Assortative Matching with Large Firms*

Jan Eeckhout† and Philipp Kircher‡

June 2016

Abstract

Technological progress allows firms to scale production processes over an increasingly large number of workers. This affects the size of the firm as well as the skill level of its workforce. We propose a unifying theory of production where management resolves a tradeoff between hiring more versus better workers. The span of control or size is therefore intimately intertwined with the sorting pattern. We provide a condition for sorting that captures this tradeoff between the quantity and quality of workers and that generalizes Becker's sorting condition. A system of differential equations determines the equilibrium allocation, the firm size and wages. We illustrate the theory using German matched employer data, and apply it to analyze quantity-biased technological change in conjunction with skill-biased technological change. We find that quantity-biased technological change is sizable and important. Moreover, it partially dampens the skill-premium, which would have increased even more. Skill-biased technological change is therefore even larger than the increase in the skill premium indicates.

Keywords. Sorting. Large Firms. Span of Control. Firm Size. Quantity-Biased Technological Change. Skill-Biased Technological Change. Complementarities.

*We would like to thank numerous colleagues and seminar audiences for insightful discussions and comments, and Diego Restuccia and Tasso Adamopoulos for generously sharing their data. Julia Faltermeier, Cristina Lafuente and David Pugh provided invaluable research assistance. Eeckhout gratefully acknowledges support from the ERC, Advanced grant 339186, and from ECO2012-36200, and Kircher from the ERC, grant 184119.

†University College London and Barcelona GSE-UPF, `j.eeckhout@ucl.ac.uk`.

‡Department of Economics, University of Edinburgh, `philipp.kircher@ed.ac.uk`.

1 Introduction

The firm size is an important aspect of a firm's productivity. Given wages, managers determine the optimal size of their workforce. This has far reaching aggregate implications. Researchers have used the firm size distribution to identify the role of market frictions and to explain how the misallocation of resources can lead to large differences in productivity across economies that have access to similar technologies.[1] At the same time, inputs in production are heterogeneous. Workers have different skills, managers and capital have different levels of productivity. This means that the sorting of workers into firms is an important determinant of the efficient production of output. What is missing in the literature is a tractable framework that allows for the standard firm size choice but also allows us to think about the sorting of heterogeneous workers into such large firms.[2] In this paper we take a first step towards proposing such a framework and show how the sorting of workers into firms interplays with the size of the firm. In particular, we show how the firm's management trades off the quality of its workforce against the quantity, and how wages and allocations are determined in equilibrium. A key insight is that the firm size distribution feeds back into the choice of labor inputs and affects the skill premium, as well as that the input heterogeneity feeds back into the firm size distribution.

We introduce a model of the firm where the span of control – the number of workers under the control of management within a firm – attributes an essential role to the firm. Just like in the canonical macroeconomic context, firms in our model predominantly make quantity decisions. Endowed with different management, technologies, or capital, companies choose the span of control accordingly, and this has important implications for the size of firms (Lucas (1978); Hopenhayn and Rogerson (1993)). This labor factor intensity decision is both realistic and a convenient modeling device. Yet, firms typically face a more complex tradeoff. They simultaneously choose the *quality* of the workers as well as the quantity. A retail arm of a company that sells electronics products for example faces the tradeoff between hiring skilled shop floor assistants who have extensive experience with a wide range of its products versus more unskilled assistants who can only be of help with the most basic features. Heterogeneity in skills and jobs is without doubt an important component of the labor market. Without the quantity dimension, the allocation process of differently skilled workers to jobs has extensively been analyzed, both with search frictions and without. In the standard frictionless matching model (Becker (1973)), each firm consists of exactly one job which leads to sorting since the firm's choice is in effect about which worker to hire, the *extensive margin*, rather than how many, the *intensive margin*.

By simultaneously solving the quantity and the quality dimension within the same model, we not

[1] See amongst many others Hsieh and Klenow (2009), Restuccia and Rogerson (2008) and Guner, Ventura, and Xu (2008).

[2] Of course the literature has proposed tractable models of sorting in settings with many-to-one matching, for example Sattinger (1993), or Garicano (2000)). These models have provided useful insights, deriving their tractability from the fact that the number of workers (i.e., the firm size) is fully determined once the manager has chosen the type of his workforce. Our framework instead endogenizes the size decision, which is standard in most macroeconomic environments (see for example Lucas (1978)). We discuss in Section 4 how existing models solve as special cases in our general framework, as well as the relation of our setup to the broader literature on many-to-one matching.

only nest other well known models of sorting and of firm size. Most importantly, we also analyze how the different technological determinants interact in general equilibrium with endogenous prices. Within this framework, we pin down the features of the equilibrium allocation: the sorting pattern, firm size distribution and the wage distribution. We find a surprisingly simple condition for assortative matching that captures both the quality and quantity considerations. This condition is new and compares the different degrees of complementarity[3] along four margins: (1) *type complementarity* captures the interaction between firm and worker types. Clearly, if better firms receive a exceptionally high return only from better workers, then they will end up hiring those workers. This is the only effect present in standard quality-sorting models in the spirit of Becker (1973)). Additionally, there is the (2) *complementarity in quantities* of workers and resources, just as in the standard model with quantity choices only. There is the (3) *span-of-control complementarity* between the firm or manager type and the number of workers that features in Lucas (1978); how much of a higher marginal product do better managers have from supervising more workers of a given skill? Finally, there is the (4) *managerial resource complementarity*, the complementarity between worker skills and managerial or firm resources: do better workers have a higher marginal product when receiving more supervision time? A simple tradeoff between these four forces determines the pattern of sorting. It characterizes the efficient equilibrium outcome and is a measure of the efficiency losses that would result from misallocation.

We also precisely pin down the composition of the workforce across different firm types, i.e., how firms resolve the tradeoff of span of control over more versus better workers. The equilibrium allocation of types and quantities is entirely governed by a system of three differential equations. In particular, this gives a prediction for the firm's span of control, and therefore, for the firm size distribution. It also determines the equilibrium allocation of skills and the wage distribution. This system also makes explicit how firm size interacts with the skill premium.

The combination of size and quality sorting allows us to study how changes in the size distribution affect wage inequality and the skill premium. We can also investigate how changes in the inequality of inputs affects the firm size distribution. Obviously, neither of these questions can be answered in models where all inputs are homogeneous or where all firms have equal size, as in most of the previous literature.

We illustrate our theory with an application to technological change. We ask how technology has changed over the last two decades and in which way technological change is driven by forces that affect the determinants of worker-firm complementarity, Skill Biased Technological Change (SBTC), as well as the determinants of Quantity Biased Technological Change (QBTC). Technological change drives economic progress, but it does not affect all factors of production equally. For example, the introduction of computers affects high skill workers differently than low skill workers. In light of this

[3]We will use the term complementarity and supermodularity interchangeably. For our purposes, it can best be thought of as the fact that the marginal contribution of higher input (quantity or quality) to output is higher when matched with other high inputs, i.e. there are synergies. In mathematical terms, the cross-partial of the output generated is positive (negative in the case of substitutes or submodularity).

biased technological change, relative prices adjust and resources are reallocated. A lot of focus has been on SBTC: the marginal product of the high skilled relative to the low skilled has increased substantially. This has strong implications for the wage premium, and therefore for wage inequality (see amongst many others Krusell, Ohanian, Ríos-Rull, and Violante (2000), Acemoglu (1998), and Acemoglu (2002)). But technological change also affects how firms organize their production and thereby how they determine their size. In this framework, we can analyze the quantity dimension of technological change or QBTC. Technologies such as scanning devices and GPS emitters enable management to supervise a larger *number* of workers simultaneously. For example, if the chief operating officer, say at FedEx, streamlines the production processes, then this affects thousands of workers. Technological change is quantity biased if already large firms grow relatively more than small firms. If such firms employ mostly high skilled workers, the size expansion might reduce their marginal product, counteracting other drivers of technological change in the wage premium. The wage premium would then understate the importance of these other drivers in terms of technological development.

