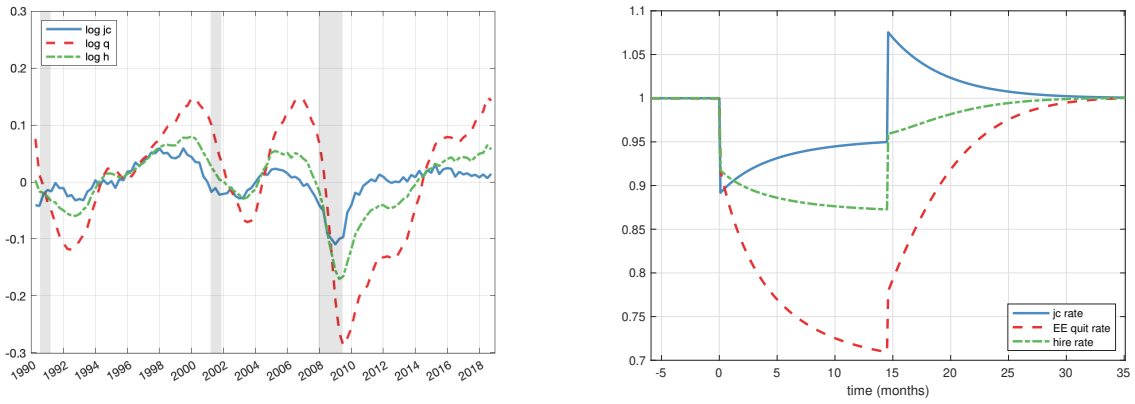
explains why estimated reduced form unemployment dynamics are typically close to containing a unit root. Figure 2 in the Introduction shows the unemployment process is stable and so there is no unit root. However it also confirms that recessionary shocks generate large and persistent unemployment loops. And by generating qualitatively identical net job creation dynamics, the estimated model also generates large and persistent unemployment loops with a volatility and persistence of unemployment comparable to that of the data. Of course the central question is why does this occur? According to the model it is due to the efficiency wage structure, that wages are not competitively determined. The underlying structure is not dissimilar to the insider/outside approach where each firm pays its employees the same wage. Except here this is optimal because the firm faces a trade-off between reduced employee quit rates and higher turnover costs. But a standard criticism of the typical insider/outsider approach is why doesn't the firm charge "outsiders" a job fee? An important difference here is firm productivity is private information. For example if workers are willing to pay jobs fees, what is to stop a struggling entrant from scamming workers by initially paying high gazelle wages, collecting high entry fees from new employees, only to subsequently declare an unfavourable productivity shock and so close down? Given the natural reticence of workers to pay job fees, the efficiency wage structure reduces wage flexibility over the cycle. The resulting net job creation dynamics then ensure large recessionary shocks generate large unemployment loops.

5.1 Replacement hiring and vacancy chains

This section describes the interaction between quit turnover, replacement hiring and vacancy chains over the cycle. Figure 9(a) plots the (log) job creation jc, gross hire h and gross quit q rates for the US (1990Q2-2018Q4) while Figure 9(b) describes the model's "impulse" response to a recessionary shock. Specifically Figure 9(b) supposes the model is initially in its conditional steady state $s = 2$ (the intermediate aggregate state). At date 0 there is a recessionary shock to low productivity state $s = 1$ which lasts for 15 months [the expected duration of the low state], after which aggregate productivity permanently reverts to $s = 2$. Of course agent expectations are always consistent with the full model and so Figure 9(b) describes the economy response to a particular sequence of realised

productivities a_t which we refer to as a "typical recession".[16]

Figure 9: JC, hires, quits in the data and model



(a) Data: HP-filtered (b) Model: typical recession

Left panel plots Davis, Faberman and Haltiwanger (2012) data on the job creation, hire, and quit rates in the data, which have been logged and HP-filtered with parameter 10^5. Right panel plots the path for these variables in the model, during a "typical recession" experiment (see text for details), expressed as deviations from their initial values.

Figure 9(a) demonstrates that job creation, hire and quit rates are highly positively autocorrelated where quits vary the most and job creation varies the least. Table 2 also reveals that quit rates are more highly correlated with unemployment U. To understand why our model reproduces these features, note first estimated employed worker search intensity is $\widehat{\phi} = 0.1$ and so any increase in unemployment implies an increase in aggregate search intensity. This has important crowding out effects: for example in the "typical recession" 51% of hires are from unemployment in the initial conditional steady state which rises to 60% at the trough of the recession. Furthermore job finding rates collapse as the number of jobs created per unemployed worker, JC/U, collapses in the recession [see Figure 3 in the Introduction]. Thus like unemployed worker job finding rates, a "typical recession" finds quit rates also fall steeply [and the job ladder collapses].

The "typical recession" shows that job creation rates are the first to start recovering (before hires and quits and so is the least persistent process). This behaviour is clearly reflected in the data: for example Figure 9(a) shows the 2008 recession led to a steep fall in all three series but job creation was the first to recover. The "typical recession" finds

[16] In the Online Appendix we provide the impulse responses of the main aggregate variables, describing their behaviour under a typical recession.

in the subsequent recovery, i.e. once the economy returns to $s = 2$ at 15 months, that job creation rates increase by a discrete amount, as do quit rates (because higher job creation makes finding work easier). Nevertheless quit rates continue to be suppressed while unemployment remains above trend. And because unemployment is the most volatile and highly persistent process, job search crowding out implies quit rates are more volatile and more persistent than job creation rates.

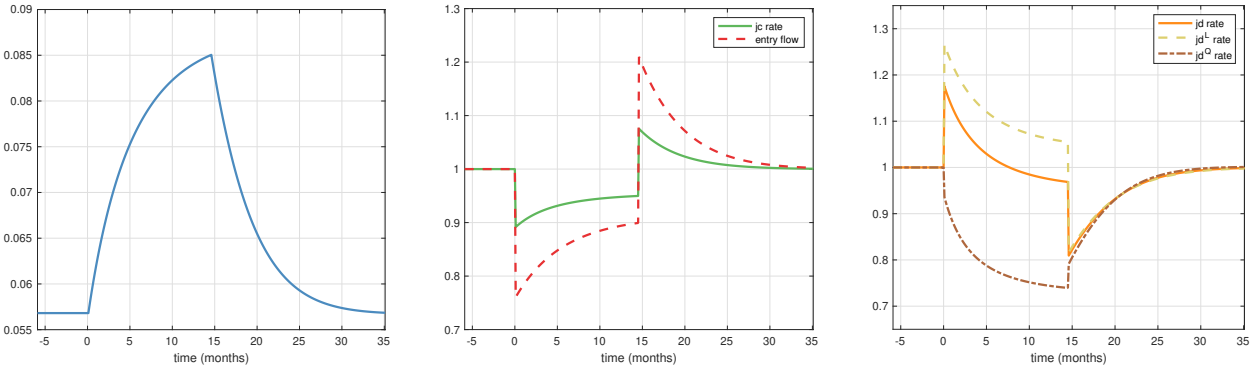
An interesting insight due to Figure 9(a) is that when quit rates are above trend [i.e. when $q > 0$] then [detrended] gross hire rates typically exceed job creation rates, and vice versa. This reflects the underlying replacement hiring process: that replacement hiring increases as quits increase, and gross hires equal job creation plus replacement hiring. The large quit dynamic described above thus explains why gross hires are more volatile and more persistent than job creation. And, of course, replacement hiring causes vacancy chains, where if a new job created is filled by an already employed worker then a "new job" continues to exist should the previous employer choose to hire a replacement worker. As described in Elsby et al. (2021), this vacancy chain effect magnifies the gap between gross hires and job creation.

5.2 Job creation and job destruction

Job creation and destruction are the fundamental drivers of unemployment over the cycle. Figure 1 in the Introduction describes how actual *jc* and *jd* evolve in the US economy and in the Online Appendix we present its model equivalent. Instead Figure 10 plots the behaviours of *jc* and *jd* over a "typical recession". Job destruction is additionally decomposed into jd^L, jobs destroyed where the worker is laid off into unemployment (mostly) as a result of a δ_D shock, and jd^Q, jobs destroyed where the firm instead declines to replace a worker who quits.

