Sorting in the Labor Market: Theory and Measurement

Rafael Lopes de Melo†

First Draft: February 15, 2008
This Version: November 10, 2008

Job Market Paper

Abstract

Are more skilled workers employed by more productive firms? Are complementarities important in production? We provide four contributions to the measurement of sorting. First, we introduce a frictional sorting model to show that the standard empirical method used to measure sorting in the labor market can be biased in favor of not detecting sorting. Second, we isolate the economic mechanism responsible for this bias. Third, we propose an alternative method to detect sorting that is immune from this bias. Finally, we apply both methods to a Brazilian matched employer-employee dataset, RAIS. We confirm the absence of sorting when using the first method, but the second method reveals strong sorting. According to the model, these two apparently contradictory empirical findings suggest that sorting is widespread in the labor market, and complementarities are important in production.

*This paper is part of my dissertation. I am indebted to Giuseppe Moscarini for the extensive advice and support. I would like to thank Fabian Lange, Fabien Postel-Vinay, Björn Brügemann, Jeremy Lise, Pieter Gautier, Rasmus Lentz, Joe Altonji, Jesus Fernandez-Villaverde, Philipp Kircher, Jan Eeckhout and Theodore Papageorgiou for their valuable comments and help. Very useful feedback was also received from participants at the macroeconomics and labor/public finance seminars at Yale University, at the Structural Models of the Labor Market Conference, UCL, 2008, at the 2008 Summer Meeting of the Society for Economic Dynamics, at the Search and Matching Workshop at University of Pennsylvania, at a seminar at PUC-Rio, and at a seminar at Ipea-Brasilia. Confidential data collected by the Brazilian labor Ministry were used in this paper. I thank Ipea-Brasilia for granting my access to the data and for all the support given while I was doing my research there. I alone am responsible for any shortcomings in the paper.

†Department of Economics, Yale University (e-mail: Rafael.Melo@yale.edu).
1 Introduction

Are more skilled workers employed by more (or less) productive firms? Are complementarities important in production? This paper analyzes the assortative matchup between heterogeneous firms and workers in the labor market. An extensive literature in economic theory studies the sorting patterns of heterogeneous agents. Important examples are partners in a marriage, buyers and sellers negotiating a product, students and teachers, players and teams, and workers and firms. A common feature among all these examples is that positive (negative) complementarities in production between agents induce positive (negative) sorting in equilibrium. For example, Becker [5] shows that if the production function is supermodular, the equilibrium exhibits perfect sorting between the partners in a marriage: the most desirable individuals get together, and vice-versa.¹

The current empirical evidence suggests, however, that little sorting actually takes place between workers and firms in labor markets—which would reflect the absence of complementarities.² The standard empirical methodology adopts the framework of Abowd, Kramarz et al [3, 2, 1] (hereafter, AKM), which consists of estimating a log-linear wage regression with worker and firm fixed effects. The procedure then uses the estimated fixed effects of workers and firms as proxies for unobserved heterogeneity. One standard—and very robust—result in this literature is that the correlation between these fixed effects is zero (or even negative). Taken at face value, this means that there is little sorting in labor markets.

This paper revisits this result by introducing a frictional model of the labor market with complementarities in production (worker-firm) that exhibits positive sorting between workers and firms in equilibrium. The model is consistent with the current empirical evidence, as well as new evidence, collected using a novel Brazilian matched employer-employee dataset. Moreover, we argue that it is hard to reconcile our new evidence with an economy that features no complementarities in production and little sorting.

The answers to our questions are important as the nature of production and allocation processes can matter a great deal for efficiency and inequality. For example, if complementarities are important in production, frictions that prevent workers from finding their "right job" can lead to large output losses.³ These losses are much less relevant if inputs are substitutable.⁴

¹If \( f(x,y) \) is smooth, then supermodularity is equivalent to \( f_{xy} > 0 \). Also, he uses the core concept of equilibrium.
²This statement refers to sorting conditional on observable characteristics. Mendes et al [17] find a positive association between workforce education and firm productivity measures using Portuguese manufacturing data.
³This is only relevant for policy if the equilibrium is constrained inefficient—i.e, given the level of frictions. Gautier and Teulings [26] discuss reasons for inefficiencies in an environment with search frictions, and heterogeneous workers and firms.
⁴For example, suppose the output from a match of worker \( h \) with firm \( p \) is \( F(h,p) \). One can easily see that if \( F(h,p) = h + p \), there is little output loss if resources are randomly allocated.
We construct a model of the labor market with four main ingredients: heterogeneity of both workers and firms; complementarities in production, which induce positive sorting in equilibrium; search frictions and on-the-job search in order to make it more realistic and comparable to the data. These frictions add noise to the sorting process, causing the agents to accept suboptimal partners in order to avoid idleness. Finally, in the model, firms have limitations in their capacity to post new vacancies. This creates ex-ante rents for vacancies and creates a reasons for firms to reject some workers in equilibrium.

We use a Brazilian matched employer-employee dataset to produce empirical moments that are relevant to the question of sorting in the labor market. This type of dataset is very useful, if not essential, to our questions. We begin by using the AKM methodology, and confirm their standard result in our Brazilian dataset. In addition, we use the same output of the wage regression to compute a new moment: the correlation between the fixed effects of workers and the average fixed effects of his/her coworkers. We find that this new moment has a value close to 0.3, which is much larger than the traditional worker-firm correlation. This fact suggests sorting and appears in contradiction with the previous evidence. The value of both correlations persist, even controlling for observables such as gender, education, occupation, industry and location. For example, the value of the new correlation is 0.39 for managers and 0.25 for waiters.

Our model provides an explanation for these facts. First, the model exhibits positive sorting between workers and firms in equilibrium, which explains the positive correlation among coworkers: the "good" workers work for the "good" firms, so they end up with other "good" workers as well. The surprising result is that the model can account for the zero correlation between worker and firm fixed effects, despite featuring positive sorting. This is due to non-monotonicities in the wage equation caused by the interaction of wage bargaining and limitations in the capacity of the firms to post new vacancies. High productivity firms have better outside options than their low-productivity counterparts, which causes downward pressure on the wages of their workers. This is particularly relevant for low-skilled workers, who may be paid less by the more productive of two firms. These non-monotonicities do not affect the ordering of wages across workers, which explains why our new moment correctly captures the degree of sorting while the standard one does not. Finally, shutting down complementarities in production or the capacity constraints of firms compromises the ability of the model to explain these facts.

The rest of the paper is structured as follows. Section 2 describes the related literature, Section 3 the theoretical model, Section 4 the data and facts about sorting patterns and Section 5 some preliminary results on matching the model to the data. Section 6 shows how to endogeneize the vacancy formation in the theoretical model. Section 7 concludes.
2 Related Literature

This paper relates to several strands of literature. First, the basis of the empirical methodology used throughout the paper comes from a series of papers by Abowd, Kramarz et al [3, 1, 2]. With the advent of datasets that follow workers and firms at the same time, they use the mobility of workers across firms to estimate a wage equation with worker and firm fixed effects. The results from these papers provided a number of insights and puzzles.\(^5\)

Second, our model is derived from the theoretical assignment literature. Becker [5] presents a model with two sided heterogeneity, and shows that if there are enough complementarities in production the equilibrium exhibits perfect sorting and is efficient. Shimer and Smith [25], Lu and McAfee [16] and Sattinger [16] introduce search frictions in the Becker model, which adds noise to the equilibrium allocations and requires stronger complementarities for the equilibrium to exhibit positive assortative matching. Eeckhout and Kircher [10] develop a directed search model that is an intermediate case between the frictionless model and the economy with search frictions. That framework requires less complementarities than the search case, but more than the frictionless economy, to induce positive sorting.

Thirdly, a recent stream of papers applies models with two-sided heterogeneity and bilateral matches to the labor market. Woodcock [27] uses a search model with no complementarities in production and match specific learning, which features negative sorting in equilibrium, as a way to justify the negative correlation in wage fixed effects in the AKM \(^6\) regressions. Lentz [14] and Bagger and Lentz [4] formulate a search model with endogenous search intensity, and structurally estimate it using the Danish matched employer-employee dataset. They need weaker complementarities in production to have positive sorting, as the high-skill worker search with a higher intensity. Lise, Meghir and Robin [15] build on the model of Shimer and Smith [25], introducing on-the-job search, a wage mechanism as in Cahuc, Postel-Vinay and Robin [8] and a dynamic process for firms that causes them to change productivity over time. Eeckhout and Kircher [9] use variations from the Becker model [5] to argue that, from wage data alone, one cannot distinguish a model that features positive sorting from a model of negative sorting. However, they also argue that wages can give information about the strength of sorting, which is consistent with our results. Moreover, they argue that the strength is more important than the sign for questions related to efficiency and inequality.

\(^6\)His model implies a specification like in AKM in levels, not in logs.
3 The Model

The model shares the basic features of Shimer and Smith [25], plus specific features of labor markets such as wages, unemployment compensation and asymmetries between worker and jobs.\(^7\) It is a continuous time economy with heterogeneous agents, complementarities in production and search frictions. Production in this economy happens in bilateral matches. Workers meet jobs according to a stochastic process, i.e. matching is not instantaneous because of search frictions. They observe each other’s types, decide if they want to pair up or not, and, upon matching, produce output and split the proceeds according to a pre-designed rule. Similarly as in their model, matches are dissolved exogenously at a given rate. However, we introduce a modification to the model, on-the-job search. This has been acknowledged as an important feature of labor markets and lets us talk about endogenous separations.

One important assumption is that there is a fixed stock of workers of heterogeneous skills and and a fixed stock of jobs of different productivity. Whereas the former assumption is standard, the assumption of a fixed stock of jobs is not. In fact, other authors model vacancy creation by assuming free entry at each firm type.\(^8\) We depart from this tradition, and follow strictly Shimer and Smith [25], in order to introduce a mechanism in the model which allows us to better explain the data. In Section 6 we specify a technology for entry and vacancy creation which preserves our results without assuming a fixed stock of jobs.