Using matched employer-employee data for Germany, we find that two technological determinants have changed dramatically: the marginal product of skilled labor, and the span of control of firms. Instead, the marginal productivity of firms, the elasticity of substitution between skills and firm productivity as well as the skill and productivity distributions have not or barely evolved. The estimates of the technology indeed show that SBTC has a sizable effects on the skill premium, but that some of this is mitigated due to QBTC. In its absence, the increase in wage inequality would have been even bigger. Likewise, the effect of QBTC on the firm size distribution would have been much larger if it were not because of the increase in the skill premium which reduces the impact on firm size due to higher wages. Not only does this exercise uncover novel aspects of technological determinants and their evolution, it also highlights the important role of the equilibrium interplay of sorting and firm size.

This example highlights how changes in the firm size distribution feed back into the inequality of payoffs amongst heterogeneous production inputs. But there is also a feedback channel from input heterogeneity to the firm size distribution. For some research questions, the latter might be even more central, for example in the misallocation debate as we show in section 4. One strand concerns the misallocation debate which is based on the observation that there are too many small firms in developing countries relative to developed countries, discussed for example in Hsieh and Klenow (2009), Restuccia and Rogerson (2008), Guner, Ventura, and Xu (2008), and Adamopoulos and Restuccia (2014). This literature either allows no input heterogeneity or assumes efficiency units of labor. We briefly showcase how our setup allows the introduction of input heterogeneity into this debate, e.g., such that better inputs (either land or labor) facilitate the use of capital. We illustrate how such input heterogeneity can affect the firm size distribution, and outline how our framework might be useful to think about the role of input heterogeneity within the misallocation debate.

In addition to a discussion on the misallocation debate, Section 4 is devoted to how the relevant literature relates to the mathematics of our model. We have chosen to give detailed credit to the related literature only after we have introduced our model because it allows us to we combine the discussion of related papers with a some simple analytical arguments how our model nests a number of well known matching models as special cases. Most notably, we show that Becker's one-to-one matching model is the limit case of a multiplicatively separable version of ours where the quantities enter as a Constant Elasticity of Substitution (CES) technology that converges to Leontief, i.e., with an elasticity of substitution equal to zero. We show similar connections to the influential papers by Sattinger (1975) and Garicano (2000) who embed specific forms in which firm size depends on worker and firm types but do not allow this to be a choice variable. We also discuss other related work from various literatures. We further show how our framework lends itself to introducing search frictions. To our knowledge, this extension with search frictions provides the first model that combines three essential features of labor market data: two-sided heterogeneity with complementarities, unemployment due to search frictions, and large firms. Existing models have combined two of those three, but not all three at the same time. Most surprisingly, we find that in this model the condition for assortative matching is independent of the matching technology and thus holds even if we move away from a Walrasian setting.

The paper is organized as follows. In the next section we lay out the model. In Section 3 we first solve the model and derive the general sorting condition, and then we characterize the equilibrium assignment, the firm size distribution and the wage profile. We then discuss in Section 4 the special cases that are nested in the model, we briefly refer to the extensions in the Appendix, and we review the related literature. Section 5 elaborates on the application of the theory: Quantity Biased and Skill Biased Technological Change. We conclude in Section 6.

2 The Model

We consider a static assignment problem in the tradition of Monge-Kantorovich, except that the allocation is not limited to one-to-one matching. To preview the basic economic situation that the model intends to capture, we consider an economy with two sides, which we mostly label as firms and workers, even though the labels managers/workers, farms/land, and capital/labor would be equally appropriate. There is heterogeneity on both sides: workers differ by skills, and firms are heterogeneous in terms of the quality of some proprietary resource that is exclusive to the firm, such as scarce managerial talent or particular proprietary capital goods. These scarce internal resources limit the scope of the firm. In a modern business setting, the resource might reflect the time endowment of an entrepreneur that she spends interacting with and supervising her employees, and quality can refer to the value of the final output or the ability during such supervision. If she supervises different workers, she might adjust supervision time to suit each worker's skill. Output depends on the type of worker and of the supervisor and the time they interact. The setup is formalized as follows.

Agents. The economy consists of firms and workers. Workers are indexed by their skill $x \in \mathcal{X} = \mathbb{R}_+$, and $H^w(x)$ denotes the measure of workers with skills below x. Firms are indexed by their productivity type $y \in \mathcal{Y} = \mathbb{R}_+$, where $H^f(y)$ denotes the measure of firms with type below y. Unless otherwise stated, we focus on distributions H^f and H^w with non-zero continuous densities h^f and h^w on the compact subsets $[\underline{x}, \overline{x}] \subset \mathcal{X}$ and $[\underline{y}, \overline{y}] \subset \mathcal{Y}$, respectively, but especially for our main characterization result we also provide a proof for arbitrary distribution functions.

Preferences and Production. The main primitive of our model is the output function $F : \mathbb{R}^4_+ \to \mathbb{R}_{++}$ that describes how the firm combines labor and its resources to produce output. Output is perfectly transferable, and firms maximize profits while workers maximize wage income. A firm has a fixed amount of proprietary resources. If a firm of type y hires an amount of labor l of type x, it has to choose a fraction of its resources r that it dedicates to this worker type. This allows the firm y to produce output

$$F(x, y, l, r) \tag{1}$$

with this worker type x, where the first two arguments (x, y) are *quality variables* describing the worker and firm types while the latter two arguments (l, r) are *quantity variables* describing the level of inputs. We assume that output is twice differentiable, but place no further restrictions on the quantity variables, even though we often refer to higher types as "better" types which is more appropriate for output functions that are increasing in types. Our main assumptions on the production functions concern the quantity variables. For technical reason we assume that it is strictly concave in each quantity variable in the interior of the type space, no output is produced without resources, and standard Inada conditions apply.[4]

Of economic relevance is the assumption that production displays constant returns to scale in the quantity variables. For example, if the output of each worker depends only on his own type x, the type of the firm y, and how many resources the worker receives, then constant returns to scale arise as twice the workers produce twice the output if the resources per worker stay constant. Constant returns implies that the output in (1) can be expressed as the product of the amount of resource r and the output per unit of resource:[5]

$$f(x, y, \theta) := F(x, y, \theta, 1) \tag{2}$$

where $\theta = l/r$ represents the amount of workers per unit of resource, which we often call the *intensity*. Importantly for the later analysis, function $f(x, y, \theta)$ also represents the production of a firm that only hires one type of worker, in which case θ represents the *firm's size*. Because of the tight link between f and F in (2), either can be used as the primitive of the model.

The action of a firm y is to choose two distributions, the number of workers of each type and the

[4]The requirement that $F(x, y, l, 0) = 0$ is made for convenience as it rules out that workers are hired by firms that devote no resources to them. This is only weakly concave in l, and therefore we can only assume strict concavity in the interior. Finally, Inada conditions on labor are $\lim_{l \to 0} F_r(x, y, l, r) = \infty$ for given $x, y, r > 0$, and $\lim_{l \to \infty} F_r(x, y, l, r) = 0$. Similar conditions can be placed on resources.

[5]If total output $F(x, y, l, r)$ has constant returns to scale, we can write it as $F(x, y, l, r) = rF(x, y, l/r, 1) = rf(x, y, \theta)$.

amount of resources devoted to them. Let labor demand $\mathcal{L}^y(x)$ denote the cumulative distribution of the number of workers that firm y hires of type x or lower, and let resource allocation $\mathcal{R}^y(x)$ denote the cumulative distribution of the resources that the firm dedicates to all workers of type x or lower. There is no loss to the assumption that firms hire workers only if they devote resources to them, as workers without resources produce no output (formally this means that labor demand is absolutely continuous in the resource allocation). The choices of $\mathcal{L}^y(x)$ and $\mathcal{R}^y(x)$ then determine the number of workers per unit of resources $\theta^y(x)$ relevant in (2) through the Radon-Nikodym derivative $\theta^y(x) := d\mathcal{L}^y(x)/d\mathcal{R}^y(x)$ almost everywhere. Conversely, the number of workers per resource $\theta^y(x)$ and the allocation of resources $\mathcal{R}^y(x)$ fully summarize the firm's labor demand as the sum of workers-per-resource over all resources: $\mathcal{L}^y(x) = \int_{\underline{x}}^x \theta(\tilde{x}) d\mathcal{R}^y(\tilde{x})$.[6] We can therefore interchangably use ($\mathcal{L}^y(\cdot)$,$\mathcal{R}^y(\cdot)$) and $(\theta^y(\cdot), \mathcal{R}^y(\cdot))$ to represent the firm's choices.