Consider first the *jc* process in a "typical recession": panel (b) plots both the job creation rate and also the firm entry flow [the number of start up firms], as deviations from their initial values. In line with the data, the typical recession finds firm entry is more volatile than gross job creation but firm entry only has a modest effect on total job creation because new start-ups begin small [though more gazelles is clearly good for future growth]. The efficiency wage structure implies firm value depends directly on worker

Figure 10: Impulse response to typical recession: Unemployment and quit dynamics



(a) Unemployment (b) Job Creation (c) Job Destruction

Panels plot the paths for variables during a "typical recession" experiment (see text for details). Job creation and destruction are given as rates to the employment stock, which are expressed as deviations from their pre-recession values. jd^L and jd^Q refer to job destruction from layoffs and unreplaced quits respectively.

quit rates and the previous section has shown quit rates fall directly with unemployment. Equilibrium unemployment is thus stable. However to generate large variations in unemployment consistent with the data, the feedback of higher unemployment into greater net job creation has to be relatively weak. The framework thus estimates the elasticity of job creation rates to firm value. This is important because the free entry approach assumes a penny increase in value yields an infinite number of entrants and this creation process is far too elastic for the data. The estimated elasticities of firm entry ($\xi_e = 4.3727$) and job creation ($\xi_{JC} = 2.8495$) to firm value are more than one, and so relatively elastic (with start-up entry being the more elastic), but both are a long way from infinity (the free entry case).

The endogenous job destruction process is more complicated for it has two separate channels. Similar to Mortensen and Pissarides (1994), a small (unfavourable) aggregate productivity shock has a large, immediate impact on layoffs precisely because the lowest productivity firms have small surplus. The onset of the typical recession causes a [30%] spike the job destruction rate through increased layoffs. Although this spike in layoffs has an immediate and large impact on unemployment, the cleansing effect of recessions implies this spike is relatively short-lived. In contrast the job destruction channel via unreplaced quits is more persistent precisely because quits are a highly persistent process. The fall in quit rates in the recession results in lower job destruction via unreplaced quits

which partly compensates for the spike in layoffs.

6 Conclusion

This paper has developed a new fully microfounded equilibrium business cycle model of the US labor market which is consistent both with the underlying distribution of firm growth rates across firms [by age and size] and macro-evidence regarding gross job creation and job destruction flows over the cycle. The framework not only successfully generates the (targeted) average firm size distribution by age but also the (untargeted) distributions of firms and of employment by firm size. The approach also provides an important new insight - that net job creation is uncorrelated with unemployment. We have shown that it is this property of the data which is central to explaining the business cycle frequencies of unemployment.

The approach has used aggregate productivity shocks, rather than discounting shocks, as the driver of the economy. It would seem unlikely that changing to discounting shocks would much affect our insights. For example any negative aggregate shock will always lead to a spike in layoffs because a key component of the job destruction process is low productivity firms have small surplus. Furthermore as described above, the efficiency wage distortion will always imply it is the quit process, rather than wages, which move the most over the cycle. That is, high unemployment always causes a steep fall in quit rates, the consequent collapse of the job ladder and a slow recovery because net job creation rates respond (at most) weakly to high unemployment.

References

[1] Audoly, R. 2020. "Firm Dynamics and Random Search over the Business Cycle," *mimeo*.

[2] Bertolla, G. and R. Caballero. 1994. "Cross-Sectional Efficiency and Labour Hoarding in a Matching Model of Unemployment," *Review of Economic Studies*, 61: 435-456.

[3] Bilal, A, N. Engbom, S. Mongey and G. Violante. 2021. "Firm and worker dynamics in a frictional labor market," *Econometrica*, Forthcoming.

[4] Brown, C. and J. Medoff. 1989. "The Employer Size-wage Effect," *Journal of Political Economy*, 97(5): 1027-1059.

[5] Burdett, K. and D. Mortensen. 1998. "Wage Differentials, Employer size and Unemployment," *International Economic Review*, 39: 257-273.

[6] Coles, M. G. and E. Smith. 1998. "Marketplaces and Matching," *International Economic Review*, 39(1): 239-254.

[7] Coles, M. G. and A. Moghaddasi. 2018. "Do Job Destruction Shocks Matter in the Theory of Unemployment?," *American Economic Journal: Macroeconomics*, 10(3): 118-136.

[8] Coles, M. G. and D. Mortensen. 2016. "Equilibrium Labor Turnover, Firm Growth, and Unemployment," *Econometrica*, 84: 347-363.

[9] Coles, M. G. 2001. "Equilibrium Wage Dispersion, Firm Size and Growth," *Review of Economic Dynamics*, 4(1): 159-187.

[10] Cooper, R and J. Haltiwanger. 2006. "On the Nature of Capital Adjustment Costs," *Review of Economic Studies*, 73(3): 611-633.

[11] Davis, S. J., R. J. Faberman and J. Haltiwanger. 2012. "Labor market flows in the cross section and over time," *Journal of Monetary Economics*, 59(1): 1-18.

[12] Davis, S. J., R. J. Faberman and J. Haltiwanger. 2013. "The Establishment-Level Behavior of Vacancies and Hiring," *The Quarterly Journal of Economics*, 128(2): 581-622.

[13] Davis S. J. and J. Haltiwanger. 1992. "Gross Job Creation, Gross Job Destruction, and Employment Reallocation," *The Quarterly Journal of Economics*, 107(3): 819-863.

[14] Diamond, P. A. 1982. "Aggregate Demand Management in Search Equilibrium," *Journal of Political Economy*, 90(5): 881-894.

[15] Elsby, M. and A. Gottfries. 2021. "Firm Dynamics, On-the-Job Search, and Labor Market Fluctuations", *Review of Economic studies*. Forthcoming.

[16] Elsby, M., A. Gottfries, R. Michaels and D. Ratner. 2021. "Vacancy Chains". University of Edinburgh, Department of Economics, mimeo.

[17] Faberman, J. and E. Nagypal. 2008. "Quits, Worker Recruitment, and Firm Growth: Theory and Evidence", Working Paper No. 08-13, Federal Reserve Bank of Philadelphia, USA.

[18] Haltiwanger, J. C., H. R. Hyatt, L. B. Kahn and E. McEntarfer. 2018. "Cyclical Job Ladders by Firm Size and Firm Wage," *American Economic Journal: Macroeconomics*, 10(2): 52-85.

[19] Haltiwanger, J. C., R. S. Jarmin, R. Kulick and J. Miranda. 2017. "High Growth Young Firms: Contribution to Job, Output, and Productivity Growth,". In J. Haltiwanger, E. Hurst, J. Miranda, and A. Schoar, (Eds). *Measuring Entrepreneurial Businesses: Current Knowledge and Challenges*, University of Chicago Press, Chicago, USA.

[20] Hosios, A. 1991. "On the Efficiency of Matching and Related Models of Search and Unemployment". *Review of Economic Studies*, 57 (2): 279-298.

[21] Klette, T. J. and S. Kortum. 2004. "Innovating Firms and Aggregate Innovation," *Journal of Political Economy*, 112(5): 986-1018.

[22] Mercan, Y. and B. Schoefer. 2020. "Jobs and Matches: Quits, Replacement Hiring, and Vacancy Chains", *American Economic Review: Insights*, 2(1): 101-124.

[23] Mortensen, D. and C. Pissarides. 1994. "Job Creation and Job Destruction in the Theory of Unemployment". *Review of Economic Studies*, 61(3): 397-415.

[24] Moscarini, G. and F. Postel-Vinay. 2013. "Stochastic Search Equilibrium," *Review of Economic Studies*, 80(4): 1545-1581.

[25] Lee, Y. and T. Mukoyama. 2015. "Entry and Exit of Manufacturing Plants over the Business Cycle," *European Economic Review*, 77: 20-27.

[26] Lentz, R. and D. T. Mortensen. 2008. "An Empirical Model of Growth Through Product Innovation," *Econometrica* 76(6): 1317-1373.