3.1 The Environment

There is a continuum of workers and jobs, each indexed by a type. We exogenously assign a value \(h \in \mathbb{R}\) for each worker, which can be interpreted as his/her human capital, and value \(p \in \mathbb{R}\) for each job, which can be interpreted as that firm’s productivity. These exogenous random variables have atomless distributions \(L : \mathbb{R} \to [0, 1]\) and \(G : \mathbb{R} \to [0, 1]\), respectively, with density functions \(l(h)\) and \(g(p)\). These types are assumed to be perfectly observable and the distributions are known to all agents. The mass of workers is normalized to 1, and the mass of jobs per worker is \(\bar{N}\).

Workers and firms discount time at rate \(r\) and have linear preferences.

When a worker of type \(h\) and a firm of type \(p\) match they produce a flow output of \(F(h, p)\) at every instant. The production function is an essential element to our model because it has strong implications for the sorting patterns of workers. Shimer and Smith [25] provide a full character-

---

\(^7\)Shimer and Smith [25] study the problem of generic symmetric agents.

\(^8\)For example, Gautier and Teulings [26, 11] and Lentz [14].
ization of the necessary and sufficient conditions for positive assortative matching in their model. In particular, they show that if the production function, the log of its first partial derivatives and the log of its cross derivative are all supermodular then the economy exhibits positive assortative matching. Although our framework is not identical to theirs—we do not assume bounded support and we introduce on-the-job search—which prevents us from relying on their results, we use a production function that satisfies these requirements in our empirical specification: \( F(h, p) = hp \).

Unemployed workers receive a flow value \( b(h) \), which may depend on their type, \( h \), and vacant jobs receive zero flow value.

We assume a random search technology where worker/jobs meet each other at a finite Poisson rate. Upon meeting, they take a draw from the corresponding idle type distribution—e.g., if an unemployed worker meets an employer he/she takes a draw from the vacancies distribution. Unemployed workers meet vacancies at the rate \( \lambda^W \). Employed workers contact vacancies at rate \( \kappa \lambda^W \), where \( \kappa \) represents the relative efficiency of the on-the-job search, as compared to search when unemployed. On the firm side, vacancies meet unemployed workers at a rate \( \lambda^F \) and employed workers at rate \( \phi \lambda^F \). We do not allow firms to practice on-the-job search.\(^9\) Finally, all matches are exogenously destroyed at rate \( \delta \). For our purposes, these rates are treated as exogenous. Moreover, we assume that the economy is in steady state.

The presence of search frictions creates a temporary bilateral monopoly power in a match. We follow several others and adopt the generalized Nash bargaining solution to determine wages. We choose this particular mechanism because it simplifies the solution of the problem, while embedding the economic forces that we believe are important to explain the observed phenomena. In the presence of on-the-job search, this solution raises a complication. When an employed worker contacts another firm, a three-agents negotiation problem arises, requiring us to specify a rule for this situation. We follow Moscarini [20]. Workers and firms split the surplus at every instant according to the standard Nash Bargaining solution, with \( \beta \) being the bargaining power of the workers. However, when a new firm shows up, the two firms engage in an ascending first price auction for the services of the worker. The firms bid a non-negative lump-sum transfer to the worker. Whoever wins, pays the transfer, retains the worker and resumes the match with that worker splitting proceeds using the Nash rule. These lump-sum transfers can be interpreted as a signing bonus if

\(^9\)Kiyotaki and Lagos [13] introduce a theoretical framework that includes on-the-job search by the firms (replacement hiring). Similarly as in our model, capacity constraints on the firms provide the incentives for the firms to engage in such activities. Adding on-the-job search by firms would be analogous to adding on-the-job search by workers, with the caveat that we need to specify a solution for a quad-lateral negotiation problem (when employed worker meets filled jobs). It would cause the firms to be less picky when considering a new match, but it would provide a new avenue for firms to find the "right worker". Furthermore, It adds an extra compensation for workers through bargaining for the fact that the firm has the option of searching for new workers.
the competing firm wins the auction, or a retention bonus if the incumbent firm wins.

Before describing how the equilibrium of the model is determined, we introduce the following notation. An employed worker of type \( h \) working for a firm of type \( p \) has value \( V^E(h,p) \), or, if this workers is unemployed, \( V^U(h) \). A job of type \( p \) has value \( J^E(h,p) \) if matched with a worker of type \( h \), or \( J^U(p) \) if vacant. The surplus of a match between worker \( h \) and job \( p \) is \( S^E(h,p) \equiv [V^E(h,p) - V^U(h)] + [J^E(h,p) - J^U(p)] \). The wage in any match \((h,p)\) is \( w(h,p) \).

Matching sets are characterized by indicators functions. For an unemployed worker of type \( h \) who meets a vacant job of type \( p \), the indicator \( \alpha^U(h,p) \) assumes a value of 1 if they match, and 0 otherwise. Alternatively, if worker \( h \) is currently working for firm \( p \), but contacts firm \( p' \), then \( \alpha^E(h,p,p') \) equals 1 iff the worker switches firms. The steady state distribution of employed matches is \( \Gamma(h,p) \), with density \( \gamma(h,p) \). The density of unemployed workers for each type is \( l(h) \), and the density of vacancies is \( g(p) \). Finally, the unemployment rate, \( u \), and the number of vacancies in the economy, \( v \), are also determined endogenously.

### 3.2 Values and Decisions

An employed worker \( h \) at a firm \( p \) earns a flow wage of \( w(h,p) \). At a rate \( \delta \) his/her match is destroyed exogenously. At a rate \( \kappa \lambda^W \) the employed worker meets another firm, upon which he/she may move or stay at the current job, and she earns a transfer of value \( T(h,p,p^*) \) from the winner of the auction. To simplify notation, I suppress the arguments of \( T(h,p,p^*) \). This implies the following value equation:

\[
\begin{align*}
  rV^E(h,p) &= w(h,p) + \delta [V^U(h) - V^E(h,p)] + \\
  &\kappa \lambda^W \int_{-\infty}^{\infty} \alpha^E(h,p,p^*) [V^E(h,p^*) - V^E(h,p)] + T \right] g(p^*) dp^*. \quad (1)
\end{align*}
\]

An unemployed worker of type \( h \) enjoys flow unemployment benefits \( b(h) \) and receives job offers at rate \( \lambda^W \):

\[
\begin{align*}
  rV^U(h) = b(h) + \lambda^W \int_{-\infty}^{\infty} \alpha^U(h,p^*) [V^E(h,p^*) - V^U(h)] g(p^*) dp^*. \quad (2)
\end{align*}
\]

A productive job of type \( p \), employing a worker of type \( h \), enjoys flow profits \( F(h,p) - w(h,p) \), but can be dissolved either by the \( \delta \) shock or through poaching by other firms. Moreover, if another
firm appears but loses the auction, the firm pays $T$ to the worker:

$$rJ^E(h,p) = F(h,p) - w(h,p) + \left[ \delta + \kappa \lambda^W \int_{-\infty}^{\infty} \alpha^E(h,p,p^*) g(p^*) dp^* \right] [J^U(p) - J^E(h,p)]$$

$$- \kappa \lambda^W \int_{-\infty}^{\infty} T (1 - \alpha^E(h,p,p^*)) g(p^*) dp^*. \quad (3)$$

Finally, a vacancy does not earn any flow values, but can hire both unemployed or employed workers (upon paying a transfer $T$):

$$rJ^U(p) = \lambda \int_{-\infty}^{\infty} \alpha^U(h^*,p)[J^E(h^*,p) - J^U(p)] l(h^*) dh^*$$

$$+ \phi \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \alpha^E(h^*,p^*,p)[J^E(h^*,p) - J^U(p) - T] \gamma(h^*,p^*) dh^* dp^*. \quad (4)$$

As previously described, wages are determined by the Generalized Nash bargaining solution with repeated first price auctions. Recall that this type of mechanism involves the standard Nash bargaining solution while a match goes on, but when the worker contacts a second firm, the two firms engage in an auction for the services of the worker. Although this game features multiple equilibria, Moscarini [20] argues that the most plausible equilibrium for this game is the one where the firm that produces the highest surplus with that worker wins the auction, pays no lump-sum transfer and immediately starts producing with the worker, splitting the proceeds according to the same rule. This occurs because the firm with the highest surplus can always outbid the other firm and, knowing that, the other firm does not even bother bidding. It is easy to see that this equilibrium is the only one that survives if we introduce an arbitrarily small cost of bidding. Therefore, $T(h,p,p^*) = 0, \forall (h,p,p^*)$. Standard results then imply that$^{10}$

$$S^E(h,p) = \frac{[J^E(h,p) - J^U(p)]}{(1 - \beta)} = \frac{V^E(h,p) - V^U(h)}{\beta}. \quad (5)$$

Plugging Equations (1) and (3) in this formula and rearranging we solve for the wage

$$w(h,p) = \beta [F(h,p) - rJ^U(p)]$$

$$+ (1 - \beta) \left[ rV^U(h) - \beta \kappa \lambda^W \int_{-\infty}^{\infty} \alpha^E(h,p,p^*) S^E(h,p^*) g(p^*) dp^* \right]. \quad (6)$$

From this expression, we can see that wages are a function of output, the outside options of the

worker and the firm, and an extra term that reflects the amount the worker has to compensate the firm for having the option to search for other jobs.

We use these results to rewrite the surplus function in recursive form. Plug Equations (1) to (5) in the definition of the surplus and rearrange to obtain:

\[
S^E(h, p) = F(h, p) - b(h) - (1 - \beta) \lambda^F \int_{-\infty}^{\infty} \alpha^U(h^*, p) S^E(h^*, p) l(h^*) dh^* - r + \delta + \kappa \lambda^W \int_{-\infty}^{\infty} \alpha^E(h, p, p^*) g(p^*) dp^*
\]

\[
= \beta \lambda W \int_{-\infty}^{\infty} \left[ \alpha^U(h, p^*) - \kappa \alpha^E(h, p, p^*) \right] S^E(h, p^*) g(p^*) dp^* + (1 - \beta) \varphi \lambda^F \int_{-\infty}^{\infty} \alpha^E(h^*, p^*, p) S^E(h^*, p) \gamma(h^*, p^*) dh^* dp^*.
\]

In intuitive terms, the surplus is the present discounted value of output, plus the expected gains from on-the-job search minus the unemployment benefits and losses in outside options. In this case, the worker loses the option of searching while unemployed and the firm loses the option of searching for alternative workers (employed or unemployed) while vacant. It is easy to see that all the other values and wages can be calculated from \(S^E(h, p)\) solving (7).