When a firm hires workers of multiple types we assume that its total output is the sum of the outputs across all its types. Additive separability again arises naturally if the output of each worker depends only on his and the firm's types and on the amount of resources available to him. Such formulations allows for interactions between firm and worker type, but abstracts from interactions amongst workers except through the limited resources. This abstraction is restrictive, but implies existence and – more importantly – tractability for the analysis of all the other cross-complementarities between quantities and qualities. Since $F(x, y, l, r) = rf(x, y, \theta)$ is the output of one worker type, the sum across all worker types can formally be represented as $\int f(x, y, \theta^y(x))d\mathcal{R}^y(x)$ where, as mentioned above, $d\mathcal{R}^y(x)$ represents how the firm allocates resources across different worker types.

Competitive Market Equilibrium. We consider a competitive equilibrium where firms can hire a worker of type x at wage $w(x)$. In equilibrium, firms' hiring decisions must be optimal and markets for each worker type must clear.[7]

Profit maximization of a firm of type y entails a choice of a production plan that maximizes output minus wage costs. For resource devoted to workers of type x at intensity θ the output is $f(x, y, \theta)$ but the firm has to pay the wage $w(x)$ to each of the θ workers that produce with this resource. The optimal production strategy therefore solves:

$$\max_{\theta^y, \mathcal{R}^y} \int [f(x, y, \theta^y(x)) - w(x)\theta^y(x)] d\mathcal{R}^y(x). \tag{3}$$

The firm's total wage bill $\int w(x)\theta d\mathcal{R}^y$ consists of the wage $w(x)$ integrated over its labor demand $\mathcal{L}^y(x) = \int_{\underline{x}}^x \theta(\tilde{x}) d\mathcal{R}^y(\tilde{x})$. For later reference it is useful to note that a firm that only hires one worker type x has a workforce size of $l(y) = \theta^y(x)$.

[6]Since $\theta^y(x) = d\mathcal{L}^y(x)/d\mathcal{R}^y(x)$, one can use $\theta^y(x)$ to reconstruct labor demand as: $\mathcal{L}^y(x) = \int_{(x,\theta):x\leq\tilde{x}} \theta(\tilde{x}) d\mathcal{R}^y(\tilde{x})$.

[7]We require wages to be non-negative in order not to violate the workers' outside option, which is normalized to zero for all agents. Firms can achieve their outside option simply by hiring no workers. We will call worker types with a zero wage and firm types with zero profits as inactive, while all other agents are called active.

Feasibility of the total allocation of resources requires that firms attempt to hire no more workers than there are in the population. Consider an interval of worker types $(x', x]$. A firm of type y has a demand for such workers of $\mathcal{L}^y(x) - \mathcal{L}^y(x')$. Integrated over all firms this yields the aggregate demand for such worker types. Therefore, labor demand schedules $\mathcal{L} = \{\mathcal{L}^y\}_{y \in \mathcal{Y}}$ are feasible if for any such interval of worker types the implied aggregate demand does not exceed the economy's endowment with such worker types:

$$\int_y \left[\mathcal{L}^y(x) - \mathcal{L}^y(x')\right] dH^f \leq H^w(x) - H^w(x'). \tag{4}$$

We can now define an equilibrium as follows:

Definition 1 *An equilibrium is a tuple functions* $(w, \theta^y, \mathcal{R}^y, \mathcal{L}^y)$ *consisting of a non-negative wage schedule* $w(x)$ *as well as intensity functions* $\theta^y(x)$ *and resource allocations* $\mathcal{R}^y(x)$ *with associated feasible labor demands* $\mathcal{L}^y(x)$ *such that*

1. *Optimality: For any* y *the combination* $(\theta^y, \mathcal{R}^y)$ *solves (3).*

2. *Market Clearing:* (4) *holds with equality if wages are strictly positive a.e. on* $(x', x]$.

The market clearing condition simply states that if wages for some worker types are positive, their markets clear. A useful feature of our setup is that firm's preferences over workers are convex, as shown in the appendix, so that we can draw on classical results on existence and welfare theorems in, e.g., Ostroy (1984) and Khan and Yannelis (1991). Our main focus here is on characterization: When do better firms hire better workers? How are the wages determined? When do better firms employ more employees? How is that effected by quantity-biased technological change?

Assortative Matching. Our focus is on labor demands that are monotonic in x and y. There is positive assortative matching (PAM) if higher firm types employ higher worker types in their production, i.e., for almost all firm types y and y' with $y > y'$ it holds that x is in the support of $\mathcal{L}^y$ and x' is in the support of $\mathcal{L}^{y'}$ only if $x \geq x'$. Negative assortative matching (NAM) can be defined by reversing the last inequality, capturing that lower type workers are employed in higher type firms. This definition is suitable in the presence of mass points in the type distributions.

A more natural and more tractable formulation of assortative matching arises if higher types produce strictly more output and the type distributions have non-zero continuous densities. We will focus on this case for expositional convenience, but our main sorting result in the Proposition 1 holds without these restrictions. With these restrictions higher types are more valuable and therefore there exist boundary types $\hat{x}$ and $\hat{y}$ such that all higher types are active. Assume that almost all active firm types y hire exactly one worker type $\nu(y)$ and reach size $l(y)$. We prove in the Appendix (Lemma 3) that this has to hold if there is assortative matching. An equivalent but simpler notion of assortative matching is therefore that $\nu(y)$ exists and is strictly monotone for almost all active types.

Traditionally models are solved from the perspective of the workers, for which the above discussions imply that for almost all active types x we can define the inverse $\mu = \nu^{-1}$ so that we can interpret $\mu(x)$ as the firm type that hires worker type x. The intensity for this worker is the *worker intensity* $\theta(x) := \theta^{\mu(x)}(x)$. Clearly μ inherits the strict monotonicity of ν, and as mentioned earlier, intensity equals firm size so that $\theta(x) = l(\mu(x))$. The market clearing condition now becomes particularly tractable. For the case of PAM, for example, it reduces to

$$\int_{\mu(x)}^{\overline{y}} \theta(s) h^f(s) ds = \int_x^{\overline{x}} h^w(s) ds \tag{5}$$

where right hand size sums up all workers above x and the left hand side sums over all firms that hire these workers times the number of workers each hires. In the case of one-to-one matching as in Becker, $\theta = 1$ and therefore $\int_x^{\overline{x}} h^w(s)ds = \int_{\mu(x)}^{\overline{y}} h^f(s)ds$ implies $H^w(x) = H^f(\mu(x))$. With firm size θ, this now means that we are matching one firm to θ workers.[8]

3 The Main Results

Models of assortative matching are in general difficult to characterize completely. Therefore, the literature has tried to identify conditions under which sorting is assortative. These conditions help our understanding of the underlying driving sources of sorting. And if the appropriate conditions are fulfilled, they substantially reduce the complexity of the assignment problem and allow further characterization of the equilibrium. In this section we first derive necessary and sufficient conditions for assortative matching, and then we characterize the assortative equilibrium allocation.