[27] Pissarides, C. 2000. "Equilibrium Unemployment," MIT press, Cambridge, USA.

[28] Postel-Vinay, F. and J-M. Robin. 2002. "Equilibrium Wage Dispersion with Worker and Employer Heterogeneity," *Econometrica* 70(6): 2296-2350.

[29] Schaal, E. 2017. "Uncertainty and Unemployment," *Econometrica*, 85(6): 1675-1721.

[30] Sedlacek, P. and V. Sterk. 2017. "The Growth Potential of Startups over the Business Cycle," *American Economic Review*, 107(10): 3182-3210.

[31] Shapiro, C. and J. E. Stiglitz. 1984. "Equilibrium Unemployment as a Worker Discipline Device", *American Economic Review*, 74(3): 433-444.

[32] Shimer, R. 2005. "The Cyclical Behavior of Equilibrium Unemployment and Vacancies," *American Economic Review*, 95(1): 25-49.

ONLINE APPENDIX

A Proofs

Proof of Lemma 2: For convenience we suppress reference to Ω. It is immediate that $q(i) = \lambda_u + \lambda_1$ for $i \leq i^h$. Consider now $i > i^h$ where (13) implies equilibrium quit rate

$$q(i) = \lambda_u + \frac{\lambda_1}{\lambda} \int_i^1 \frac{h(j)[1-U]G'(j)}{\alpha + (1-\alpha)G(j)} dj.$$

Now $\lambda_1 = \phi\lambda_0$ and $\lambda = \lambda_0 U + \lambda_1(1-U)$ implies $\lambda_1/\lambda = \phi/[U + \phi(1-U)]$ while $\alpha \equiv \lambda_0 U/\lambda = U/[U + \phi(1-U)]$. Substituting out λ_1/λ and α in the above yields

$$q(i) = \lambda_u + \int_i^1 \frac{h(j)G'(j)}{\frac{U}{\phi(1-U)} + G(j)} dj.$$

Because $h(j) = JC(j) + q(j)$ for $j \geq i > i^h$ we also have

$$h'(j) = JC'(j) - \frac{h(j)G'(j)}{\frac{U}{\phi(1-U)} + G(j)}.$$

Now define $Z(j) = \frac{U}{\phi(1-U)} + G(j)$ and so $Z'(j) = G'(j)$. Integration by parts establishes:

$$\int_i^1 Z'(j)h(j)dj = [Z(j)h(j)]_i^1 - \int_i^1 Z(j)\left[JC'(j) - \frac{h(j)G'(j)}{\frac{U}{\phi(1-U)} + G(j)}\right] dj$$

and simplifying yields:

$$Z(1)h(1) - Z(i)h(i) = \int_i^1 Z(j)JC'(j)dj.$$

Integrating by parts then yields:

$$Z(1)h(1) - Z(i)h(i) = Z(1)JC(1) - Z(i)JC(i) - \int_i^1 Z'(j)JC(j)dj.$$

Substituting out $h(1) = \lambda_u + JC(1)$, $h(i) = JC(i) + q(i)$ implies:

$$\begin{aligned} Z(i)q(i) &= Z(1)\lambda_u + \int_i^1 Z'(j)JC(j)dj \\ &= Z(i)\lambda_u + \int_i^1 Z'(j)[\lambda_u + JC(j)]dj. \end{aligned}$$

Using $Z(i) = \frac{U}{\phi(1-U)} + G(i)$ and $Z'(j) = G'(j)$ then establishes the Lemma.

B Quantitative model appendix

B.1 Finite productivity model summary

In this section we briefly summarise the equations of the finite productivity model, which is used in our quantitative work. We also explain the minor additions to the model made relative to the continuous productivity model in the text. Our calibrated model features $i^c(\Omega) = 1$ at all times, and we present the equations for this case of the model. Recall that we specialise to a finite number of productivities $i = 1, ..., I$, where within each productivity level firms additionally separate into different wage ranks $\chi \in [0, 1]$. We then define the overall wage rank across all firms as $x \in [0, 1]$, as specified in the text.

Firm HJB and policy functions: All firms with the same productivity p_i achieve the same value v_i, regardless of their wage rank. With aggregate shocks the HJB includes the aggregate state $\Omega = (s, N_1, ..., N_I)$. If $i^c(\Omega) = 1$ at all times, the N_i evolve continuously over time. In this case, the HJB can be written

$$\begin{aligned}(r+\delta_F)v_i(\Omega) = a_s p_i - c_f - w_{\min} - (\lambda_1(\Omega)+\lambda_u)\min[v_i(\Omega), c_0] + \mu_1 E_c \max[v_i(\Omega) - [c_0 + c], 0] \\ - \delta_D E_c \min[v_i(\Omega), c] + \alpha_\gamma \sum_j \gamma_{ij}(v_j(\Omega) - v_i(\Omega)) \\ + \alpha_a \sum_{s'} \gamma_{s,s'}(v_i(s', \underline{N}) - v_i(\Omega)) + \sum_{j=1}^{I} \frac{\partial v_i(\Omega)}{\partial N_j}\dot{N}_j(\Omega). \quad (22)\end{aligned}$$

Notice that we extend the model relative to the main text by introducing a flow cost of capital maintenance, c_f. This is a cost which must be paid each period to maintain each existing unit of capital. The introduction of c_f does not change the economics of the model, but it is useful as it allows us to more easily partition the productivity bins into those which do and do not replace quits (see Section B.4 for more details). The only aggregate "price" which affects firm value is the scalar $\lambda_1(\Omega)$. The expectations over JC and JD cost draws have closed form solutions under the assumed distributions. The hiring threshold $i^h(\Omega)$ is defined as the first i for which $v_i(\Omega) > c_0$.

In the finite productivity model, most policy functions – and in particular those which relate to net employment dynamics – depend only on i and not the wage rank. Specifically, the job creation rate per employee is $jc_i(\Omega) = \mu_1 H^{JC}(v_i(\Omega) - c_0)$. The job de-

struction rate is $jd_i(\Omega) = \delta_D[1 - H^{JD}(v_i(\Omega))]$ for firms with $i > i^h(\Omega)$ and $jd_i(\Omega) = \delta_D[1 - H^{JD}(v_i(\Omega))] + \lambda_1(\Omega) + \lambda_u$ otherwise. Entrants who draw productivity i have average initial employment $\bar{n}_{0,i}(\Omega) = n_u + \frac{N_0}{\mu_1} jc_i(\Omega)$. Here, n_u is the number of initial workers a firm can draw from unemployment for free upon startup. We define $\hat{n}_{0,i}(\Omega) = \frac{N_0}{\mu_1} jc_i(\Omega)$ as average entrant size excluding the initial n_u free hires from unemployment.

Evolution of employment distribution: The total mass of employment at each productivity bin evolves according to:

$$\dot{N}_i(\Omega) = \mu_0 P^E(\Omega)\gamma_{0i}\bar{n}_{0,i}(\Omega) + N_i \left[jc_i(\Omega) - jd_i(\Omega) - \delta_F - \alpha_\gamma \sum_{j \neq i} \gamma_{ij} \right] + \alpha_\gamma \sum_{j \neq i} \gamma_{ji} N_j. \quad (23)$$

The first term on the right hand side is the inflow of employment from firm entry. The term in square brackets gives net job creation accounting for job creation and destruction including the firm exit shock. The terms proceeded by α_γ gives the transition of firms across productivity bins. Total unemployment is $U = 1 - \sum_i N_i$. The distribution across firm wage ranks can then be calculated from our closed form solution:

$$G(x, \Omega) = \frac{\sum_{j=1}^{i-1} N_j + \frac{x - x_{i-1}}{x_i - x_{i-1}} N_i}{1 - U} \text{ for all } x \in [x_{i-1}, x_i). \quad (24)$$

Recall that we define the boundaries $x_0 = 0$, $x_i = x_{i-1} + \gamma_{0i}$ and $x_I = 1$. A firm in state $i \in \{1, 2, ..., I\}$ with wage rank $\chi \in [0, 1)$ is correspondingly defined as being in state $x \in [0, 1]$ where $x = x_{i-1} + \chi[x_i - x_{i-1}]$.