When idle agents meet, they decide to pair up if doing so increases their values. It is easy to see from Equation (5) that this is equivalent to choosing the option that provides the largest surplus. Thus, when an unemployed worker meets a vacancy, the decision is given by

\[
\alpha^U(h, p) = 1 \left[ S^E(h, p) > 0 \right]
\]

whereas when an employed worker meets a second firm the decision is

\[
\alpha^E(h, p, p') = 1 \left[ S^E(h, p') > S^E(h, p) \right].
\]

### 3.3 Steady State Flows

We now describe the equilibrium equations that jointly determine the stationary distributions \(\Gamma(h, p), l(h)\) and \(g(p)\), and the idleness rates \(u\) and \(v\).

The first equilibrium equation that we describe is between the flows in and out of employed matches of type \((h, p)\). The derivation is in the appendix. In the outflows, some matches are
dissolved exogenously and some are dissolved due to on-the-job search. In the inflows, matches are formed from idle pairs and with workers retained from other matches. If \( \alpha_U(h,p) = 0 \) then \( \gamma(h,p) = 0 \). Otherwise, \( \gamma(h,p) \) is determined by equating the inflows to the outflows:

\[
\delta (1-u) \gamma(h,p) + \kappa \lambda^W (1-u) \left[ \int_{-\infty}^{\infty} \alpha^E(h,p,p^*) g(p^*) dp^* \right] \gamma(h,p) = u \lambda^W l(h) g(p) + \kappa \lambda^W (1-u) g(p) \int_{-\infty}^{\infty} \alpha^E(h,p^*,p) \gamma(h,p^*) dp^*.
\]

Employed workers of type \( h \) have to equal the workers of that type minus the unemployed ones. This implies that

\[
\bar{l}(h) - ul(h) = (1-u) \int_{-\infty}^{\infty} \gamma(h,p^*) dp^*.
\]

The same holds for jobs of type \( p \)

\[
\bar{N}_G(p) - vg(p) = (N-v) \int_{-\infty}^{\infty} \gamma(h^*,p) dh^*.
\]

Next, the unemployment rate is determined by integrating (10) over the full support of \( h \) and \( p \), and its stationarity requires

\[
\delta (1-u) = u \lambda^W \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \alpha^U(h^*,p^*) l(h^*) g(p^*) dh^* dp^*.
\]

Finally, the last requirement for equilibrium is that the total number of employed workers has to equal filled jobs.

\[
1-u = N-v.
\]

### 3.4 Equilibrium

**Definition 1** A steady state equilibrium in this economy consists of values for \( \gamma(h,p) \), \( l(h) \), \( g(p) \), \( u \), \( v \), \( S^E(h,p) \), \( \alpha^U(h,p) \) and \( \alpha^E(h,p,p') \) such that Equations (7), (8), (9), (10), (11), (12), (13) and (14) are satisfied.

Shimer and Smith [25] have a proof of equilibrium existence, but not uniqueness, for the model without on-the-job search.\(^{11}\) We do not have an analogous proof of existence for the augmented model. It is on the research agenda to investigate further existence, uniqueness and efficient computation of equilibrium of this economy.

\(^{11}\)In the case with a quadratic matching function.
3.5 Proposed Algorithm

This model does not have a closed form solution. Thus, we have to solve this model computationally. For our empirical implementation we use the following algorithm to solve the model. First, we discretize the support of the type distributions. Second, we guess an initial value for all endogenous objects. Then we take the following steps:

1. Values and matching sets: given all the guesses we iterate on (7) until finding a fixed point, which yields a value for $S^E(h, p)$. In each step of the iteration, we update the matching sets, $\alpha^U(h, p)$ and $\alpha^E(h, p, p')$, using (8) and (9).

2. Steady state flow equations: given the output from Step 1 we iterate (10) to (14) until we find a fixed point. This gives a value for $\gamma(h, p), \lambda(h), g(p), u$ and $v$.

3. Repeat Steps 1 and 2 until the endogenous distributions converge.

Because we do not have results showing that these mappings are contractions, we cannot anticipate that this algorithm works. However, in practice, this algorithm does work over a wide range of parameters.

3.6 Illustration of the Equilibrium

As an example, Figures 1 and 2 illustrate the equilibrium of a symmetric economy—same distribution of worker and job types and Nash Bargaining parameter equal to 0.5—and production function $F(h, p) = hp$.

The first panel in Figure 1 illustrates the matching set for the idle agents—the X axis has workers and the Y axis has firms. The upward sloping bands are a direct consequence of the equilibrium with positive assortative matching: the low types are rejected by the corresponding high types. However, since workers can search on-the-job they are less picky than firms, which is reflected by the higher area "to the right" of the 45 degrees line. The bottom left panel contrasts the distribution of all workers in this economy with the distribution of the unemployed. The distribution of the idle types is stochastically dominated by the distribution of all workers. This is a clear implication of the positive assortative matching, where the low types are systematically rejected by

---

12 We use the following specification for this graph. The distributions of types are $N(0, 0.5)$ for both workers and jobs. We assume that $N = 1$, which implies that $v = u$. Also, we normalize the transition rates the following way: $\lambda^W = \lambda u = \lambda^F$. Also, $\phi \lambda^F = \psi \lambda (1 - u)$. This reduces the transition parameters of the model to $\lambda$, $\psi$ and $\delta$. We choose $\lambda = 0.75$, $\psi = 0.1$ and $\delta = 0.025$. Unemployment insurances: $b(h) = 0, \forall h$. 


the high corresponding types. We would obtain the same pattern if we plotted the jobs/vacancies distributions, but a slightly smaller discrepancy, again because of on-the-job search.

The upper right panel shows the outside option of workers (or firms). This is clearly convex, reflecting the nature of the production technology. The bottom right panel illustrates the unemployment rate by worker type. Low type workers have a much higher unemployment rate, since the high productivity firms reject them. The unemployment rate increases for the very high types because the high skill workers prefer not to match with low types in order to assure a good match.

Figure 2 plots log-output and log-wages as functions of the worker and the firm’s types. The first thing to notice is that both figures have “valleys”. Those represent the regions outside the matching sets, i.e., whenever $\alpha^U(h,p) = 0$.\(^{13}\) Now, recall that $F(h,p) = hp$. Because of that, log-output grows linearly both in the worker’s and the firm’s type on the valid regions of the matching set. This is in contrast to log-wages, where the relationship can be non-monotonic, for example for low worker types in the graph.

We can observe from the wage function that wages are unambiguously increasing in $h$. However, that is not true when increasing $p$. In this case, for workers with low $h$, increasing $p$ can

\(^{13}\)We input a value slightly smaller than the minimum wage produced by the equilibrium of the model.
 actually reduce wages. For workers with high $h$, increasing $p$ always increase wages. This non-monotonicity of wages with respect to $p$ suggest that the firm fixed effects in a wage regression are unlikely to capture the true ordering across firms productivities. However, the same graph suggests that wage data can still be useful to infer sorting. This is because of two reasons. First, in our model the high-skill workers work for the high-productivity firms. A consequence of that is that these workers work with other high skill workers as well.\textsuperscript{14} Second, wages are monotonic in $h$, suggesting that the relationship between worker and coworker wages reflects that of the primitives.

The fact that $w(h, p)$ is monotonically increasing in $h$, but not in $p$ is more general than this example. In the Appendix, we show, in the case without on-the-job search, that $w_h(h, p) > 0$, for all points in the matching set, and $w_p(h, p) < 0$, for at least one point in the matching set. We do not have yet a proof that this holds in the general case, with on-the-job search, but in all our numerical examples that holds true. It is important to note that the aforementioned non-monotonicities are present even without on-the-job search—i.e., in Shimer and Smith [25].\textsuperscript{15}

\textsuperscript{14}So far the model has been formulated in terms of workers and jobs. In Section 5 we show how to introduce the concept of a firm to the model.

\textsuperscript{15}The reason why the wages can be asymmetric (always increasing in $h$, but not in $p$) in a completely symmetric environment (e.g. if $\Psi = 0$ in the example above) is due to the fact that wages are paid by the firms to the workers and not vice versa. Similar asymmetries are present in $F(h, p) - w(h, p)$. 

---

Figure 2: Log-Output and Log-Wages in the Model
4 Sorting in the Data

4.1 Data Description and Sample Selection

We use the labor market census RAIS (Relação Annual de Informações Sociais), an administrative dataset collected annually by the Brazilian labor ministry, which includes all firms in the Brazilian formal sector and provides information for all their workers. The ministry collects demographic information for workers, such as age, education and sex, some information about establishments, such as sector and location, and provides information about the job, such as the average wage earned during that year, the wage in December, the average number of hours worked, occupation, dates of admission and separation, type of contract, causes for separation.\textsuperscript{16}

We restrict the sample to the Southeast region of Brazil, which includes the states of São Paulo, Rio de Janeiro, Minas Gerais and Espírito Santo. We choose to do so for three reasons. First, this region is the most important economic region of the country, producing approximately 50\% of the country’s GDP. Second, the RAIS is an enormous dataset. Therefore, this reduction in the number of states makes the dataset more manageable. Third, this southeast region has a much smaller level of informal employment than all the other regions of the country (Table 1).