3.1 Assortative Matching

Our main result on sorting provides a necessary and sufficient condition that applies to arbitrary type distributions, and that places no restrictions on how types influence output. To build up intuition, though, it will be convenient to focus on necessary conditions for assortative matching in the case discussed at the end of the previous section: higher types produce more output and distributions have non-zero continuous densities. As outlined earlier, in an assortative equilibrium we can define for almost all active worker types the function $\mu(x)$ that denotes the firm type that hires worker x. Employment is at intensity $\theta(x) = \theta^{\mu(x)}(x) > 0$ at the equilibrium wage $w(x) > 0$. The strict inequalities arise because otherwise either worker or firm payoff would be zero, which would violate that these types are active. In an equilibrium with positive sorting $\mu(x)$ is strictly increasing. When output is increasing in x, also $w(x)$ is increasing as better types necessarily earn higher wages. Monotone functions are differentiable

[8] Like in our model, the mechanical relation that pins down matching in Becker (1973) no longer holds even in the one-to-one matching model when types are multi dimensional. See Lindenlaub (2016).

almost everywhere. Therefore, for almost any active x there exists an open neighborhood in which the following arguments based on differentiability are valid.

For this to be an equilibrium outcome, the firms' choices have to maximize their optimization problem (3). The next Lemma establishes that we can focus on a simplified problem.

Lemma 1 *Consider an active firm with strategy $(\theta^y, \mathcal{R}^y)$ that maximizes (3). Almost everywhere in the support of $\mathcal{R}^y$ it has to hold that $(x, \theta^y(x))$ solves*

$$\max_{\tilde{x},\tilde{\theta}} f(\tilde{x}, y, \tilde{\theta}) - \tilde{\theta} w(\tilde{x}). \tag{6}$$

Proof. Proceed by contradiction. Assume a positive measure of resources is placed by active firm y on a set of worker types $\tilde{\mathcal{X}}$ such that $(x, \theta^y(x))$ does not solve (3). Let (x^*, θ^*) be an optimizer of (3). Firm profits can be decomposed into the sum of $\int_{x\in\mathcal{X}\setminus\tilde{\mathcal{X}}}[f(x, y, \theta^y(x)) - w(x)\theta^y(x)]d\mathcal{R}^y(x)$ and $\int_{x\in\tilde{\mathcal{X}}}[f(x, y, \theta^y(x)) - w(x)\theta^y(x)]d\mathcal{R}^y(x)$ where the first term captures the profits with worker types in $\mathcal{X} \setminus \tilde{\mathcal{X}}$ and the second term captures the profits with worker types in $\tilde{\mathcal{X}}$. Placing all resources that the firm places on types in $\tilde{\mathcal{X}}$ instead on type x^* and choosing an intensity at x^* of θ^* leaves the first term unchanged but changes the second term to $\int_{x\in\tilde{\mathcal{X}}}[f(x^*, y, \theta^*) - w(x^*)\theta^*]d\mathcal{R}^y(x)$, which strictly improves profits since the integrant has strictly increased. ■

This Lemma states that firms do not choose worker type and intensity unless the combination maximizes the return per unit of resource. It implies that firms with a unique optimizer for (6) hire only one worker type. Optimality requires that the choices solve the first order conditions with respect to x and θ :

$$f_\theta(x, \mu(x), \theta(x)) - w(x) = 0, \tag{7}$$

$$f_x(x, \mu(x), \theta(x)) - \theta(x)w'(x) = 0, \tag{8}$$

where functions with lower case letters denote partial derivatives (e.g., $f_x = \partial f/\partial x$). Note that these equalities hold within the neighborhood around x. The implicit function theorem applied to (7) establishes that $\theta(x)$ is locally differentiable, and then the implicit function theorem applied to (8) implies that $w'(x)$ is once more locally differentiable. A necessary condition for optimality of the first order conditions is that the Hessian is positive definite, and in particular that its determinant is positive:

$$f_{\theta\theta}\left[f_{xx} - \theta w''(x)\right] - \left[f_{x\theta} - w'(x)\right]^2 \geq 0,$$

where the argument $(x, \mu(x), \theta(x))$ of f and its derivatives is suppressed for notational convenience. While this still entails the endogenous wage schedule, one can differentiate the first order conditions along the equilibrium path and use this to substitute out the wage schedule to obtain equivalently (see appendix for the derivation):

$$\mu'(x)\left[f_{xy}-\frac{f_{y\theta}(\frac{f_x}{\theta}-f_{x\theta})}{f_{\theta\theta}}\right]\geq 0.$$

Since PAM requires $\mu'(x) > 0$, a necessary condition is that the square bracket is weakly positive. This places restrictions on the production technology f of firms with only one worker type. The term f_{xy} is familiar from one-to-one matching models and – if positive – captures that higher firm types value higher worker types more. This is not enough to ensure PAM. It also matters to which extent higher types value the size of the firm. Intuitively, if higher type firms get high value out of being large but higher worker types are most productive in small firms, then this counteracts the familiar force. This can be seen even easier when using (2) to express this in terms of the original production function F. The next proposition makes this point, states this as necessary as well as sufficient condition, and dispenses with assumptions on the type distribution nor requires output to increase in types:

Proposition 1 *A necessary condition to have equilibria with positive assortative matching under any arbitrary distribution of types is that the following inequality holds:*

$$F_{xy}\geq\frac{F_{yl}F_{xr}}{F_{lr}} \tag{9}$$

for all $(x, y, l, r) \in \mathbb{R}^4_{++}$. *With a strict inequality, it is also sufficient to ensure that any equilibrium entails positive assortative matching. The opposite inequality provides a necessary and sufficient condition for negative assortative matching.*

Proof. In Appendix. ■

The proof has to deal with possible mass points in the type distributions which can lead to multiple firm types choosing a given work type in equilibrium. More importantly, the argument above only shows that the derivative of the matching function $\mu(x)$ has to be positive when (9) holds wherever this derivative is defined. In a positive assortative equilibrium this derivative is defined almost everywhere and (9) is necessary at these points. To ensure this under all type distributions, we show in the Appendix that (9) is necessary everywhere. But this does not prove sufficiency since it does not rule out that the matching can have a discontinuity with a discrete jump downward, which requires a global rather than a local argument. In the Appendix we deal with these issues by exploiting the implication of the First Welfare Theorem that any equilibrium allocation maximizes output in this economy with quasi-linear utility. If (9) fails but allocations have mass around points that are positive assortative, there are strict efficiency gains from re-arranging production in a negative assortative way. If (9) holds strictly but mass is placed around negative sorting, efficiency can be improved by re-arranging production in a positive assortative way. These properties are easy to show in one-to-one matching models where production always requires $r = l = 1$ and the local PAM requirement of $F_{xy} > 0$ can easily be integrated out to yield the gobal implication that $F(x_h, y_h) + F(x_l, y_l) > F(x_h, y_l, 1, 1) + F(x_l, y_h, 1, 1)$ for any $x_h > x_l$ and $y_h > y_l$, meaning that output increases when types are matched positively assorted. When the quantity dimension is active and r can differ from l, the new sorting condition (9) is more involved and cannot simply be integrated, requiring a substantially more involved argument despite the similarity in spirit.

Interpretation: Condition (9) embodies the quantity-quality trade-off that the firm makes, and this is captured by all four possible combinations of pairwise complementarities: one within qualities, two across quality-quantity dimensions, and one within quantities. Observe that, as in the one-to-one matching model (Becker (1973)), both positive and negative assorted allocations constitute an equilibrium if the condition holds with equality. Hence, the condition is only sufficient when it holds strictly.