Quits and hires across the x distribution: To close the model, we need to calculate the offer arrival rate $\lambda_1(\Omega)$. To do this, we must solve the quit rates across the wage rank distribution. As in the continuous productivity model, the quit rate for incumbent firms at any wage rank x is given by (20), which we rewrite in our x notation as

$$q(x, \Omega) = \begin{cases} \lambda_u + \lambda_1(\Omega) & \text{if } x \in [0, x^h(\Omega)) \\ \lambda_u + \frac{\phi \int_x^1 \{JC(y,.) + \lambda_u [1-U] G'(y)\} dy}{U + \phi[1-U]G(y)} & \text{if } x \geq x^h(\Omega) \end{cases} \quad (25)$$

where $x^h(\Omega) = x_{i^h(\Omega)-1}$ corresponds to the lowest ranked hiring firm, who has $i = i^h(\Omega)$ and $\chi = 0$. The total job creation flow at each x is $JC(x, \Omega) = \left\{ [1 - U] G'(x) + \frac{\mu_0}{\mu_1} N_0 P^E(\Omega) \right\} jc_i(\Omega)$. The closed form solution for $G(x)$ similarly defines a closed form solution for $G'(x)$,

which is well defined except at the x_i boundaries where $G(x)$ is non-differentiable. Performing the integration in (25) yields a closed form solution for $q(x,\Omega)$ for any $x \geq x^h(\Omega)$:

$$q(x,\Omega) = \lambda_u + \frac{(1-\chi)\left[(jc_i + \lambda_u)N_i + \mu_0 P^E(\Omega)\gamma_{0,i}\hat{n}_{0,i}\right] + \sum_{j=i+1}^{I}\left[(jc_j + \lambda_u)N_j + \mu_0 P^E(\Omega)\gamma_{0,j}\hat{n}_{0,j}\right]}{U/\phi + \sum_{j=1}^{i-1} N_j + \chi N_i} \tag{26}$$

This equation uses the reverse mapping $\chi(x)$ to find the χ associated with the current x from $\chi = \frac{x - x_{i-1}}{\gamma_{0i}}$, and similarly the productivity level $i(x)$ associated with the current x. The hiring rate for incumbent firms can then be simply computed as $h(x,\Omega) = jc_i(\Omega) + \mathbf{1}(i \geq i^h(\Omega))q(x,\Omega)$. Finally, $\lambda_1(\Omega)$ is just $q(x,\Omega)$ evaluated at $x = x^h(\Omega)$ and then subtracting λ_u, which gives

$$\lambda_1(\Omega) = \frac{\sum_{j=i^h(\Omega)}^{I}\left[(jc_j(\Omega) + \lambda_u)N_j + \mu_0 P^E(\Omega)\gamma_{0,j}\hat{n}_{0,j}(\Omega)\right]}{U/\phi + \sum_{j=1}^{i^h(\Omega)-1} N_j} \tag{27}$$

This closes the model, and provides sufficient information to simulate the model keeping track only of the finite employment vector $\underline{N}$. In particular, the model can be solved and simulated using only the HJB (22), the evolution of total employment by productivity bin (23), and the closed-form solution for the job offer arrival rate (27).

B.2 Data sources and treatment

We use the following data throughout the paper. Our main sample period used for estimation purposes is the one for which we can get data on all variables simultaneously, which is 1990Q2 to 2018Q4.

Business Dynamics Statistics (BDS): We use the BDS to construct data stratified by firm age. We use the 2018 release, which is available yearly from 1978 to 2018, and take the national data, split by firm age. We use data on the number of firms (and total employment in firms) of each age bin to calibrate our model. We also measure firm entry using this dataset, as the number of firms aged 0 in each year, and firm exit, as given in the dataset. We also use the data on Job Creation and Destruction by age for Figure 1 in the Introduction of the paper, but instead calibrate our model to the quarterly JC and JD data from Davis, et al. (2012).

Davis, Faberman, and Haltiwanger (2012, DFH): We extensively use the data provided by Davis et al. (2012) which underlies the analysis in their paper. We are grateful to

the authors for sharing the (updated) data which underlies the plots in their paper with us. This data consists of quarterly data from 1990Q2 to 2018Q4. We use their estimates of aggregate job creation and job destruction, as well as layoffs and quits. We add their measure of "other separations" to layoffs. Their data are given as rates of total employment. We calibrate our model to match the average of these data over our sample, as well as the HP-filtered time series. As well as this, we use data from this paper to form an estimate of the fraction of worker quits that firms replace by undertaking a replacement hire. We discuss this further below.

Bureau of Labor Statistics (BLS): We use monthly aggregate data from the Current Population Survey. We aggregate these data up to a quarterly frequency by taking the simple average. We use data on total employment (CE16OV) and unemployment (UNEMPLOY), in levels and seasonally adjusted. We additionally use data on the total number of people unemployed for less than five weeks (UEMPLT5) to construct the unemployed worker UE rate, following the approach of Shimer (2005).

Bureau of Economic Analysis (BEA): To construct our measure of labor productivity (output per worker) we use data on quarterly real GDP (GDPC1) from the BEA. Labor productivity is calculated as real GDP divided by total employment from the BLS data.

Compustat: Since our model assumes constant returns to scale, we do not allow for permanent productivity differences across firms, as these would lead to permanent differences in employment growth rates (rather than levels, as would be true in a model with decreasing returns to scale). Therefore, we calibrate our firm-level productivity process to the within-firm standard deviation of productivity shocks, rather than the across firm standard deviation. To compute this measure, we use data from Compustat. We use data on all US based firms in their sample, and use data only on sales and total employment. We deflate sales using the GDP deflator (GDPDEF, from the BEA) to create a measure of real sales, and then define firm-level labor productivity each year as real sales over employment. We drop firm-year observations with missing or negative sales or employment, and winsorize the data by dropping the top and bottom 1% of data by both yearly sales growth and employment growth. We take the log of labor productivity, and regress it on firm and year fixed effects, and take the residual as our measure of firm-level productivity, corrected for firm-level averages and aggregate changes. We take the standard deviation of this measure, which yields a value of 28.38%, computed from 291,703

firm-year observations.

Estimating the fraction of quits which are replaced: To estimate the amount of replacement hiring in the model, we draw information both from gross flows and from underlying firm-level data. Firstly, we note that the amount of replacement hiring is not simple to observe from aggregate flows, due to the fact that firms may do replacement hiring for two reasons: either to replace workers who quit, or to replace workers they lay off for being a bad match but where the firm wants to keep the job open. Through the lens of the model, the total hiring rate is equal to

$$h_t = jc_t + qfr_t \left(q_t + \lambda^u\right) \tag{28}$$

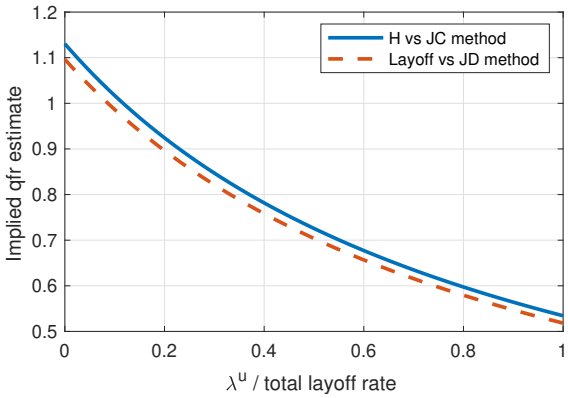
where we define qfr_t as the fraction of worker quits (and layoffs due to bad worker match) which are replaced. Recall that λ^u is the rate at which workers are fired for being a bad match, but where the firm's capital remains intact so the firm has the option to hire to replace them. qfr serves as our calibration target for the amount of replacement hiring in the model. Notice that three objects in this equation are observable in the DFH dataset: hiring (h_t), job creation (jc_t), and quits (q_t). If we assumed that all layoffs were due to job destruction (and firms never fired workers with the aim of replacing them with another worker) then $\lambda^u = 0$, and estimating the degree of replacement hiring would be simple using this aggregate data alone. In this case, simply rearrange (28) to yield $qfr_t = (h_t - jc_t)/q_t$.