The informality of the labor market is an important feature in Brazil, just as it is in many developing countries. Hereafter, we will not deal explicitly with this feature of the labor market, but we will lump it into a non-employment category. Thus, whenever we say employment, we are referring to formal employment.\textsuperscript{17}

<table>
<thead>
<tr>
<th>Table 1 – Employed Workers in Brazil in 1999</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Southeast Region</strong></td>
</tr>
<tr>
<td>Formal Sector</td>
</tr>
<tr>
<td>13,084,202</td>
</tr>
<tr>
<td>Informal</td>
</tr>
<tr>
<td>6,028,245</td>
</tr>
<tr>
<td>Indeterminate</td>
</tr>
<tr>
<td>7,387,019</td>
</tr>
<tr>
<td>Total</td>
</tr>
<tr>
<td>26,489,466</td>
</tr>
<tr>
<td><strong>Other Regions</strong></td>
</tr>
<tr>
<td>11,312,328</td>
</tr>
<tr>
<td>7,421,816</td>
</tr>
<tr>
<td>9,491,019</td>
</tr>
<tr>
<td>28,224,963</td>
</tr>
<tr>
<td><strong>Workers in RAIS</strong></td>
</tr>
<tr>
<td>15,856,851</td>
</tr>
<tr>
<td>Source: Cacciarmali e Braga [7] and RAIS</td>
</tr>
</tbody>
</table>

In our analysis, we use 11 years of data, covering the years from 1995 to 2005. Although data from previous years were available, we choose this time-frame because until 1994 Brazil suffered

\textsuperscript{16}The remaining variables are race, nationality, a measure of disability and the juridic nature of the firm.

\textsuperscript{17}It is worthwhile noting that Menezes, Muendler and Ramey [18], who also use RAIS, did not find significant differences in their wage regression results after controlling for selectivity into the formal sector.
from extremely high inflation, which caused serious measurement problems in variables such as wages, and also had structural implications for the macroeconomy.

The entire RAIS database contains over 400 million observations, which can lead to data manipulation problems.\footnote{This is just for the Southeast region of the country.} Studies that utilized analogous datasets in other countries used sampling techniques to deal with this problem. For example, AKM utilized data from 1/25 of the workers in France, and when using American data, they used data from 1/10th of the workers in the state of Washington. For the part of our study that requires longitudinal data we draw 1/10 of the workers in the formal sector in the Southeast region of Brazil. We collect this subsample drawing without replacement, from the pool of workers. So as not to lose information about the dynamics, we collect all observations from the selected workers. This procedure yields 22,133,168 observations.

Administrative datasets are sometimes known to have problems, such as errors in coding, missing observations and inconsistent information. In order to alleviate these problems, we perform the standard sample corrections adopted in the related literature. The procedure is described in the Appendix. After performing these steps we are left with 17,030,856 observations (77% of initial subsample). Table 2 shows some statistics from this sample.

<table>
<thead>
<tr>
<th>Table 2: Summary Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Longitudinal Sample: 10% of workers, 1995-2005 (18,117,645 obs)</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Total number</td>
</tr>
<tr>
<td>Avg. # obs</td>
</tr>
<tr>
<td>Avg. # of firms</td>
</tr>
<tr>
<td>Avg. # of years in sample</td>
</tr>
<tr>
<td>Education</td>
</tr>
<tr>
<td>Below High School</td>
</tr>
<tr>
<td>High School or College Dropout</td>
</tr>
<tr>
<td>College Graduate</td>
</tr>
<tr>
<td>Male</td>
</tr>
</tbody>
</table>

In our analysis we will also consider cross-sectional aspects of the data. For this part we utilize the full sample in year 2000.

### 4.2 Comparison With Other Datasets

Before proceeding, it is useful to compare the results of common procedures applied to other administrative datasets to those from the same procedures applied to our dataset. The most common exercise applied to a matched employer-employee datasets is the linear error decomposition
proposed by AKM. Their results motivated much of the subsequent empirical and theoretical literature in this field. Thus, it is to the very least informative to compare the results of applying this methodology in our data to the results found in other datasets.

This procedure consists of estimating the following wage equation.

\[ w_{it} = x_{it} \beta + \theta_i + \psi_{J(i,t)} + \epsilon_{it} \]  

(15)

where: \( w_{it} \) denotes the measure of earnings, \( x_{it} \) the vector of time-varying covariates, \( \theta \) the worker’s fixed effects and \( \psi_{J(i,t)} \) the fixed effect of firm \( J \), which employs worker \( i \) at time \( t \). Note that \( x_{it} \) only includes time-varying variables, because of the collinearity with the fixed effects.

In addition, to control for the observed time-fixed characteristics such as sex and education, they consider the additional auxiliary regression

\[ \theta_i = u_i \eta + \alpha_i \]  

(16)

where \( u_i \) are the time-fixed observables.\(^{19}\)

Following AKM, we use experience, time dummies and region dummies \( x_{it} \), and sex and education dummies for \( u_i \).\(^{20}\)

To estimate this model, a sample with a significant amount of mobility was necessary, because only workers who switch firms allow disentangling fixed effects of workers and firms. Another identifying assumption of this model is that the mobility patterns of workers are exogenous. It is clear that these are strong assumptions, and our model sheds light on them. However, the output from these regressions is still useful to summarize some aspects of the data.

The most informative aspects of this regression comes from the variances and correlations between each component. Table 3 compares the standard deviations and correlations of each component of this regression for France, the US and Brazil.\(^{21}\)

The table shows that the results utilizing the Brazilian data agree with the results found in the US and in France. In particular, two aspects are worth highlighting. First, the standard deviations of \( \alpha_i \) and \( \psi_{J(i,t)} \) are, respectively 0.49/0.46 in France, 0.81/0.35 in the US and 0.49/0.37 in Brazil. Thus, similar to other countries, \( \text{Var} \left( \psi_{J(i,t)} \right) \) in Brazil is relatively close to \( \text{Var} \left( \alpha_i \right) \). This result

---

\(^{19}\)Note that this step requires the assumption that \( \text{Corr} \left( u_i, \eta, \alpha_i \right) = 0 \).

\(^{20}\)We use the first four powers of experience as regressors. Experience is computed as potential experience for the first year and then actual experience for the subsequent years. We assign one dummy for each state of the southeast region.

\(^{21}\)The France and US information is reproduced from Abowd et al [1].
### Table 3

**A) AKM**

<table>
<thead>
<tr>
<th>Component</th>
<th>France</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log Real Wage</td>
<td>STD</td>
<td>STD</td>
</tr>
<tr>
<td>Time-Varying Characteristics</td>
<td>in w</td>
<td>in w</td>
</tr>
<tr>
<td>Person Effects</td>
<td>$\times \beta$</td>
<td>$\theta$</td>
</tr>
<tr>
<td>schooling and sex</td>
<td>$\delta$</td>
<td>$\omega$</td>
</tr>
<tr>
<td>Unobservable</td>
<td>$\alpha$</td>
<td>$\psi$</td>
</tr>
<tr>
<td>Firm Effect</td>
<td>$\epsilon$</td>
<td>$\epsilon$</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### France

<table>
<thead>
<tr>
<th>Component</th>
<th>STD</th>
<th>in w</th>
<th>$\times \beta$</th>
<th>$\theta$</th>
<th>$\omega$</th>
<th>$\alpha$</th>
<th>$\psi$</th>
<th>$\epsilon$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log Real Wage</td>
<td>0.977</td>
<td>1.000</td>
<td>0.538</td>
<td>0.457</td>
<td>0.151</td>
<td>0.432</td>
<td>0.429</td>
<td></td>
</tr>
<tr>
<td>Time-Varying Characteristics</td>
<td>0.409</td>
<td>0.538</td>
<td>1.000</td>
<td>0.070</td>
<td>-0.047</td>
<td>0.087</td>
<td>0.167</td>
<td>0.000</td>
</tr>
<tr>
<td>Person Effects</td>
<td>0.522</td>
<td>0.457</td>
<td>0.070</td>
<td>1.000</td>
<td>0.292</td>
<td>0.957</td>
<td>-0.223</td>
<td>0.000</td>
</tr>
<tr>
<td>schooling and sex</td>
<td>0.152</td>
<td>0.151</td>
<td>-0.047</td>
<td>0.292</td>
<td>1.000</td>
<td>0.000</td>
<td>0.029</td>
<td>0.000</td>
</tr>
<tr>
<td>Unobservable</td>
<td>0.499</td>
<td>0.432</td>
<td>0.087</td>
<td>0.957</td>
<td>0.000</td>
<td>1.000</td>
<td>-0.242</td>
<td>0.000</td>
</tr>
<tr>
<td>Firm Effect</td>
<td>0.467</td>
<td>0.429</td>
<td>0.167</td>
<td>-0.223</td>
<td>0.029</td>
<td>-0.242</td>
<td>1.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual</td>
<td>0.555</td>
<td>0.568</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

#### US

<table>
<thead>
<tr>
<th>Component</th>
<th>STD</th>
<th>in w</th>
<th>$\times \beta$</th>
<th>$\theta$</th>
<th>$\omega$</th>
<th>$\alpha$</th>
<th>$\psi$</th>
<th>$\epsilon$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log Real Wage</td>
<td>0.894</td>
<td>1.000</td>
<td>0.231</td>
<td>0.487</td>
<td>0.173</td>
<td>0.457</td>
<td>0.493</td>
<td>0.404</td>
</tr>
<tr>
<td>Time-Varying Characteristics</td>
<td>0.697</td>
<td>0.231</td>
<td>1.000</td>
<td>-0.609</td>
<td>-0.153</td>
<td>-0.589</td>
<td>0.064</td>
<td>0.000</td>
</tr>
<tr>
<td>Person Effects</td>
<td>0.843</td>
<td>0.487</td>
<td>-0.609</td>
<td>1.000</td>
<td>0.275</td>
<td>0.962</td>
<td>0.045</td>
<td>0.000</td>
</tr>
<tr>
<td>schooling and sex</td>
<td>0.232</td>
<td>0.173</td>
<td>-0.153</td>
<td>0.275</td>
<td>1.000</td>
<td>0.000</td>
<td>0.082</td>
<td>0.000</td>
</tr>
<tr>
<td>Unobservable</td>
<td>0.811</td>
<td>0.457</td>
<td>-0.589</td>
<td>0.962</td>
<td>0.000</td>
<td>1.000</td>
<td>0.023</td>
<td>0.000</td>
</tr>
<tr>
<td>Firm Effect</td>
<td>0.359</td>
<td>0.493</td>
<td>0.064</td>
<td>0.045</td>
<td>0.082</td>
<td>0.023</td>
<td>1.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual</td>
<td>0.361</td>
<td>0.404</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