On the left-hand side, a large value of the cross-partial on the quality dimensions (F_{xy}) captures strong *type complementarity* and means that higher firm types have ceteris paribus a higher marginal return for matching with higher worker types. The two terms in the numerator on the right-hand side represent the complementary interaction across qualities and quantities. The cross-partial F_{yl} captures the *span-of-control complementarity.* If it is large, it means that higher firm types have a higher marginal valuation for the quantity of workers. That is, better firms value the number of "bodies" that work for them especially high. In this case better firms would like to employ many workers. The *managerial resource complementarity* F_{xr} expresses how the marginal product of managerial time varies across better workers. If managerial time is particularly productive when spent with high skilled types, then it is positive and large. This would be the case for example if the learning by high types is faster.[9] Noteably, if better firms particularly value large firms but better workers particularly excel with lots of resources (implying few co-workers and therefore small firms) this creates a tension that counteracts positive assortative matching. The term in the denominator captures the *complementarity in quantities*, and acts mostly as a normalization since only its magnitude varies but its sign is always strictly positive due to constant returns to scale in quantities. Since it is tightly linked to the concavity in labor holding resources fixed, it captures the extent to which additional labor decreases the value of output.[10] The overall condition (9) can interpreted like the Spence-Mirrlees single crossing condition, adjusted for the additional complication that there are three goods that firms care about: the number of workers, the type of worker, and the numeraire.[11]

[9]This type of complementarity is often discussed in the context of teaching in the classroom. If a low-ability student reaches his limits earlier than a high-ability student, then additional instructor time might be more worth-while when it is devoted to the high-ability student ($F_{xr} > 0$). If high-ability students do well without further input while low-ability students crucially need the instructors time, then additional time by the instructor might be more worthwhile with the low-ability students ($F_{xr} < 0$). Clearly, in this context the output measure is not as clear as in a production setting, and considerations of fairness and equity play an additional role.

[10]Due to constant returns to scale $F_{rl}(x, y, l, r) = -F_{ll}(x, y, l, r)l/r$.

[11]In a standard Spence-Mirlees analysis, agents care only about two dimensions. For example, think about an alternative model in which agents of type y maximize $f(x, y, \theta)$ and have a budget set M and feasible (x, θ)-combinations that only

The condition for positive assortative matching then compares the within-complementarities with the across-complementarities. In the absence of a quantity dimension the right-hand side of the inequality is zero. With a quantity dimension, the requirements for positive assortative matching now depend on how much substitutability there is of quality for quantity, i.e., the ability to substitute additional workers to make up for their lower quality, and which worker types are most loosing out when additional workers are added. If size is important and better workers loose most in productivity when they receive little resources, then the traditional type complementarity F_{xy} must be strong enough for good firms to still employ these types. Substitutability along the quantity dimensions are key to this trade-off. The discussion in Section 4 reveals that as the elasticity of substitution on the quantity dimension goes to zero – in the limit there is no substitution and agents can only be matched into pairs – the right hand side goes to zero.

Finally, one may wonder what happens when our homogeneity assumption does not hold and output is not proportional to the ratio θ of the labor force l to the amount of resources r. Conceptually, the problem is identical to the one we solve here (see the Appendix for the derivation). While the interpretation is much less transparent, the main sorting condition (16) is still necessary for differential positive assortative matching under increasing returns to scale, only the steps that require homogeneity do not apply.

3.2 Equilibrium Assignment, Firm Size Distribution and Wage Profile

In contrast to models with pairwise matching where assortativeness immediately implies who matches with whom (the best with the best, the second best with the second best, and so forth), the matching patter is not immediate in this framework as particular firms may hire more or less workers in equilibrium. Our main focus is the characterization. For the following we will consider output functions that are strictly increasing in types and distributions with continuous non-zero densities, which ensures that all types above some cutoff are matched. If output can fall for higher types, holding all other variables constant, than there might be holes in the matching set, and the following characterization can only be applied on each connected component. The results hold even if output can fall, as long as it is ensured that on the equilibrium path all agents above some cut-off trade. The next proposition fully characterizes the equilibrium.

Proposition 2 *If matching is assortative and output is strictly increasing in types, then the factor intensity (firm size), equilibrium assignment, and wages are determined by the following system of*

include those that satisfy $\theta w(x) = M$. In this case the standard single-crossing condition on f would suffice. Our condition can be thought of as a three-good extension of the Spence-Mirlees condition, where firms can choose different budget levels in terms of the numeraire on top of choosing θ and x.

differential equations evaluated along the equilibrium allocation at almost all types :

$$\text{PAM:} \qquad \theta'(x) = \frac{\mathcal{H}(x)F_{yl} - F_{xr}}{F_{lr}}; \quad \mu'(x) = \frac{\mathcal{H}(x)}{\theta(x)}; \quad w'(x) = \frac{F_x}{\theta(x)}, \tag{10}$$

$$\text{NAM:} \qquad \theta'(x) = -\frac{\mathcal{H}(x)F_{yl} + F_{xr}}{F_{lr}}; \quad \mu'(x) = -\frac{\mathcal{H}(x)}{\theta(x)}; \quad w'(x) = \frac{F_x}{\theta(x)}, \tag{11}$$

where $\mathcal{H}(x) = h^w(x)/h^f(\mu(x))$*.*

Proof. Consider the case of PAM – the case of NAM can be derived in a similar way. Differentiating market clearing condition (5) readily establishes the equation for $\mu'(x)$ in (10). The equation for $w'(x)$ follows from (8) since $F_x = f_x$. Finally, totally differentiating (7) with respect to x and substituting for w' and μ' we obtain:

$$f_{x\theta} + f_{y\theta}\mathcal{H}(x)/\theta(x) + f_{\theta\theta}\theta'(x) - f_x/\theta(x) = 0$$

where we again suppressed the arguments $(x, \mu(x), \theta(x))$ of the production function and its derivatives. This defines $\theta'(x)$. Using (2) we can replace $f_{x\theta}(x, y, \theta)$ with $F_{xl}(x, y, \theta, 1)$, and similarly for the other derivatives. Moreover, in the Appendix (between (15) and (16)) we review that $F_x = \theta F_{xl} + F_{xr}$ and $F_{lr} = \theta F_{ll}$ when evaluated at $(x, y, \theta, 1)$. Substitution then yields the condition for $\theta'(x)$ in (10). ■

This first order differential equation system in μ and θ together with appropriate boundary conditions can be used to compute an equilibrium.[12] Proposition 2 is stated from the point of view of workers, and $\theta(x)$ is the size of the firm in which worker type x is employed. From the firms' perspective, the firm size is $l\,(y) = \theta(\nu(y))$ where $\nu(y)$ is the inverse of $\mu(x)$. Applying the chain rule then immediately implies that $l'(y) = \theta'(\nu(y))\theta(\nu(y))/\mathcal{H}(\nu(y))$ in the case of PAM and the same but with opposite sign in the case of NAM. This immediately generates the following corollary on the size of different firm types:

Corollary 1 *If matching is assortative and and output is increasing in types, better firms hire more workers if and only if along the equilibrium path:*

1. $\mathcal{H}(\nu(y))F_{yl} > F_{xr}$ *under PAM,*
2. $\mathcal{H}(\nu(y))F_{yl} > -F_{xr}$ *under NAM.*

To gain intuition, these results can be interpreted as follows. Consider the case of PAM, and to simplify the exposition we set $\mathcal{H}(x) = 1$ by assuming uniform type distributions, which can be

[12]For PAM, one boundary condition is $\mu(\bar{x}) = \bar{y}$. For a guess of $\theta(\bar{x})$, an equilibrium allocation has to solve the first order differential equation system in μ and θ for all lower worker types. Along the differential equation wages $w(x)$ and firm profits $\pi(\mu(x)) \equiv f(x, \mu(x), \theta(x)) - \theta(x)w(x)$ have to be positive. The guess for $\theta(\bar{x})$ has to be such that at the lowest active type $\hat{x}$ the differential equation stops at one of three possible end-point conditions: $\hat{x} > \underline{x}$ and $w(\hat{x}) = 0$ as not all worker types are used in production, $\hat{x} = \underline{x}$ and $\mu(\hat{x}) > \underline{y}$ with $\pi(\mu(\hat{x})) = 0$ as not all firm types are used in production, or $\hat{x} = \underline{x}$ and $\mu(\hat{x}) = \underline{y}$.

interpreted as a normalization.[13] First, if better firms have a higher marginal value of hiring many workers (the span-of-control complementarity F_{yl} is large), this gives rise to better firms being large. Nevertheless, under PAM they also hire better workers. If these workers have a high marginal value from getting many resources of the firm (F_{xr} large), then the firm will tend to be small. Clearly, if F_{xr} is negative, meaning that better workers need less resources, this generates an even stronger force for firm size to increase in y. Under NAM, the first effect is the same, but now better firms are matched with worse workers. In this case, firms become exceptionally large if better workers need more resources, meaning that worse workers need less resources.