However, the data in DFH suggest that firms do indeed replace some of their workers who leave due to layoff, so this approach is likely not valid. In particular, in their well-known "hockeystick" plot (their Figure 7(b)) we observe that firms who have positive employment growth, and hence are expanding, still extensively use worker layoffs. In fact, our calculations below suggest that the average layoff rate for non-contracting firms appears to be around 2.73% per quarter. Given that these firms are expanding their employment on net, it is likely that they are replacing some of the workers who they have laid off, meaning that $\lambda^u > 0$. Indeed, through the lens of our model, firms which perform job creation necessarily replace worker quits.

Given that $\lambda^u > 0$ therefore seems like a more reasonable assumption, we can then return to (28) to understand the impact this has on estimated quit replacement. Express-

ing the relationship in steady state gives $h = jc + qfr\,(q + \lambda^u)$, where we take h, jc, and q directly from the average values in DFH's data. This gives the implied value of qfr for any assumed λ^u as $qfr = (h - jc)/(q + \lambda^u)$. Without any further information, λ^u is constrained to lie within the range 0 (in which case all layoffs are due to job destruction) and the total layoff rate in the data, l (in which case all layoffs are replaced, and not due to job destruction) [see equation (30) below]. We plot the implied value of qfr in this range in the blue line in Figure B.1 below.

Figure B.1: Estimated fraction of quits replaced vs. assumed λ^u



This procedure bounds the fraction of quits and layoffs which are replaced to be between around 50%, if all layoffs are assumed to be replaced, and 100%, if only 10% of layoffs are assumed to be replaced. Notice that for values below 10% the data implies that more than 100% of quits are replaced. Before going further, we therefore note that our chosen value for the estimation, $qfr = 80\%$, happens to lie approximately in the middle of the upper and lower bounds implied by the aggregate flow data.

One could potentially use other aggregate flow relationships, such as those between job destruction and layoffs, might help estimate the fraction of layoffs which are replaced, and hence pin down qfr. However, we found this challenging as the aggregate relationships are by definition collinear, and more information is needed. To see this, consider that job destruction is given by

$$jd_t = jd_t^l + jd_t^q = jd_t^l + (1 - qfr_t)(q_t + \lambda^u) \tag{29}$$

and layoffs by

$$l_t = jd_t^l + \lambda^u \tag{30}$$

where l_t is layoffs in the data, jd_t^l is job destruction shocks which induce layoffs, and jd_t^q is job destruction due to unreplaced quits and layoffs. Taking jc, jd, q, and l as data, (28), (29), and (30) appear to provide three equations which can solve for the three unknowns qfr_t, λ^u, and jd_t^l. However, the equations actually contain the same economic content and are collinear. Combining (29) and (30) to yield

$$qfr_t(q_t + \lambda^u) = l_t - jd_t + q_t \tag{31}$$

and rearrange (28) to yield

$$qfr_t(q_t + \lambda^u) = h_t - jc_t. \tag{32}$$

As the left hand sides of these equations are identical, the three equations together cannot be solved for a unique solution for qfr_t, λ^u, and jd_t^l. Instead, combining these two equations implies an adding-up condition which should hold in the data in theory: $jc_t - jd_t = h_t - l_t - q_t$. In practice, the adding up condition is very slightly violated, meaning that (31) and (32) provide very slightly different estimates of qfr_t for a given assumed λ^u. The estimate from (31) is given in Figure B.1 as the dashed red line, which is very similar the the previous estimate.

The discussion above shows that additional data must be included to estimate the fraction of quits which are replaced, and we investigate two approaches.

As a first approach, note that with knowledge of λ^u, the value of qfr can be calculated using the accounting relationship above. λ^u is the rate at which workers are fired for being a bad match, with the firm having the option to replace them if desired. Through the lens of the model, this can be identified as the layoff rate at expanding firms, who perform no job destruction and so any layoffs must be due to the λ^u shock. To estimate this in the data, we use the hockeystick and growth rate distribution plots in DFH.[17] We have access only to the growth rate distributions from 2006 and 2008-9 (as plotted in DFH) and so use data for the distribution and hockeysticks from 2006 to construct our estimate. Accordingly, this is data from a single year, which corresponds to an estimate for a typical non-recession year.[18] For each growth rate bin $i = -200, 199, ..., 200$, we have data on the

[17]The authors very kindly provided us with the data behind these plots, which consists of the hires, layoff, total separation, and quit rate at each growth rate bin (their Figures 6 and 8), and the (employment weighted) kernal density function of firms in each bin (their Figure 5). The data are provided on slightly different grids for each plot, and we interpolate the data onto an integer grid from -200 to 200.

[18]The results are robust to using the hockeysticks from all years (their Figure 6) integrated using the

mass of employment at establishments with that growth rate (d_i) and the construct a layoff rate at that bin (l_i). We calculate the average layoff rate in all bins with non-declining employment growth as $\sum_{i=0}^{200} l_i d_i / \sum_{i=0}^{200} d_i = 2.73\%$. Under the identifying assumption that λ^u is constant across firms (as it is in the model), this implies an estimate $\lambda^u = 2.73\%$, which is 39% of the average layoff rate of 7.0% in the DFH data in 2006. Referring back to Figure B.1, we see that 39% of layoffs being potentially replaceable implies a value of qfr of approximately 80%, which is the value used in our calibration.[19] As an alternative, we also directly calculated the fraction of quits replaced from the 2006 hockeystick data for quits and hires, and found that 79.6% of quits were replaced. Specifically, we assume that expanding firm bins replace all quits and layoffs ($qfr_i = 1$). For contracting bins, we calculated the fraction of quits replaced as $qfr_i = h_i/(q_i + \lambda^u)$. Taking the $(q_i + \lambda^u)$-weighted average of qfr_i across the whole distribution yields 79.6%.

As a second approach, we consider the notion of replacement hiring in Elsby et al. (2021). They define a broad notion of replacement hiring using JOLTS data as follows. For each establishment, they consider replacement hires as the minimum of gross hires and quits in a given quarter. They then sum across establishments, and find that, by this definition, around 45% of all hires are replacement hires. Doing the same exercise on simulated data from our model finds that 40% of hires are replacement hires. As mentioned in the text, our model also generates that around 50% of firms have zero net employment change over 3 months, and since these firms also lose workers to quits, this serves as another measure of replacement hiring. Elsby et al. (2021) find this number to be around 55% and 65% in the QCEW and JOLTS data respectively. Finally, their strictest measure of replacement hiring is the total hiring at firms with zero net employment change as a fraction of total hiring. This number is 7.5% in their data, and 7.1% in our model. By all these measures our model generates a substantial amount of replacement hiring, close to the measures in the data. This provides an alternative justification for our calibrated value of $qfr = 0.8$, which delivers a sensible amount of replacement hiring by these alternative measures, and suggests that our results would be robust to instead using these measures as targets in our estimation.

average of the 2006 and 2008-9 growth rate distributions to roughly attempt to form an estimate for all years. However, since the match of hockeysticks and growth rate distribution sample is not exact in this case, we prefer to use the data from 2006 only.

[19] If roughly 40% of layoffs are replaceable and the replacement rate is 80%, this implies that 32% of layoffs are actually replaced.

B.3 Numerical methods

Steady state: For given parameter values, solving the core equations of the model in steady state reduces to solving for a vector of I values v_i and employment stocks N_i, as well as the arrival rate λ_1. This is a simple problem to solve using the steady state versions of (22), (23), and (27). Intuitively, one can guess a value of λ_1, solve the HJB for the values v_i, use the implied policy functions to calculate the employment stocks N_i, and use these to update your guess for λ_1. In practice, we solve the model in steady state at the same time as calibrating our parameters, which involves calculating other statistics, which we detail further below.

Calculating average quit and hiring rates involves integrating over the wage rank distribution, x. To do this, we build an uniform grid over $\chi \in [0, 1]$ with 1,000 nodes. This is then combined with the x_i to build a grid over x with $I \times 1,000$ nodes. We calculate all integrals on this grid using trapezoid integration.