**Sources:** DADS (1976-1996) for France, LEHD (1990-1999) for the US.

### B) Our Results

**Correlations of Components of Real Annual Wage Rates – Brazil**

<table>
<thead>
<tr>
<th>Component</th>
<th>STD</th>
<th>in w</th>
<th>$\times \beta$</th>
<th>$\theta$</th>
<th>$\omega$</th>
<th>$\alpha$</th>
<th>$\psi$</th>
<th>$\epsilon$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log Real Wage</td>
<td>0.800</td>
<td>1.000</td>
<td>0.205</td>
<td>0.533</td>
<td>0.439</td>
<td>0.313</td>
<td>0.604</td>
<td>0.330</td>
</tr>
<tr>
<td>Time-Varying Characteristics</td>
<td>0.520</td>
<td>0.205</td>
<td>1.000</td>
<td>-0.551</td>
<td>-0.413</td>
<td>-0.366</td>
<td>0.087</td>
<td>0.000</td>
</tr>
<tr>
<td>Person Effects</td>
<td>0.680</td>
<td>0.533</td>
<td>-0.551</td>
<td>1.000</td>
<td>0.710</td>
<td>0.704</td>
<td>0.109</td>
<td>0.000</td>
</tr>
<tr>
<td>schooling and sex</td>
<td>0.490</td>
<td>0.439</td>
<td>-0.413</td>
<td>0.710</td>
<td>1.000</td>
<td>0.000</td>
<td>0.185</td>
<td>0.000</td>
</tr>
<tr>
<td>Unobservable</td>
<td>0.480</td>
<td>0.331</td>
<td>-0.366</td>
<td>0.704</td>
<td>0.000</td>
<td>1.000</td>
<td>-0.032</td>
<td>0.000</td>
</tr>
<tr>
<td>Firm Effect</td>
<td>0.370</td>
<td>0.604</td>
<td>0.087</td>
<td>0.109</td>
<td>0.185</td>
<td>-0.032</td>
<td>1.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual</td>
<td>0.320</td>
<td>0.330</td>
<td>-0.041</td>
<td>-0.040</td>
<td>0.036</td>
<td>-0.094</td>
<td>-0.016</td>
<td>1.000</td>
</tr>
</tbody>
</table>
has been interpreted as being supportive of the existence of frictions in the labor market, because it suggests that workers with similar characteristics, even if they are unobservable, earn different wages in different firms. Second the correlation between \( \alpha_i \) and \( \psi_{J(i,t)} \) is \(-0.24\) in France, \(0.02\) in the US and \(-0.03\) in Brazil. This shows that our dataset yields the same result verified in several other datasets that this correlation is either zero or negative. This correlation, if taken at face value, can be interpreted as a measure of sorting in the labor market. This would suggest that, although there is substantial evidence that both the heterogeneity of workers and firms are important, there is no positive association between them.

Menezes, Muendler and Ramey [18], who also utilize RAIS, estimate wage regressions using a cross-section of workers, with establishment fixed effects only, and compare their results with results obtained from American and French data. They conclude that

Overall, our results establish a close similarity between wage structures in a major developing country and two major industrialized countries. Key differences between the countries, in particular the high variability of Brazilian wages, emerge from the way worker characteristics are compensated and not from differences in establishment-level wage policies.

### 4.3 Subsamples

To have a sample compatible with a possible realization of our theoretical model we have to make some adjustments to the sample. It is well known that observable aspects of the labor market such as region, gender and education are relevant for labor market outcomes. Since we do not explicitly model these, we deal with these types of heterogeneity by restricting our sample. The idea is to assume that there are separate labor markets. Thus, we select a sample of males living in the state of São Paulo, and we stratify this sample in two levels of education: low (high school or college dropout) and high (college degree).\(^{22}\) This combination of restrictions drops 88% of the longitudinal sample and 92% of the cross-sectional sample.\(^ {23}\) It is worth noting that each of the procedures below will be applied individually to each of this subsamples. For example, the average wage of a firm in the low education subsample will include only the low-educated workers.

\(^{22}\) This drops the workers with less than a high school degree.

\(^{23}\) The facts mentioned in this Section also hold when using more representative samples, with fewer restrictions.
4.4 Wages and Turnover - Descriptive Statistics

Before proceeding we provide some descriptive statistics on wages and turnover for our subsamples.

Figure 3 plots the cumulative distribution function for both the wage distribution and the firm-wage distribution. The plot shows the firm-wage distribution is more compressed than the wage distribution. Additionally, we include, in the graph, the distribution of wages among entrants, which we label "wage offers". This distribution is clearly stochastically dominated by the distribution of wages in the economy. This is consistent with results found in several datasets worldwide, and on-the-job search, a feature present in our model, has been shown to successfully explain it.\(^{24}\)

Table 4 provide statistics on worker turnover patterns. The number of job-to-job transitions relative to job-to-unemployment seems very low in this database.\(^{25}\) This number is comparable to the numbers found in Mediterranean countries such as Spain (25.4%), Italy (29.4%) and Portugal (36.9%).\(^{26}\) Finally, to check the hypothesis that the State of São Paulo is a "closed economy", we include migration numbers in the table. Out of the workers who began as employed in our sample, around 1% find a job outside of São Paulo in that same year. The number is smaller than 1% if we consider jobs outside of the Southeast region.

\(^{24}\text{See Mortensen [19].}\)
\(^{25}\text{We define a job-to-job transition if the worker shows in another firm within a month after being listed as being separated from his old job. This definition is similar to the one used in Jolivet, Postel-Vinay and Robin [12].}\)
\(^{26}\text{See Jolivet, Postel-Vinay and Robin [12].}\)
4.5 Patterns of Sorting in the Data

We look at two aspects of the data that are related to sorting. The first is standard in the literature, while the second is novel. Both can only be computed with a matched dataset.

First, we follow the standard empirical methodology, that uses wage fixed effects as proxies for the worker and firm heterogeneities. As shown above, this procedure consists in estimating the wage equations (15) and (16). In our case, since we are subsampling by sex and education, estimating Equation (16) is not required.

As previously mentioned, a very robust and surprising result in this literature is that the correlation between fixed effects is 0 or sometimes even negative, which is apparently at odds with the equilibrium of our model. However, non-monotonicities in Equation (6), discussed in Section 3, may distort the mapping between the firms’ productivities and the wage fixed effects. The same non-monotonicities do not affect the ordering of wages across workers. Moreover, in the model, workers of similar skill tend to work together as they sort in similar firms. This motivates us to use a new moment to measure of sorting: the correlation between $\theta_i$ and the average $\theta$ among his coworkers. Thus, we construct the measure

$$\theta^\text{coworkers}_{i,t} \equiv \frac{\sum_{k \in J_{i,t}} \theta_k}{N_{J_{i,t}} - 1},$$

where $J_{i,t}$ denotes the set of workers at firm $J$, who employs $i$ at $t$, minus worker $i$; $N_{J_{i,t}}$ is the number of coworkers. If $N_{J_{i,t}} = 1$, then $\theta^\text{coworkers}_{i,t} \equiv 0$\textsuperscript{27}

\textsuperscript{27}The mean of $\theta$ is normalized to zero in the AKM methodology.
Table 5 represents the results from this wage regression applied to our two subsamples.\textsuperscript{28} First, the $\text{Corr}(\theta_i, \psi_{J(i,t)})$ is slightly negative in both samples, which is not surprising and is consistent with the literature. Second, $\text{Corr}(\theta_i, \theta_{\text{coworkers}}) \approx 0.3$, which is much larger than any value previously found for $\text{Corr}(\theta_i, \psi_{J(i,t)})$ in the literature. If our model’s interpretation is correct, the second correlation indicates that workers and firms sort positively, and the first correlation is zero because of non-monotonicities in the wage equation.\textsuperscript{29}

<table>
<thead>
<tr>
<th>Table 5: Correlation between elements of Wage Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Low Education Sample</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Log Wages</td>
</tr>
<tr>
<td>STD</td>
</tr>
<tr>
<td>0.78</td>
</tr>
<tr>
<td>time-varying characteristics</td>
</tr>
<tr>
<td>STD</td>
</tr>
<tr>
<td>0.41</td>
</tr>
<tr>
<td>Worker Fixed Effects</td>
</tr>
<tr>
<td>STD</td>
</tr>
<tr>
<td>0.46</td>
</tr>
<tr>
<td>Mean Coworker Effect</td>
</tr>
<tr>
<td>STD</td>
</tr>
<tr>
<td>0.25</td>
</tr>
<tr>
<td>Firm fixed effects</td>
</tr>
<tr>
<td>STD</td>
</tr>
<tr>
<td>0.39</td>
</tr>
</tbody>
</table>

Ps: Sample has 1,564,817 obs, with 252,604 workers and 144,066 firms.

| **High Education Sample**                              |
|                                                        |
| Log Wages                                              |
| STD | $ln w$ | $x_{\beta}$ | $\theta$ | $\theta_{\text{coworkers}}$ | $\psi$ |
| 0.89 | 1.0    | 0.35        | 0.63     | 0.16                         | 0.49   |
| time-varying characteristics                           |
| STD | $ln w$ | $x_{\beta}$ | $\theta$ | $\theta_{\text{coworkers}}$ | $\psi$ |
| 0.26 | 0.35   | 1.0         | 0.11     | 0.03                         | -0.05  |
| Worker Fixed Effects                                   |
| STD | $ln w$ | $x_{\beta}$ | $\theta$ | $\theta_{\text{coworkers}}$ | $\psi$ |
| 0.56 | 0.63   | 0.11        | 1.0      | 0.3                          | -0.09  |
| Mean Coworker Effect                                   |
| STD | $ln w$ | $x_{\beta}$ | $\theta$ | $\theta_{\text{coworkers}}$ | $\psi$ |
| 0.3  | 0.16   | 0.03        | 0.3      | 1.0                          | -0.1   |
| Firm fixed effects                                     |
| STD | $ln w$ | $x_{\beta}$ | $\theta$ | $\theta_{\text{coworkers}}$ | $\psi$ |
| 0.53 | 0.49   | -0.05       | -0.09    | -0.1                         | 1.0    |

Ps: Sample has 495,159 obs, with 71,816 workers and 33,516 firms.