Propositions 1 and 2 provide us with a description of the economy expressed in four interaction terms: F_{xy}, F_{xr}, F_{yl} and F_{lr}, which determine the sorting patterns and the size distribution. These can be used to discuss the determinants that are likely to drive matching in various industries. For example, the most productive firms in the retail market have invested heavily in information technologies to monitor cash registers, the logistics of stocks, and employee performance. This allows a single store manager to supervise a large number of employees, which in our model is captured by a large F_{yl} term. Since top retailers such as Walmart actually pay low wages and hire low skilled employees compared to smaller and less profitable mum-and-pop stores, NAM seems to prevail. From condition (9) we therefore infer that the type complementarity F_{xy} is not too high relative to the span-of-control complementarity F_{yl}. Since top retailers also operate much larger businesses, by the previous corollary we would infer that their span-of-control complementarity must be larger than the negative of the managerial resource complementarity F_{xr}.

In other industries such as management consulting or in law firms, matching appears positive assortative, since top firms hire top graduates. From this we infer that the type complementarityF_{xy} must be large. While it seems natural that the best managers benefit more from having many team members in order to leverage their skills ($F_{yl} > 0$), it is also very beneficial to spend time with the very talented team that they assembled to transfer their knowledge ($F_{xr} > 0$). The type complementarity must be large to outweigh the product $F_{yl}F_{xr}$. Firm size changes according to $F_{yl} - F_{xr}$. The fact that top consultancy firms do not operate much larger groups than lower level ones indicates that the difference between these two complementarities is small.

Interestingly, if matching is PAM and $F_{yl} = F_{xr}$ holds exactly, then the economy operates as in a one-to-one matching model: the ratio of workers to resources is constant, the assignment and the wages are as in Becker (1973). The reason is that the improvements of the firm in taking on more workers are exactly offset by the advantages of the workers to obtain more resources. Since the size

[13] Intuitively, we can always call workers and firms by their rank in the type distribution. Start with an economy with production function F and type distributions H^w and H^f with continuous non-zero densities. Give each worker x a new name $\hat{x}$ that corresponds to his rank in the type distribution: $\hat{x}(x) = H^w(x)$. For firms use similarly $\hat{y}(y) = H^f(y)$. Now production is $\hat{F}(\hat{x}, \hat{y}, l, r) = F((H^w)^{-1}(\hat{x}), (H^f)^{-1}(\hat{y}), l, r)$, where $(H^w)^{-1}$ and $(H^f)^{-1}$ are the inverse of H^w, and H^f, respectively. Clearly, the new economy with "names" $\hat{x}$ and $\hat{y}$ and production $\hat{F}$ generates exactly the same output, but type distributions are by construction uniform.

distribution does not vary across types, the remuneration also does not stray from the one that arises if we exogenously imposed a one-to-one matching ratio.

A final observation concerns the role of the type distribution when it is not normalized. An immediate implication of interest of these equilibrium conditions is that the size distribution $l(y)$ may change even if we hold the production function and the distribution of firms type constant. This occurs when the distribution of workers changes. In particular, for some distributions of worker skills better firms will be smaller, while for other distributions better firms might be larger. That means that even if the technological determinants of firms and their production capabilities are identical in two economies, as is often assumed in the misallocation debate mentioned in the introduction, the firm size distribution can vary even without distortions in the economy, once the skill distribution is taken into account. We will return to this and how the model can be used to analyze some issues within this debate below.

4 Discussion: Special Cases and Relation to the Literature

Before we proceed to the main application of our framework in a tangible economic setting, in this Section we discuss the relation to the existing literature as well as possible simple extensions of our setup. Wherever possible, we derive existing models as special cases within our own setup. This documents how our model nests a number of models that have been heavily used in the literature. It also highlights that our model can capture new settings that have not been analyzed before.

Efficiency units of labor have been a long-standing instrument to incorporate differences in labor productivity (see, e.g., Stigler (1961)) and are still a prevalent assumption in many models in macro and labor economics. It relies on the assumption that workers of a given skill are exactly replaceable by a number of workers of a different skill proportional to their skill difference: workers with half the skill level are perfect substitutes as long as there are twice as many of them. Sorting is then arbitrary. Each firm cares only about the total amount of efficiency units, but not whether they are obtained by few high-type workers or many low-type workers. Our setup captures efficiency units of labor under the production function $F(x, y, l, r) = \tilde{F}(y, xl, r)$. It is easily verified that $F_{xy}F_{lr} = F_{yl}F_{xr}$ in this case.

One-to-one matching models originating from Kantorovich (1942), Koopmans and Beckmann (1957), Shapley and Shubik (1972), and Becker (1973), introduced a meaningful interaction between worker and firm types and have been informative for analyzing interactions in markets with two-sided heterogeneity.[14] They restrict attention to settings where agents have to be matched in pairs, which limits insights into the size of the firm and its capital intensity.[15] Within our setup, one-to-one matching can be captured with the functional form $F(x, y, \min\{l, r\}, \min\{r, l\}) = F(x, y, 1, 1) \min\{l, r\}$ so that

[14] For a recent review article, see Chade, Eeckhout, and Smith (2016).

[15] Notice that the matching models by Tervio (2008), and Gabaix and Landier (2008) which explain the changes of CEO compensation are of this kind. While they use firm size to determine the type of firm, only one worker (the CEO) is matched to one firm, where the firm size is exogenously given.

each unit of labor needs exactly one unit of resource to be productive, and vice versa. This Leontief formulation is only weakly concave, but its well-known sorting condition $F_{xy} > 0$ arises in the limit as our production function approaches this, which is most easily shown in the limit of the following two special cases.

MULTIPLICATIVE SEPARABILITY of the form $F(x, y, l, r) = A(x, y)B(l, r)$ provides particular tractability, and the one-to-one matching case in the previous section can be viewed as a special case where the quantity dimension $B(l, r)$ is Leontief. In the multiplicative case the condition (9) for positive assortative matching is also multiplicatively separable and can be written as $[AA_{xy}/(A_xA_y)][BB_{lr}/(B_lB_r)] \geq 1$.

CONSTANT ELASTICITY OF SUBSTITUTION IN QUANTITIES for the multiplicatively separable case arises if $B(l, r)$'s elasticity is constant and equal to ε. Then the sorting condition reduces to $AA_{xy}/(A_xA_y) \geq \varepsilon$, or equivalently $FF_{xy}/(F_xF_y) \geq \varepsilon$ as the quantity term B cancels. This allows us to capture two special cases of particular relevance. The one-to-one matching model discussed earlier arises as the elasticity of substitution approaches zero, in which case B becomes Leontief and the sorting condition reduces to the well-known $F_{xy} \geq 0$. Another special case is the Cobb-Douglas specification where $B = l^{\gamma}r^{1-\gamma}$. This arises either by assumption as in Grossman, Helpman, and Kircher (2016) that builds on this special case of our work, or when output is linear in the amount of workers but is valued in the market at decreasing returns due to CES preferences of final consumers, as e.g. in Costinot (2009). Either case generates an elasticity of substitution of unity and the sorting condition reduces to $FF_{xy}/(F_xF_y) \geq 1$, or equivalently log-supermodularity of F in worker and firm types, which is the well-known condition in this literature.