To calculate the firm age and size distribution we solve for the densities of firms on grids for age and size. This is thus reminiscent of the non-stochastic simulation approach of Young (2010), or the methods of Achdou et al. (2021). To calculate the size distribution, we build a grid over firm sizes. Recalling that the number of employees in a given firm is an integer, we define a size grid as integer values from 0 to 20,000 employees. We solve for the mass of firms at each size s and productivity i. We use the firm dynamics processes (job creation, destruction, entry, exit, and so on) to construct flow rates across these joint size-productivity bins, which we use to build a matrix of transitions. We can then solve for the steady-state density of the number of firms at each size-productivity bin by inverting this matrix.

To calculate the age distribution we follow a similar process. However, since age is a continuous number, we discretise the age grid on a uniform grid from age 0 to age 26 years old (since this is the maximum age bin recorded in the BDS data) with 1,000 nodes. We solve for the mass of firms and employment at each age-productivity bin. Finally, to compute the mass of active firms at each age, we actually need to compute the joint age-size distribution, since we define firms with 0 employees as having exited. To do this, we must solve for the joint age-size-productivity distribution, which we do using the same methods. Given the high dimension of this object, we solve for this distribution using a reduced firm size grid from 0 to 1,000 employees, and confirm that raising this maximum

has no impact on the moments for which this distribution is used.

To compute the growth rate density and hockeystick plots (Figure 7 in the main text) we simulate a panel of firms for one quarter, with their initial states drawn from the steady-state size-productivity distribution. We simulate a panel of one million firms and calculate net employment growth and gross flows from the beginning to the end of the quarter.

Business cycle: For given parameter values, solving the core equations of the model over the business cycle reduces to solving for the functions $v_i(\Omega)$, $\dot{N}_i(\Omega)$, and $\lambda_1(\Omega)$ over the state space $\Omega = (s, N_1, ..., N_I)$, using the equations (22), (23), and (27). In terms of approximation, it actually suffices to approximate only the function $v_i(\Omega)$, as the values of $\dot{N}_i(\Omega)$, and $\lambda_1(\Omega)$ can then be calculated exactly using (23) and (27) at any grid point or point in a simulation.

We approximate $v_i(\Omega)$ using second order polynomials in $N_1, ..., N_I$, with different coefficients for each of the discrete productivity level pairs i, s. In particular, first adjust the value function notation to $v_i(\Omega) = v_{i,s}(N_1, ..., N_I)$ to acknowledge that aggregate productivity s is also a discrete state. Then for each i, s we approximate $v_{i,s}(N_1, ..., N_I)$ as

$$v_{i,s}(N_1, ..., N_I) \simeq h^0_{i,s} + \sum_{j=1}^{I} \left(h^1_{i,s,j} N_j + h^2_{i,s,j} N_j^2 \right) \tag{33}$$

where $h^0_{i,s}$ $h^1_{i,s,j}$ and $h^2_{i,s,j}$ are scalar coefficients to be estimated. $h^0_{i,s}$ is the intercept, and $h^1_{i,s,j}$ and $h^2_{i,s,j}$ capture the first and second order effect on the value of firms with state i of changing total employment of firms with productivity j . Notice that we exclude cross terms in the second-order approximation, since these are known to typically be unstable given that the $N_1, ..., N_I$ tend to be highly correlated. For each i, s this approximation uses $1 + 2 \times I = 1 + 2 \times 5 = 11$ coefficients. Since we use $I = 5$ idiosyncratic and $S = 3$ aggregate productivity nodes, this gives $5 \times 3 \times 11 = 165$ coefficients to estimate.

To solve the HJB, we need to know both the level of value and its derivative with respect to the aggregate employment bins. The derivatives are easy to compute given our approximation, as

$$\frac{\partial v_{i,s}(N_1, ..., N_I)}{\partial N_j} = h^1_{i,s,j} + 2 h^2_{i,s,j} N_j \tag{34}$$

In order to solve the full business-cycle version of the model, we use the following proce-

dure:

1. We need a grid of values for $(N_1, ..., N_I)$ to approximate our second-order polynomial on. To generate this, we use a Sobol set, which generates values of the $(N_1, ..., N_I)$ vector which are roughly equally spaced between a minimum and maximum value for each N_i. We generate 50 of such vectors, denoting the values of $(N_1, ..., N_I)$ at each candidate z as $\underline{N}_z = (N_{1,z}, ..., N_{I,z})$ for $z = 1, ..., 50$. Note that the aggregate state at any grid point is now denoted s, z, where s corresponds to aggregate productivity and z to the vector of $\bar{N}$ values.

2. Generate initial guesses for the parameters $h^0_{i,s}$ $h^1_{i,s,j}$ and $h^2_{i,s,j}$. Generate an initial guess for $\lambda_1(\Omega) = \lambda_{1,s,z}$ at each aggregate state. Generate an initial guess for $\dot{N}_i(\Omega) = \dot{N}_{i,s,z}$ at each aggregate state.

3. Given these guesses, solve the value function (22) for values $v_{i,s}(\Omega) = v_{i,s,z}$ at each idiosyncratic productivity node and aggregate state node. In solving (22), replace $\lambda_1(\Omega)$ with the current guess $\lambda_{1,s,z}$, and the drift term, $\sum_{j=1}^{I} \frac{\partial v_i(\Omega)}{\partial N_j} \dot{N}_j(\Omega)$, using i) the current guesses for the value function derivative implied by $h^1_{i,s,j}$ and $h^2_{i,s,j}$ and ii) the current guess for the drifts $\dot{N}_{i,s,z}$.

4. Using the new values of $v_{i,s,z}$, perform OLS regressions on (33) to update the parameters $h^0_{i,s}$ $h^1_{i,s,j}$ and $h^2_{i,s,j}$ with dampening.

5. Using the new values of $v_{i,s,z}$, calculate the new policy functions for job creation and destruction. Use these to update the drifts $\dot{N}_{i,s,z}$ using (23) and the offer arrival rates $\lambda_{1,s,z}$ using (27), both with dampening.

6. Return to step 3 and iterate to convergence.

As a measure of the accuracy of our second-order approximation, the R^2 of the regressions used to fit the polynomial is 99% on average across the $I \times S = 15$ regressions. This R^2 is a measure of the error between the predicted value from the second-order polynomial and the exactly computed value from the HJB of the $v_{i,s,z}$ on the nodes where the HJB is evaluated.

With our approximated policy function parameters $h^0_{i,s}$, $h^1_{i,s,j}$, and $h^2_{i,s,j}$ in hand, we can simulate the aggregate model, calculating all other objects exactly using the true nonlinear equations of the model. When estimating the model, we simulate using a one month

aggregate time step $\Delta t = 1$, but finer grids do not affect the results. For most data comparisons, we aggregate up to quarterly data via averaging and HP-filter the model data as in the data.

Post estimation, we also simulate our impulse response to a typical recession using the same procedure. We additionally simulate the age and size distributions over time, by first solving for the aggregate dynamics, and then extending our age and size distribution codes to allow for aggregate dynamics.

Overview of estimation procedure: We pre-set six parameters, and our estimation procedure then chooses 23 parameters to minimize the distance to a large number of moments. To speed up the estimation, we split the estimation into two layers: an inner loop and an outer loop. Conditional on outer loop parameter values, in the inner loop we solve for the values of 11 parameters to exactly hit 11 moments. Intuitively, each of these 11 parameters has a tight link to a particular moment, which we are able to exploit to quickly solve for the value of these parameters. In the outer loop, we use the remaining 12 parameters to minimize the average distance to a set of 18 moments using a global minimization routine, every time repeating the inner loop procedure.