4.6 Robustness

Because our conclusions rely crucially on the empirical measures constructed in this section, we investigate further the role of observables, such as occupation and industry information, in explaining the observed patterns in the data. One example of how this may be relevant is related to occupations: suppose some firms have a higher concentration of blue-collar workers (e.g., janitors) and other firms have a higher concentration of white-collar workers (e.g., managers and CEOs). That could explain a high correlation amongst coworkers’ wages, as is observed in the data. Similar arguments could be made for industry and for locations. Here, we decompose the two correlations previously calculated within and between subgroups. We consider industry, occupations (3 digits), location (city) and the juridic nature of the firm as the possible subgroups. To perform this

\textsuperscript{28}Because the sample selection already controls for observables, the term $x_{it}$ includes only experience and time-dummies.

\textsuperscript{29}Bagger and Lentz [4] applied our method to the Danish matched employer-employee dataset and found a correlation close to 0.39.
decomposition, we use the standard formula

\[
\text{Var} \left( \begin{array}{c}
\alpha_i \\
\psi_{j(t)} \\
\alpha_{\text{coworkers}}^i 
\end{array} \right) = E \left( \begin{array}{c}
\text{Var} \left( \begin{array}{c}
\alpha_i \\
\psi_{j(t)} \\
\alpha_{\text{coworkers}}^i 
\end{array} \right) \right) + \text{Var} \left( E \left( \begin{array}{c}
\alpha_i \\
\psi_{j(t)} \\
\alpha_{\text{coworkers}}^i 
\end{array} \right) \right)
\]

where \( k \) indexes the particular subgroup. For this procedure, we calculate \( \alpha_{\text{coworkers}}^i \) only using the coworkers that belong to subgroup \( k \).\(^{30}\) Moreover, we use the full sample, and control for sex and education by using Equation (16), to guarantee a higher number of observations per subgroup.\(^{31}\)

The results are shown in Table 6. As we can see, even within industries, occupations and counties, the same pattern holds: slightly negative \( \text{Corr} (\alpha_i, \psi_{j(t)}) \) and \( \text{Corr} (\alpha_i, \alpha_{\text{coworkers}}^i) \) with a value around 0.3. Some examples within occupations are \( \text{Corr} (\alpha_i, \alpha_{\text{coworkers}}^i) \approx 0.39 \) for managers (27,974 obs), 0.35 for lawyers (12,924 obs) and 0.25 for waiters (202,929 obs). One thing to note is that \( \text{Corr} (\alpha_i, \alpha_{\text{coworkers}}^i) \) is close to 1 between subgroups.\(^{32}\) This is because, when we average \( \alpha_i \) and \( \alpha_{\text{coworkers}}^i \) within subgroups, that often includes most observations from the respective firms, generating identical variables.\(^{33}\) All this further justifies the use of a model based on unobservables to interpret the data.

<table>
<thead>
<tr>
<th>Industry</th>
<th>Occupations</th>
<th>Location</th>
<th>Juridic Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>-0.04</td>
<td>0.30</td>
<td>-0.04</td>
</tr>
<tr>
<td>Within</td>
<td>-0.06</td>
<td>0.28</td>
<td>-0.09</td>
</tr>
<tr>
<td>Between</td>
<td>0.36</td>
<td>1.00</td>
<td>0.45</td>
</tr>
</tbody>
</table>

5 Preliminary Results

We now present some preliminary results that illustrate the ability of the model to match the proposed moments.

\(^{30}\)For example, if \( k \) stands for the occupation “journalist”, \( \alpha_{\text{coworkers}}^i \) will only include other journalists at that same firm.

\(^{31}\)That is why we use \( \alpha_{\text{coworkers}} \) and not \( \theta_{\text{coworkers}} \) here.

\(^{32}\)Abowd, et al [1] perform this decomposition by industry and firm size and find that \( \text{Corr} (\alpha, \psi_{j(t)}) \) is close to zero or negative for the great majority of subgroups.

\(^{33}\)Note that \( \sum_{j,t} \frac{\alpha_{\text{coworkers}}^j}{N_j} = \sum_{j} \frac{N_j \alpha_j}{\sum_{j} N_j} \), where \( J_t \) stands for the set of workers at firm \( j \) at date \( t \), and \( N_j \) is the amount of those workers.
5.1 Calibration and Results

We calibrate the model model’s parameters to match some target moments. We choose the following configuration for our exercise:

- Distributions of types. First, for tractability we discretize the support of the type distributions using 100 points. We distribute the vector of probabilities of the (exogenous) worker and the job-type distributions proportionally to log-normal distributions. \( \log(h) \sim N(0, \sigma^h) \) for the workers and \( \log(p) \sim N(0, \sigma^p) \) for the firms.

- Transition parameters. First, we assume that \( \bar{N} = 1 \), which implies that \( v = u \). Also, we normalize the transition rates similarly as in the example of Section 3.

- Unemployment insurance: \( b(h) = 0, \forall h \).\(^{34}\)

To provide a model counterpart for the data, we provide a simulation of the model. This is completed by approximating the continuous time economy with the analogous discrete-time economy, taking the periodicity to be a week. We simulate the path of 100000 workers, whose initial labor market state is drawn from the steady state distribution, over 11 years.

Recall that our model is defined in terms of workers and jobs. However, when moving to the data, we need to introduce the concept of firms in this economy. We assume that firms only own jobs of one type, and are exogenously distributed according to the distribution \( \Upsilon(p) \), with density \( \upsilon(p) \). The mass of firms per worker is \( N_F \). Each firm owns \( \frac{\bar{N}_F g(p)}{N_F \upsilon(p)} \) jobs. In our empirical specification we take \( \Upsilon(p) \) to be Uniform over the support of \( p \). In practice, we draw 10.000 firms from the distribution \( \Upsilon(p) \), in order to have the ratio of workers per firm similar to as in the data. Thus, when a worker meets a firm, firms of type \( p \) have a chance \( \frac{\upsilon(p)}{\sum_{j=1}^{10000} \upsilon(p_j)} \) of being chosen.

Recall that the parameters of the model are \( \{r, \sigma^h, \sigma^p, \lambda, \kappa, \delta, \beta\} \). First, we set \( r \) to 0.49% per month.\(^{35}\) We then have three parameters related to turnover \( (\lambda, \kappa, \delta) \), two related to heterogeneity \( (\sigma^h, \sigma^p) \) and one related to rent sharing, \( \beta \). For this exercise, we choose to match three transition moments, one moment related to wage inequality and the two sorting moments derived from the AKM regressions. The outcome is described in Table 7.

Although we are not able to match all the targets the results show that the model performs well especially in explaining the sorting moments, which are the most important ones. The calibrated

\(^{34}\)It is well known that structural estimates of search models usually yield values for \( b \) smaller than zero (see Mortensen [19]). One interpretation for that is that \( b \) includes, in addition to unemployment benefits, non-pecuniary considerations.

\(^{35}\)The periodicity for all timed variables is monthly.
values are \( \{ \sigma^h, \sigma^p, \lambda, \kappa, \delta, \beta \} = \{ 0.44, 0.18, 0.16, 0.05, 0.009, 0.30 \} \). Interestingly, we found that the dispersion across firms was much smaller than the dispersion across workers (\( \sigma^h \approx 2\sigma^p \)), and found a higher bargaining power for the firms (\( \beta = 0.30 \)).

### Table 7 – Calibration of Model

<table>
<thead>
<tr>
<th></th>
<th>Target</th>
<th>Model Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sorting</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corr worker and firm fixed effects</td>
<td>-0.09</td>
<td>-0.03</td>
</tr>
<tr>
<td>Corr worker and coworker fixed effects</td>
<td>0.3</td>
<td>0.36</td>
</tr>
<tr>
<td><strong>Transition moments</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% of workers who separate after 1 year</td>
<td>18.18%</td>
<td>17.80%</td>
</tr>
<tr>
<td>% of separations that are job-to-job</td>
<td>26.07%</td>
<td>21.00%</td>
</tr>
<tr>
<td>% unemployment rate</td>
<td>8.00%</td>
<td>11.00%</td>
</tr>
</tbody>
</table>

| **Wage Inequality** |                 |                 |
| 90/10 percentile of log-wage distribution | 2.18            | 2.52            |

Table 8 shows, in further detail, the relationship between theoretical and empirical measures of sorting in the simulated data. The first panel shows the correlations between the primitives of the model in our simulation. It is clear that the equilibrium of this model displays positive assortative matching, in that the correlation between worker types \( h \) and job types \( p \) is 0.4. Also, because workers sort positively with firms they end up sorting positively with other workers as well. In this case, the correlation between \( h \) and the average value of \( h \) for the coworkers is 0.28. Things change when we make inference using the results of the linear decomposition with fixed effects applied to the simulated data. In that case, the correlation between the fixed effects, which would ideally capture \( \text{corr}(h, p) \), is \(-0.03\), which is much smaller than the correlation between the primitives. In our model, this is because of the downward distortion on wages introduced by the firm’s outside value, as previously discussed. However, as anticipated, this distortion does not happen when we look at the correlation between the fixed effects of the worker and his coworkers: this value is 0.36, which is fairly close to the value of \( \text{Corr}(h, p) \) and \( \text{Corr}(h, h_{other}) \). This ordering is consistent with the patterns observed in the data.

### 5.2 The Importance of the Model’s Assumptions

It is very useful to contrast our economy with the results derived from shutting down some channels of the model. This provides us insight on which elements of the model are important to explain the observed data, in particular the aspects of the data related to sorting. Our method is to shut down one element of the model at a time, and to re-calibrate the model. We choose the following specifications:
• No on-the-job search. This can be shut down by setting \((\kappa, \varphi) = 0\). On-the-job search can play a big role since endogenous mobility has been acknowledged to be a source of bias for the AKM regressions.\(^{36}\)

• Substitutability in production. We eliminate the complementarities in production by assuming the alternative production function \(F(h, p) = h + p\), which exhibits substitutability.