SUPERVISION-TIME MODELS have been amongst the first to allow sorting in the presence of interaction with more than one worker. Here the firm or its manager has a unit amount of time to supervise workers. The supervision time $t(x, y)$ needed by each worker depends on both the manager's and the worker's type. So r units of time allow the hiring of (no more than) $r/t(x, y)$ workers, and this determines firm size which is no longer a real choice variable once types are known. Sattinger (1975)'s seminal work assumes that output equals size: $F(x, y, l, r) = \min\{r/t(x, y), l\}$. If we approximate this non-differentiable output function by the inelastic limit of a CES production function with inputs $r/t(x, y)$ and l,[16] sorting condition (9) requires $t(x, y)$ to be log-supermodular in the inelastic limit, recovering Sattinger's condition. Related is Garicano (2000)'s model of problem solving that has been widely applied in the macro and trade literature (e.g., Garicano and Rossi-Hansberg (2006); Antràs, Garicano, and Rossi-Hansberg (2006)). Here supervision time $t(x)$ only depends on the worker's ability to solve problems, but managers themselves contribute directly to production by solving problems up to level y, leading to output function $F(x, y, l, r) = y \min\{r/t(x), l\}$. Again approximation through a smooth CES function recovers their condition that sorting is always positive. The beauty of these models is that they incorporate sorting and their explicit structure allows extensions for example to multiple hierarchical

[16]The function $F(x, y, l, r) = .([r/t(x, y)]^{(\varepsilon-1)/\varepsilon} + l^{(\varepsilon-1)/\varepsilon})^{\varepsilon/(\varepsilon-1)}$ approaches $\min\{(r/t(x, y), l\}$ as $\varepsilon \to 0$.

levels, but size is directly tied to types and does not allow a smooth extensive margin that is the heart of most macro models.

"Smooth" Span of Control underlies much of the work in macro-economics on firm size distributions and is inspired by Lucas (1978)'s seminal work. He assumed that managers with different types leverage their time smoothly over a (homogeneous) workforce, which can be captured through $F = yl\varphi(r/l)$, where φ summarizes decreasing returns due to span-of-control problems. With the specification $\varphi(r/l) = (r/l)^{1-\gamma}$ this recovers the common form in which a firm with a unit amount of resources has decreasing returns in labor of form yl^{γ}. Rosen (1982)'s supervision model can be interpreted as introducing heterogeneous worker types into this framework through his production function $F = yl\varphi(g(y)r/l, x)$.[17] Parametrization with $\varphi = \min\{\frac{r}{lt(x)}, 1\}$ would exactly replicate Garicano (2000), but instead, Rosen imposed the smoothness assumptions of standard neoclassical theory. Rosen never analyzed his general version, but additionally assumed efficiency units of labor. Within our framework we can apply our sorting condition (9) to study sorting in his general model, which yields: $\varphi_{12}\left[\varphi_1 - \varphi/[g(y)r/l]\right] \geq \varphi_2\varphi_{11}$, where subscripts denote partial derivatives of φ and its arguments $(g(y)r/l, x)$ are suppressed. Careful inspection of this condition yields many insights: a positive φ_{12} is intuitively conducive to positive sorting as it captures the interaction between $g(y)$ and x, but this turns out only to be true if the elasticity of φ with respect to its first argument is above unity. Otherwise decreasing returns to size kick in too much and the square bracket is negative, revealing again the importance of size considerations for sorting. The converse that sorting influences firm size distributions might also seem plausible, but again it has not received any attention, possibly due to a lack of theory that allows for sorting in conjunction with span of control. Empirical work on firm size distributions tends to use the span-of-control approach since firm heterogeneity allows them to rationalize size differences. Input heterogeneity in such studies is usually absent or restricted to efficiency units. To illustrate the role of such heterogeneity one can study changes within a country over time (which constitutes our main quantitative application in the next section) or between countries of different levels of development. Next we briefly touch on the latter.

The Misallocation Debate refers to empirical work that identified a non-trivial tail of large firms in the US and other developed countries, whereas such a tail is absent in developing countries where size is compressed at very low levels (e.g., Hsieh and Klenow (2009), Restuccia and Rogerson (2008), Guner, Ventura, and Xu (2008)). The same holds for agricultural farms (Adamopoulos and Restuccia (2014)). Span of control models have been unable to rationalize the differences, raising the worry about misallocation of inputs away from the most productive firms in developing countries. While the literature has discussed the role of input heterogeneity, we are not aware of frameworks that move beyond either homogeneity or efficiency units. To illustrate how our framework might contribute, consider

[17] Rosen (1982)'s equation (1) for output per worker can be written as $\tilde{g}(\tilde{y})\varphi(\tilde{y}l/r, x)$, where $\tilde{y}$ is the firm type. For a strictly monotone function $\tilde{g}$ we can relabel each firm type as $y = \tilde{g}(\tilde{y})$ so that output per worker is $y\varphi(g(y)r/l, x)$, where $g = \tilde{g}^{-1}$, which yields the expression in the text.

Adamopoulos and Restuccia (2014) who rely on farmer heterogeneity y and input l which represents land in their setting, as well as some generic capital k that can be rented at unit cost R.[18] Consider an extension to their production function of form: $\tilde{f}(x, y, l, k) = a(\eta(xk)^\rho + (1-\eta)(yl)^\rho)^{\frac{\gamma}{\rho}}$ with parameters a, η, ρ, γ. Farmer quality augments land holdings, but they consider only homogeneous inputs ($x = 1$) despite a discussion section on heterogeneity. In the generalized form x now simply augments capital so that better inputs use capital more efficiently, even though many other specifications would fit within our larger theory. Optimal profit given types and land holdings is $f(x, y, l) = \max_k \tilde{f}(x, y, l, k) - Rk$, which allows us to use the results from the main body that did not explicitly incorporate generic capital, and in the Appendix we show that this parametrization generates positive sorting. Taking the rental rate of generic capital and the remaining parameters from Adamopoulos and Restuccia (2014), we see that a mean-preserving spread in input heterogeneity reduces heterogeneity in the distribution of land holdings across farms, as better firms buy less but better land. There are indications that land quality in developing countries might be more dispersed, which would then limit the right tail of large firms in such countries.[19]

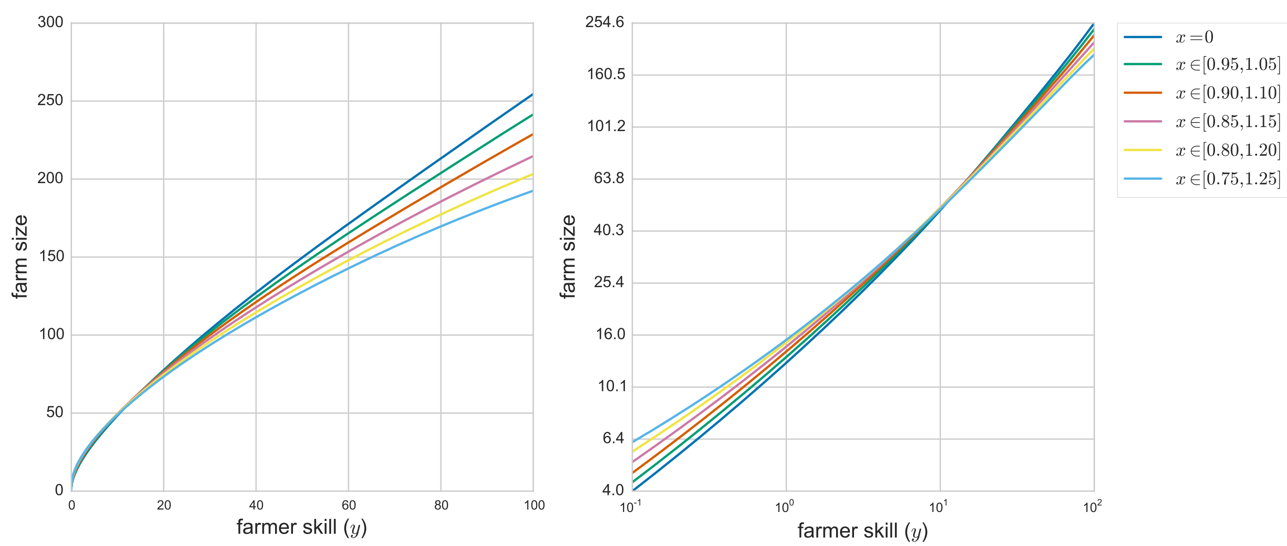


Figure 1: Firm size distribution for different dispersion in x (left in levels, right in logs).