A key step in speeding up our estimation is that we calibrate many parameters in the non-stochastic steady state of the model (i.e. a version of the model without aggregate shocks) as is standard in heterogeneous agent modelling. In brief, the procedure operates as follows:

1. Guess values for the 12 outer loop parameters.
2. Given the current guess for the outer loop parameters, use the inner loop to exactly solve for the 11 inner loop parameters, to exactly hit the inner loop moments. These moments are calculated in the non-stochastic steady state of the model.
3. Given the values of the inner and outer loop parameters, now solve the full model (out of steady state), and simulate to construct aggregate time series.
4. Calculate the moments used in the outer loop. The moments related to the firm age distribution are calculated from the non-stochastic steady state, and the moments related to the business cycle are calculated from the business cycle simulation.

5. Calculate the distance measure of the outer loop moments to the moments in the data. Update the outer loop parameters using the global minimization routine, and return to step 2. Repeat until the global minimization routine completes.

For our global minimization routine in the outer loop, we program a simplification of the "TikTak" algorithm of Arnoud et al. (2019). Specifically, we draw 6,000 initial guesses of the outer loop parameters from a Sobol set, and calculate the outer loop moments at each guess. We then choose the five best performing guesses and run a local optimizer (pattern search) at each to find the local minima, and choose the lowest error among these as our final estimate.

B.4 Further details of calibration and parameterization

As we impose a relatively small number of productivity states, we use parameter choices to impose the key behaviour (perform JC or not, perform JD or not, replacing quits or not) of each state, rather than letting the estimation decide. The estimation is allowed to affect behaviour within each node (for example the level of JC in the node, if it is positive) and the probabilities that nodes are drawn.

Additional flow cost of capital maintenance: In the estimation we impose that $i^h = 3$ in steady state, which requires that $v_2 < c_0 < v_3$. This could be imposed using a penalty function approach, which penalises the SMD error whenever either of these inequalities is violated. We take a simpler approach, which speeds up the estimation (by allowing more parameters to be chosen in the inner loop) at the cost of introducing one new parameter, the flow cost of capital maintenance. Specifically, we firstly impose $p_2 < p_3$ in the estimation, which ensures that $v_2 < v_3$. Secondly, we then choose c_f in the inner loop to ensure that $c_0 = 0.5(v_2 + v_3)$, which guarantees that $v_2 < c_0 < v_3$. Intuitively, we thus introduce the flow cost of capital to shift the average value to ensure that c_0 lies exactly in the middle of v_2 and v_3 in steady state. The estimated value of this cost is small, at 0.05, which is a small fraction of average productivity (which is one) and small compared to the hiring cost c_0.

We also impose that i^h should not move over the business cycle, so that $i^h(\Omega) = 3$ almost always during a simulation. We do this by computing the fraction of periods

where $i^h(\Omega) \neq 3$ during our business cycle simulation and adding this as a penalty in the SMD function. At the estimated parameters, $i^h(\Omega) \neq 3$ only 0.01% of the time during our long business cycle simulation, and $i^h(\Omega) = 3$ at all times in our typical recession impulse response function plots.

Placement of cost function support parameters: We place the lower support of H^{JC} and upper support of H^{JD} so that (in steady state) i) firms with $i = 1, 2, 3$ perform JD in response to the H^{JD} shock, but never JC in response to the H^{JC} shock, and ii) firms with $i = 4, 5$ perform JC but not JD. To do this requires that i) $v_3 < \bar{c}^{JD} < v_4$, so that firms with $i = 4$ have high enough value to survive any draw of c^{JD}, and ii) $v_3 < \underline{c}^{JC} + c_0 < v_4$, so that firms with $i = 3$ have low enough value so that the minimum cost of performing JC is too high. In order to not introduce additional degrees of freedom into the model, we simply set $\bar{c}^{JD} = 0.5(v_3 + v_4)$ and $\underline{c}^{JC} = 0.5(v_3 + v_4) - c_0$, which we can impose simply at every iteration of the inner loop.

***JC* and *JD* definition in the model vs. the data:** The definitions of JC and JD in the model are built to correspond as closely as possible to their notions in the data. However, practical computational limitations mean that their definitions are not identical. In particular, in the data JC is defined as the sum of employment increases across firms which saw an increase in employment between two dates, and JD is defined as the sum of employment falls at firms which saw a decrease in employment (see, e.g., DFH). Computing these measures exactly therefore would require simulating a panel of firms over the business cycle, which slows down the estimation and simulation of the model.

Instead, we are careful to segment our model so that firms in states $i = 1, 2, 3$ have declining employment at any instant of time, and firms in states $i = 4, 5$ have increasing employment at any instant in time. Abstracting for the moment from entry and exit, JC can therefore be measured as the employment change at state $i = 4, 5$ firms, which corresponds exactly to the job creation flow $\sum_{i=1}^{5} jc_i(\Omega) N_{i,t}$ in the model, since $jc_i(\Omega) = 0$ for $i = 1, 2, 3$. Similarly, JD can be measured as the employment change at state $i = 1, 2, 3$ firms, which corresponds to $\sum_{i=1}^{5} jd_i(\Omega) N_{i,t}$ since $jd_i(\Omega) = 0$ for $i = 4, 5$. This is complicated slightly by two factors. Firstly, firms may switch from being in state $i = 1, 2, 3$ to $i = 4, 5$ within the same quarter, which is the period over which JC and JD are

measured in the DFH data. This means that measured and theoretical measures may differ, as a firm might receive a JC and JD shock in the same quarter. However, this is a rare occurrence, as our productivity shocks occur on average only once per quarter. Secondly, we must account for entry and exit. Entry is simple, as all entrants create jobs and therefore can simply be added to the job creation flow. Exit complicates the analysis somewhat, as even firms with $i = 4, 5$ might receive the δ_F exit shock. However, this shock is calibrated to be very rare so this does not matter much in practice.

In order to check the applicability of our JC and JD measures, we simulate a large panel of firms after estimating the model. We focus on the steady state, and simulate the firms for one quarter of data, with the firms drawn from the ergodic productivity and size distribution. We compute the job destruction rate exactly as is done on the data, and find a quarterly rate of 6.2%, while the theoretical rate as calculated by $jd = \sum_{i=1}^{I} \left[jd_i + \delta_F\right] N_i / N$ exactly equals the targeted value of 7.0%. While not identical, this difference of roughly 10% is in line with the average error of the other moments in the outer loop of our SMD routine.

Comparison of firm-level autocorrelation to data: Our parameter values generate an autocorrelation of 0.21 for yearly productivity in a year-averaged simulated firm-level productivity series, or 0.53 for quarterly productivity (within mature firms). Elsby et al. (2017) discuss empirical estimates of the persistence of idiosyncratic productivity, and find a wide range of values. Our estimate lies within this range. Specifically, Cooper, Haltiwanger, and Willis (2015) imply a quarterly autocorrelation of 0.4, which is below our value, while Abraham and White (2008) imply 0.68 which is above our value. While our productivity process has relatively low persistence, our constant returns to scale structure means that current productivity controls the growth rate of employment, not the level. Hence, even temporary productivity shocks will generate permanent effects on a firm's employment.

Further details of the Outer loop: The 12 parameters chosen in the outer loop fall into two broad categories: those relating to the firm age distribution $(p_1, p_2, p_5, \gamma_{11}, \gamma_{55}, \gamma_{01}, \gamma_{05}, N_0)$ and those relating to business cycle moments $(\xi_e, \xi_{JC}, \xi_{JD}, c_0)$. The 12 age distribution moments are computed using the non-stochastic simulation of the age distribution in steady

state. The six business cycle moments are computed by simulating the model for 1,000 years. The simulated data are then aggregated to a quarterly frequency and HP-filtered, as in the data. We apply a simple diagonal weighting to the moments, with the total weight given to business cycle moments slightly overweighted to ensure the model performs well on both business cycle and steady state dimensions.

The estimation finds that the productivity grid is non-monotone, as $p_1 > p_2$ and $p_5 < p_4$. Nonetheless, firm *values* remain monotone, with $v_i < v_{i+1}$ for all i, which is sufficient for the job ladder to be directed monotonically by i, and hence for our notion of equilibrium to remain well defined. The disconnect between the ordering of productivities and values occurs simply because the entrant states $i = 1, 5$ are more persistent than the mature states.