• No capacity constraints on the firm side. We eliminate these by allowing the firms to costlessly replace vacancies, as in Cahuc, Postel-Vinay and Robin [8] and Lentz [14]. In practice, workers contact firms directly (at the same rate as before), drawing the specific firm from the exogenous distribution \(\Upsilon(p)\). We assume \(\Upsilon(p)\) to be uniform, but a firm of type \(p\) has a contact rate \(\tilde{G}(p)\), distributed according to a log-normal distribution.\(^{37}\) This specification implies that \(J^U(p) = 0, \forall p\).

Table 8 – Model Simulations

A – Correlation Between Primitives

<table>
<thead>
<tr>
<th></th>
<th>(h)</th>
<th>(p)</th>
<th>(h_{coworkers})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(h)</td>
<td>1.00</td>
<td>0.40</td>
<td>0.28</td>
</tr>
<tr>
<td>(p)</td>
<td>0.40</td>
<td>1.00</td>
<td>0.63</td>
</tr>
<tr>
<td>(h_{coworkers})</td>
<td>0.28</td>
<td>0.63</td>
<td>1.00</td>
</tr>
</tbody>
</table>

B – Correlation Between Components of Wage Regression

<table>
<thead>
<tr>
<th></th>
<th>(\theta)</th>
<th>(\psi)</th>
<th>(\theta_{coworkers})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Worker fe</td>
<td>1.00</td>
<td>-0.03</td>
<td>0.36</td>
</tr>
<tr>
<td>Firm fe</td>
<td>-0.03</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Coworkers fe</td>
<td>0.36</td>
<td>0.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 9 – Alternative Specifications

<table>
<thead>
<tr>
<th>Primitives</th>
<th>Data</th>
<th>Full model</th>
<th>No OJS</th>
<th>Substitutability</th>
<th>No Capacity Constraints</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Corr(h, p))</td>
<td>-</td>
<td>0.40</td>
<td>0.30</td>
<td>-0.16</td>
<td>0.00</td>
</tr>
<tr>
<td>(Corr(h, h_{coworkers}))</td>
<td>-</td>
<td>0.28</td>
<td>0.18</td>
<td>0.04</td>
<td>0.00</td>
</tr>
<tr>
<td>Wage Regression</td>
<td>-</td>
<td>-0.09</td>
<td>-0.03</td>
<td>-0.08</td>
<td>-0.12</td>
</tr>
<tr>
<td>(Corr(\theta, \psi))</td>
<td>0.3</td>
<td>0.34</td>
<td>0.25</td>
<td>0.09</td>
<td>0.00</td>
</tr>
</tbody>
</table>

\(^{36}\)For example, Postel-Vinay and Robin [22] find that endogenous mobility induced by on-the-job search biases the relative variances of the worker and the firm components of the wage equation.

\(^{37}\)The corresponding Normal has mean zero and variance \(\sigma_p^2\).
The results are shown in Table 9. First, the model without on-the-job search does a good job explaining both sorting correlations jointly. This suggests that endogeneity in transitions is not the main problem with the AKM regressions when using them to infer sorting. Second, complementarities in production play a big role, as the model with substitutability fails to explain the sorting correlations: the signs are correct, but the magnitudes are far off. This model with no complementarities and search frictions exhibits negative sorting, as one can see from \( \text{Corr}(h, p) \approx -0.16 \). This is because the high-skill workers have more to lose from being idle, but the extra compensation gained from matching with a better firm is similar for all workers.\(^{38}\) In that case, the good workers work for the bad firms, and consequently have good coworkers, which explains \( \text{Corr}(\theta_{t, t}, \theta_{i, t}^{\text{coworkers}}) \approx 0.09 \). Finally, capacity constraints on the firm side also play a big role in explaining the sorting correlations. Shutting these down eliminates sorting from the model altogether, as one can see both from the primitives and from the components from the AKM regression.

5.3 Outline of the Model’s Estimation [in progress]

Our current research effort consists in expanding the calibration methodology to an estimation of the model using the Brazilian data. Since we do not have a closed form solution for this model, we use the simulated method of moments. This procedure involves, for each parameter configuration, solving numerically the equilibrium of the model, simulating a panel of workers, and calculating the distance between the data and the simulated moments. This is a very computationally demanding problem for two reasons. First, calculating the equilibrium of the model takes a non-negligible amount of time, and there are issues with the convergence of the algorithm for some parameter regions. Second, our sample design requires simulations of a very large size: for each parameter configuration, we simulate histories for hundreds of thousands of workers, weekly, for 11 years.\(^{39}\)

There are three types of moments that are relevant to our question. First, and foremost, for the sorting moments we use the previously described correlations from the AKM regressions. Second, for the transition moments, we follow the job search literature and use duration and turnover data. Finally, we can use the moments of the wage distribution and the firm-wage distribution to gather more information on the unobserved distributions.

The matched sample design of our dataset poses problems to compute standard deviations for our parameter estimates.\(^{40}\) First, recall that, for tractability, we only use 10% of the original sam-

\(^{38}\)Note that, in the case without on-the-job search, \( w(h, p) = \beta \left[ p - rJ^U(p) \right] + \beta h + (1 - \beta) \left[ rV^U(h) \right] \).

\(^{39}\)Recall that the model is in continuous time. In practice we need to simulate it in discrete time, and we choose to do it weekly.

\(^{40}\)This paragraph applies to any matched dataset, not only RAIS.
ple. If we had a standard panel that followed workers over time the natural way of computing standard deviations for the estimates would be to use a bootstrap procedure that draws with replacement from the pool of workers, gathering all observations from the selected workers. Here, this procedure does not work because there is dependence across workers, brought by the fact that they may work for similar firms. For example, suppose we draw a bootstrap sample as previously described. In several firms the same worker will be drawed more than one time. This implies an upward bias in $\text{Corr} \left( \alpha_i, \alpha_{i, \text{coworkers}} \right)$, simply because the group of coworkers includes clones of the worker. This bias is likely to cause problems when computing standard deviations of our parameter estimates because the mapping between the primitives of the model and the proposed moments is highly non-linear. The way we deal with this problem is to go back to the original sample and re-draw new 10% samples without replacement. Another advantage of this method is that we incorporate the effect of sampling error on our estimates.

6 Endogenous Vacancy Formation

In our formulation, we strictly follow Shimer and Smith [25]: the total stock of jobs of each type is assumed to be exogenous, and the process through which workers are allocated to these vacancies is endogenous. We do this because the fixed stock of jobs causes the vacant jobs to have a positive value. More specifically, higher productivity vacancies have higher value, and this allows us to explain the data better. However, a more common assumption for models that deal with firm heterogeneity is that the distribution of firms in the economy is exogenous and then proceed from there.\footnote{For example, Burdett and Mortensen [6], Postel-Vinay and Robin [22] and Lentz [14].} In this section, we change the assumptions of the model, so that it becomes consistent with the literature in a way that preserves the features that led to the novel results. The assumption is that there is an exogenous distribution of firms in the economy, each of which endogenously choose how many vacancies to post in each period.

We assume that there is a measure $N_F$ of firms in this economy, measured in terms of labor units. These firms’ productivities are exogenously distributed according to $\Upsilon(p)$, with density $\upsilon(p)$. Each firm endogenously posts new vacancies, $n$, in each period, and is subject to a sunk cost of creating new vacancies, $C(n)$. We assume $C(.)$ to be an increasing, twice differentiable and convex cost function. This convex sunk cost is what limits the firm from posting an infinite amount of vacancies and creates a meaningful outside option for each vacancy. In this case, if the firm fills that vacancy it has to post a new one incurring the sunk cost. We then assume that there is no flow cost of keeping vacancies open. This implies that after a vacancy is posted, its value is still given
by Equation (4). We also assume that once a worker leaves a productive match this is permanently destroyed and a new vacancy has to be posted. This implies that Equation (3) becomes

\[
 rJ^E (h, p) = F(h, p) - w(h, p) - \left[ \delta + \kappa \lambda W \int_{-\infty}^{\infty} \alpha^E(h, p, p^*) g(p^*) d p^* \right] J^E(h, p). 
\] (17)

Each firm is assumed to be small, therefore, they do not internalize the effects of their own postings on the equilibrium variables of the model. The firm’s decision of posting new vacancies each period can be summarized by the problem:

\[
 \max_n J^U(p) n - C(n) 
\]

with solution

\[
 n(p) \equiv \{ n : C'(n) = J^U(p) \}. 
\] (18)

Note that under our assumptions, all firms of a given type will make the same decision when posting new vacancies. In steady state, each firm will have a stable number of outstanding vacancies, \( v(p) \), in each period. Thus, the flow of new vacancies has to equal the flow of vacancies that find workers, either from unemployment or from other jobs:

\[
 0 = \dot{v}(p) = n(p) - v(p) \lambda^F \int_{-\infty}^{\infty} \alpha^U(h^*, p) h(h^*) dh^* \\
 0 = \dot{v}(p) = n(p) - v(p) \lambda^F \int_{-\infty}^{\infty} \alpha^U(h^*, p) h(h^*) dh^* - v(p) \phi \lambda^F \int_{-\infty}^{\infty} \alpha^E(h^*, p^*, p) \gamma(h^*, p^*) dh^* dp^*.
\]

Rearranging the equation, we obtain the steady state stock of vacancies of each firm:

\[
 v(p) = \frac{n(p)}{\lambda^F \int_{-\infty}^{\infty} \alpha^U(h^*, p) h(h^*) dh^* + \phi \lambda^F \int_{-\infty}^{\infty} \alpha^E(h^*, p^*, p) \gamma(h^*, p^*) dh^* dp^*}. 
\] (19)

Next, the individual decisions of each firm imply that the vacancy rate of the economy is

\[
 v = N^F \int_{-\infty}^{\infty} v(p^*) v(p^*) dp^*.
\] (20)

Accordingly, the distribution of productivities across vacancies, i.e., the sampling distribution of job types, is given by

\[
 g(p) \equiv \frac{v(p) v(p)}{\int v(p^*) v(p^*) dp^*}.
\] (21)
We define the equilibrium in this modified economy as

**Definition 2** An equilibrium in the economy with endogenous vacancy creation consists of values for \( \gamma(h, p) \), \( I(h) \), \( g(p) \), \( u \), \( v \), \( s^E(h, p) \), \( \alpha^U(h, p) \), \( \alpha^E(h, p, p') \), \( v(p) \) and \( n(p) \) such that Equations (7), (8), (9), (10), (11), (13), (18), (19), (20) and (21) are satisfied.