Further details of the Inner loop: We present below a list of the 11 parameters (plus the additional parameter c_f) chosen in the inner loop, and how the moment used to choose the parameter is calculated. The inner loop is terminated when the error in all moments is below 10^{-8}. All moments are computed in the non-stochastic steady state of the model.

p_4 is chosen to normalise aggregate labor productivity to one. p_3 is chosen to generate a standard deviation of idiosyncratic productivity of 30%. γ_2 is chosen to match that 80% of quits (and replaceable layoffs) are replaced, calculated as $qfr = \int_{x^h}^1 q(x)dG(x) / \int_0^1 q(x)dG(x)$.

$\underline{c}_{JC}$, $\bar{c}_{JD}$, and c_f are set as discussed above. The firm entry flow μ_0 is set to hit the average size of firms, measured as N/M, where M is the mass of firms with at least one employee. μ_1, δ_D, λ_u, and ϕ are set to match the theoretically computed JC, JD, layoff, and EE quit rates. Finally, $w_{\min}$ is set to match the labor share, defined as $LS = \mathrm{E}w \times N/Y$, where $\mathrm{E}w = \int_0^1 w(x)dG(x)$.

C Additional tables and figures

Table 3: Equilibrium policies and values in steady state

	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$
v_i	0.7285	0.7316	0.8278	2.6324	2.6551
p_i	0.9265	0.6747	0.7133	1.3076	1.1379
jc_i	0	0	0	0.0445	0.0458
jd_i	0.0995	0.0987	0.0295	0	0
njc_i	-0.0996	-0.0988	-0.0296	0.0445	0.0457
N_i/N	0.0030	0.1213	0.3848	0.4849	0.0061
M_i/M	0.0471	0.1429	0.3989	0.4061	0.0050

Table summarizes the value and policy functions in steady state across productivity levels $i = 1, ..., I$. v_i is firm value, and p_i productivity. Value is monotonically increasing across states. jc_i and jd_i are job creation and destruction rates per employee for incumbent firms, excluding the δ_F exit shock. njc_i is net job creation: $njc_i = jc_i - jd_i - \delta_F$. The final two rows give the fraction of employment and active firms (with at least one employee) repectively at each i in the ergodic distribution.

Figure C.1: Impulse Response Function - Cyclical behaviour of key aggregates

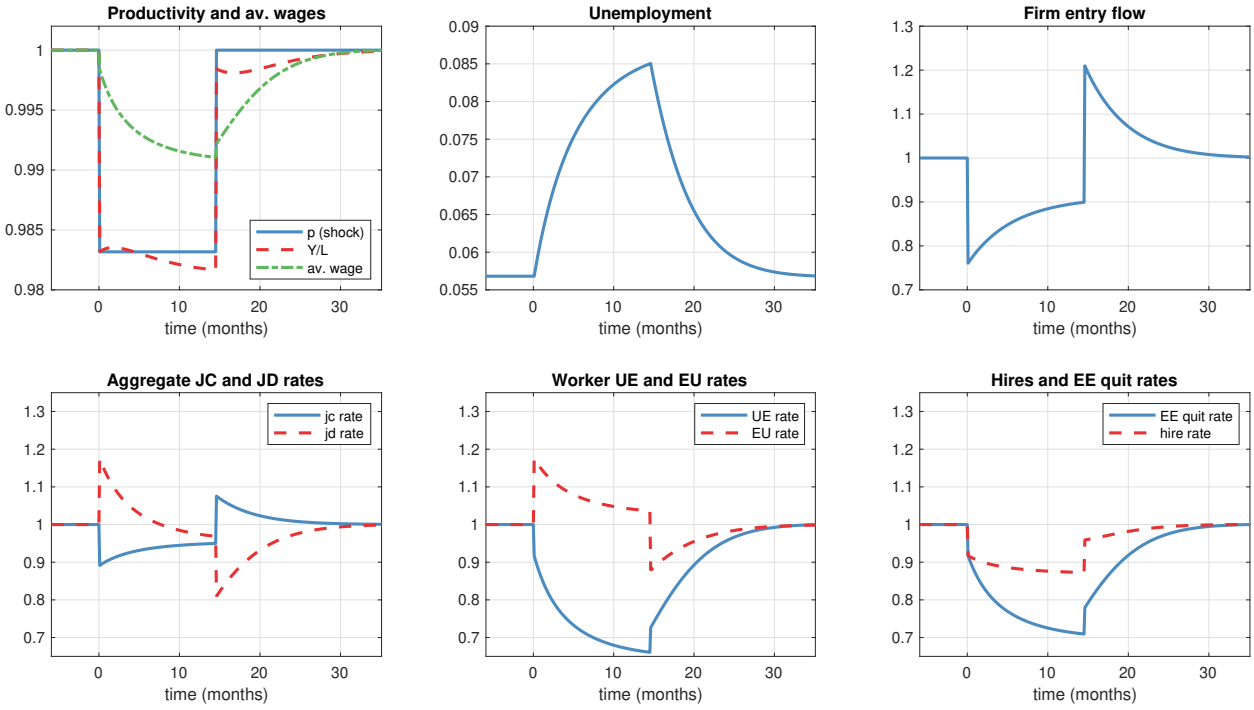


Figure plots additional aggregates from our typical recession impulse response function. See Figure 10 for further details of the experiment.

Figure C.2: Experiment 1: JC and JD by firm age, as prop. of total employment

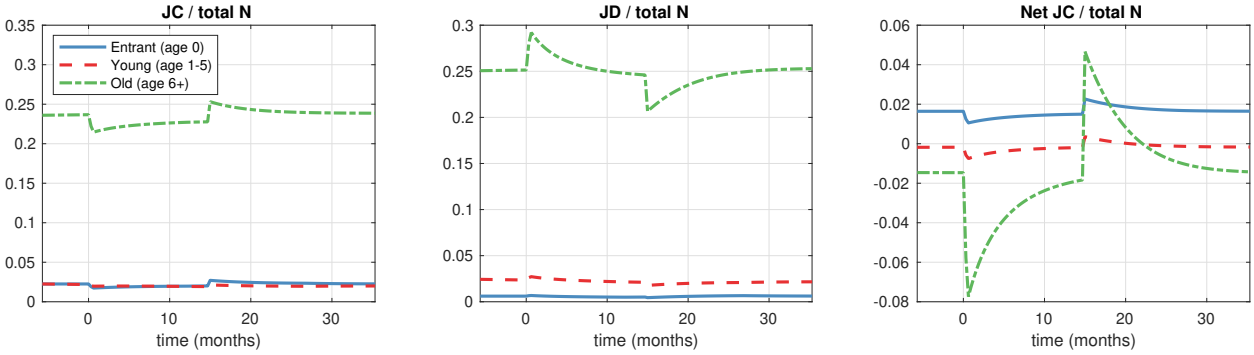


Figure plots JC and JD flows by firm age for our typical recession experiment. See Figure 10 for further details of the experiment. JC and JD flows are yearly, and computed as $1 - e^{-12r}$, where r are the average theoretical monthly rates within each bin.

References

[1] Achdou, Y, J. Han, J.-M. Lasry, P.-L. Lions, B. Moll. 2021. "Income and Wealth Distribution in Macroeconomics: A Continuous-Time Approach," *The Review of Economic Studies*, 2021.

[2] Arnoud, A, F. Guvenen, T. Kleineberg. 2019. "Benchmarking Global Optimizers," *mimeo*.

[3] Davis, S. J., R. J. Faberman and J. Haltiwanger. 2012. "Labor market flows in the cross section and over time," *Journal of Monetary Economics*, 59(1): 1-18.

[4] Elsby, M., R. Michaels and D. Ratner. 2017. "Vacancy Chains". University of Edinburgh, Department of Economics, mimeo.

[5] Shimer, R. 2005. "The Cyclical Behavior of Equilibrium Unemployment and Vacancies," *American Economic Review*, 95(1): 25-49.

[6] Young, E. R. 2010. "Solving the incomplete markets model with aggregate uncertainty using the Krusell-Smith algorithm and non-stochastic simulations," *Journal of Economic Dynamics and Control*, 34(1): 36-41.