There are two things that we would like to emphasize about this formulation. First, one can see that if \( J_U^p(p) > 0 \) then \( n'(p) > 0 \), and \( v(p) \) is most likely increasing in \( p \).\(^{42}\) This has been shown to be an important feature to explain firm-size in these types of models. Moscarini and Postel-Vinay [21] argue that, although on-the-job search, in part, explains why more productive firms are bigger, it is necessary that \( v(p) > 0 \) in order to explain the US firm-size distribution. In their exercise, they introduce \( v(p) \) into the Burdett and Mortensen [6] model through exogenous sampling weights.\(^{43}\) Here, this increasing \( v(p) \) is determined endogenously. Second, if we pick \( \psi(p) \) and \( c(p) \) "correctly", then we can get a \( g(p) \) function similar to the one generated by the exogenous jobs case. In that case, all the results previously described extend to this economy with endogenous vacancy formation.

### 7 Conclusion

We offer four contributions to the measurement of sorting in the labor market.

First, we show that the standard methodology of measuring sorting in the labor market may be biased in favor of no sorting. More specifically, we use an equilibrium model of the labor market with heterogeneous workers and firms that exhibits positive sorting in equilibrium, because of the presence of complementarities in production. Despite the presence of strong sorting in the model’s equilibrium, for plausible parametrizations the standard estimator applied to simulated data suggests, once again, no sorting, just like in the empirical data. Our second contribution is to identify the source of this bias in non-monotonocities in the wage equation (when increasing firm’s productivity) caused by the interaction of wage bargaining with limited capacity of the firms to post new vacancies. In our model, high productivity firms have better outside options than their low-productivity counterparts, which causes downward pressure on the wages of their workers. This is particularly relevant for low-skilled workers, who can be paid less when working for a more

\(^{42}\)In our attempted simulations the numerator of Equation (19) outweighs the denominator.

\(^{43}\)In work in progress, they endogeneize \( v(p) \) using our formulation in this Section.
productive firm. All this suggests that the firm wage fixed effects are not good proxies for the firms productivities. Abowd et al [1] also show an example where a theoretical model with positive sorting can imply a zero/negative correlation in wage fixed effects. However, their example requires wages to be decreasing in worker skill, and a counterfactual negative correlation between employment probability and wages.

Thirdly, we propose a new empirical method to detect sorting that is immune from this bias: the correlation between a worker’s wage fixed effect and the average fixed effects of the co-workers at the same firm. We find that this method correctly detects sorting in the simulated data from the model. Finally, we apply both methods to a novel dataset: the Brazilian matched employer-employee dataset. We confirm the absence of sorting when using the first method, but the second method reveals strong sorting. According to the model, these two apparently contradictory empirical findings suggest that sorting is widespread in the labor market.

One drawback of our approach is that, although we have argued that complementarities are indeed important in production, we cannot establish if these are positive or negative. Recall that the positive complementarities of our model imply that "good" workers end up with other "good" workers as a byproduct of working for "good" firms. Another mechanism could be that these "good" workers are clustering in "bad" firms. In fact, Eeckhout and Kircher [9] use a related model to argue that from wage data alone one cannot distinguish positive from negative sorting. They also argue, however, that for questions related to efficiency and inequality the degree of sorting is more important than the sign. Our work contributes in that direction: in the light of the model, the correlation between the worker’s fixed effects provides an approximate measure of the degree of sorting in the labor market.

8 Appendix

8.1 Computing Steady State Flows

Here, we explain how to derive Equation [10], which balances the flows in and out of employed matches. Recall that the CDF of employed matches is \( \Gamma(h, p) \). The flows in and out of matches with workers up to skill level \( h \) and jobs with productivity up to \( p \) should equal in steady state. The flows out of these group come from matches that are hit by the exogenous shock and from workers...

\[ ^{44} \text{They use the direct search model of Shimer [23].} \]
who move to jobs with productivity higher than $p$

$$
\delta (1-u) \Gamma (h, p) + \kappa \lambda^W (1-u) \int_p^\infty \int_{-\infty}^{h} \alpha^E (h', p', p^*) \gamma (h', p') g (p^*) d p' d h' d p^*.
$$

The flows into this group comes from idle agents of those types who meet and from workers of type up to $h$ who move from a job with type higher than $p$ to a job less productive than $p$

$$
u \lambda^W \int_{-\infty}^{h} \int_{-\infty}^{p} l (h') g (p') d p' d h' + 
\kappa \lambda^W (1-u) \int_{-\infty}^{h} \int_{-\infty}^{p} \alpha^E (h', p', p^*) \gamma (h', p') g (p^*) d p' d h' d p^*.
$$

If we differentiate both sides with respect to $h$ and $p$ we obtain

$$
\delta (1-u) \gamma (h, p) + \kappa \lambda^W (1-u) \int_p^\infty \alpha^E (h, p, p^*) \gamma (h, p) g (p^*) d p^*
$$

$$
-\kappa \lambda^W (1-u) \int_p^p \alpha^E (h, p', p) \gamma (h, p') g (p) d p' = 
\nu \lambda^W l (h) g (p) + \kappa \lambda^W (1-u) \int_p^\infty \alpha^E (h, p', p) \gamma (h, p') g (p) d p'
$$

$$
-\kappa \lambda^W (1-u) \int_p^p \alpha^E (h, p, p^*) \gamma (h, p) g (p^*) d p^*.
$$

Finally, if we combine the terms with integrals we obtain Equation [10].

### 8.2 Properties of the Wage Function

First, we show that $w_h (h, p) > 0, \forall \{(h, p) | \alpha^U (h, p) = 1\},$ in the case without on-the-job search—in e., $(\psi, \varphi) = 0$. Assume that $b' (h) > 0, \forall h$, the production function is log-supermodular, and the primitives of the model are such that the value functions are differentiable.\(^{45}\) If we differentiate Equation (6) with respect to $h$ we obtain

$$
w_h (h, p) = \beta F_h (h, p) + r (1 - \beta) V^U_h (h).
$$

\(^{45}\)See Shimer and Smith [25] for details.
It remains to show that $V^U_h(h) > 0$, \forall h$. Differentiating Equation (2) w.r.t. $h$, and using Equation (5)

$$rV^U_h(h) = b'(h) + \beta \lambda^W \int_{-\infty}^{\infty} \alpha^U(h,p^*) \frac{F_h(h,p^*) - rV^U_h(h)}{r + \delta} g(p^*) dp^*.$$ 

In this step, it is not necessary to differentiate the bounds of the integral because the surplus is zero at those points. Rearranging the term with $V^U_h(h)$ to the LHS it is clear that $V^U_h(h) > 0$, \forall $h$.

Next, we show that $w_p(h,p) < 0, \exists \{ (h,p) | \alpha^U(h,p) = 1 \}$. We do so in the case where $F(h,p) = hp$, and $\min(h) = 0$. If we differentiate the wage Equation we obtain

$$w_p(h,p) = \beta F_p(h,p) - rJ^U_p (p).$$ 

Now, an analogous derivation to the previous one in this Section yields

$$rJ^U_p (p) = \frac{(1 - \beta) \lambda^W}{r + \delta + (1 - \beta) \lambda^W \int_{\infty}^{\infty} \alpha^U(h^*,p) l(h^*) dh^*} \int_{-\infty}^{\infty} \alpha^U(h^*,p) F_p(h,p^*) l(h^*) dh^* dh^*$$

. Using the functional forms

$$w_p(0,p) = 0 - \frac{(1 - \beta) \lambda^W}{r + \delta + (1 - \beta) \lambda^W \int_{\infty}^{\infty} \alpha^U(h^*,p) l(h^*) dh^*} \int_{-\infty}^{\infty} \alpha^U(h^*,p) h^* l(h^*) dh^*$$

$$< 0.$$ 

This last statement is true for any $p$ that employs a worker with $h > h_{min}$.

### 8.3 Sample Dispositions

Here, we describe the procedure utilized to perform some sample corrections to the sample. We follow the procedure outlined in AKM. The numbers below refer to the 10% 11-year sample, but this correction also applies to a cross-section when this is used—using all observations.

1. We start with 22,133,168 observations. Each observation corresponds to a worker-Establishment-year cell, possibly with more than observation per cell.

2. We remove from the sample part time and non-wage workers: weekly hours $< 40$ and worker earns less or equal to a minimum wage. This discards 3,258,299 observations.

3. Because of problems with the worker identifiers, we remove the observations of workers
who show only once in the dataset and appear as working during the beginning of the year or at the end of the year—i.e, workers who also worked in other years. This drops 330,455 observations.

4. Next, we aggregate the information to the firm level. For each worker-establishment-year, we aggregate the information from all establishments of a firm. As in AKM, we use as wage the average of all the observations from the worker in that firm on that year. This drops 535,956 observations.

5. Imputations: Because of problems with the identifiers, it is common in administrative data sets for a worker to show as working in a given firm, then disappear and then show in the same firm in the following year. As in AKM, whenever this happens, we input the middle observation, using the average of the previous and following wages. This adds 211,062 observations.

6. Multiple jobs: As in AKM, we discard the information from workers who show as working in more than three jobs in a given year. For this we remove the observations of workers who show as working more than 24 months in a given year (adding over all jobs of that year) or workers who appear in more than three firms at a given year. This discards 363,587 observations.

7. We remove from the sample workers with an age smaller than 15 or higher than 65. We also discard observations with problems in the sex, age and education identifiers. This drops 825,077 observations.

References