Figure 1 uses the parameters for the developing countries and shows how firms of different types

[18]Note that k represents a generic input such as fertilizers or tractors, while r is a specific limit on the farmer such as his time endowment. Note also that agriculture might be a particularly suitable application of our theory once generic capital is included: Both in the US and in developing countries a farm is generally run by the farmer and his family and their time endowment is the relevant resource constraint (contrary to intuition seasonal help is only a minor part of overall farm labor, while generic capital in form of tractors is the main source of additional help). Also, the restriction that all land within the farm is of equal quality might not be that restrictive if farmers choose where to locate and local land has somewhat uniform quality.

[19]Variation in land quality as measured by satellite images is positively correlated with ethnolinguistic variation (Michalopoulos (2012)) while ethnolinguistic variation is typically negatively related to GDP (Michalopoulos and Papaioannou (2012)).

react to more dispersion: large ones shrink and small ones grow, in levels on the left and in percentage terms on the right (details are described in the Appendix). Similar compressions that limit the right tail might occur in the context of industrial firms and their labor inputs, since dispersion of labor inputs is negatively correlated with the overall level of education (Thomas, Wang, and Fan (2001)). We do not expect input heterogeneity to fully rationalize the differences between countries. Rather, we aim to provide a tool that allows a formal discussion of these issues and their importance. A full exploration would require a re-estimation of the generalized framework and an endogenous determination of the rental rates of capital within the larger economy beyond the single sector that is at the heart of our theory.

Economic geography has long been concerned with how individuals choose to locate across space, for example within different locations of a mono-centric city, though many models tend to abstract from spatial sorting.[20] We therefore consider a model of spatial sorting within the city. Let there be a continuum of locations y relative to the center, each with space for construction $h^f(y)$. Agents with budget x have quasi-linear preferences $u(c, s) = c + v(s)$ over consumption c and housing space s. The budget constraint is $c + p_s(y)s = xg(y)$, where x is the worker skill and $g(y)$ is an increasing function representing the time at work rather than in commute. Then we can write the individual citizen x's optimization problem as $xg(y) + v(h) - p_s(y)s$. Net of the transfers, the aggregate surplus for all l citizens is given by $F(x, y, l, r) = xg(y)l + v\left(\frac{r}{l}\right)l$. It is easily verified that sorting condition (9) is satisfied if $v(\cdot)$ is concave, so that individuals with high incomes locate centrally and those with low incomes in the periphery.[21]

Sorting in the presence of search frictions has recently attracted substantial attention, and has been explored exclusively in the setting of one-to-one matching, as in Shimer and Smith (2000) and Atakan (2006) under random search and Shi (2001), Shimer (2005) and Eeckhout and Kircher (2010) under competitive search. Under competitive search, workers see the firm characteristics and wage offers before queueing for jobs. This renders it close to competitive models, and in fact the sorting condition in Eeckhout and Kircher (2010) can be interpreted as a special case of multiplicative separability presented above, with the quantity dimension replacing the matching function. While multi-worker matching has recently attracted attention within the directed search literature (e.g., Menzio and Moen (2010), Hawkins (2011), Kaas and Kircher (2015), Schaal (2015)) we are not aware of an analysis with large firms, with search frictions *and* with two-sided heterogeneity.[22]

[20]Most work in new economic geography considers the location choice of homogeneous agents through indifference conditions. Related to our setup is in particular Lucas and Rossi-Hansberg (2002) who model the location of identical citizens and incorporate productive as well as residential land use. Though agents are identical, they earn different wages in different locations. The paper proves existence of a competitive equilibrium in this generalized location model which endogenously can generate multiple business centers. For a model with spatial sorting between cities, see Eeckhout, Pinheiro, and Schmidheiny (2014).

[21]A similar functional form is used in Van Nieuwerburgh and Weill (2010) to consider differences between cities rather than within the city, where in there model $g(y)$ is replaced by a more agnostic time-varying productivity term that differs across cities. Clearly the sorting of more talented workers to more productive cities prevails.

[22]Lentz (2010) and Bagger and Lentz (2016) study sorting when firm size is limited only by search frictions, albeit with

To illustrate that our model is amenable to this, assume that workers cannot be hired frictionlessly, and firms have to post a measure v of vacancies at a cost c per vacancy. They specify a worker type and wage offer (x, w) and workers see all offers and decide where to queue for jobs after observing these. If u workers queue for v vacancies with a particular offer, the number of matches that arise is determined by matching function $M(u, v)$ with the usual properties. In the Appendix we show that the techniques developed in this paper are suitable to analyze search frictions with large firms and sorting, and notably that the sorting condition for PAM remains unchanged at $F_{xy}F_{lr} \geq F_{xl}F_{yr}$ and is thus independent of the search frictions. Obviously the matching function does affect the equilibrium allocation and the unemployment rate, and we derive predictions about the unemployment rate of various worker types and the vacancy rate across firms in the Appendix.

Extensions that allow for capital investment are discussed in the misallocation debate and explored in the Appendix. In the Appendix, we additionally introduce monopolistic competition at the output level with a Dixit-Stiglitz setup and derive the extended conditions that arise. There we also link our model to the optimal transportation literature, and we introduce endogenous type distributions.

Limitations. While the assortative matching literature has made rather specific assumptions for multi-worker firms that we attempt to generalize, the combinatorial matching and general equilibrium literature has instead stayed general but has focussed mainly on existence theorems rather than on characterizing the sorting or the wage patterns. The classic example in the combinatorial matching literature is Kelso and Crawford (1982). They propose a many-to-one matching framework in a finite economy and allow for arbitrary production externalities, both between the firm and its workers and across the workers within the firm. In such a general setting it is well-known that the stable equilibrium or the core may not exist, and Kelso and Crawford (1982) derive a sufficient condition for existence in a finite agent model, that of gross substitutes: adding another worker decreases the marginal value of each existing worker. This condition is satisfied in our setting where externalities are mainly between the firm and the workers while across-worker externalities are due to scarcity of internal resources only, and scarcity becomes more binding when there are fewer workers. Gul and Stacchetti (1999) analyze the gross substitutes condition in the context of Walrasian equilibrium and show existence and the relation between the Walrasian price and the payment in the Vickrey-Clarke-Groves mechanism. In the context of auction design, Hatfield and Milgrom (2005) analyze package bidding as a model of many to one matching. Our model differs from settings such as the Roy (1951) model and its recent variants in e.g., Heckman and Honore (1990), where each firm (or sector) can absorb unbounded numbers of agents. In our setup, the marginal product decreases as the firm gets larger. Models that combine the Roy setup with decreasing returns due to price effects such as Costinot (2009) do share commonalities to our model that are discussed under multiplicative separability above.

---

linear production that that there is no interaction of workers within the production process of the firm.

5 Application: Quantity Biased and Skill Biased Technological Change

As an illustration of our theory we apply the model to analyze technological change. We consider quantity-biased technological change that affects firms' returns to scale and allows larger firms in equilibrium, which highlights how accounting for changing firm size within a sorting model offers a new view on technological change that might counteract some known determinants of wage dispersion, such as skill-bias technological change.

We use German matched employer-employee data to gain insights in the importance of the different forces in our model. Observing information on wages, as well as profits and firm size, we use our model to ask how the different determinants of technology have changed between 1996 and 2010. Unfortunately, the model does not accommodate within-firm distributions of workers, therefore we treat the worker within the firm as a representative agent using the average characteristics of the firm's work force. A sequence of recent empirical papers has documented in many countries that changes in wage inequality are driven nearly exclusively by between-firm inequality rather than within-firm inequality,[23] so that our focus on skill and wage heterogeneity between firm – while abstracting from within-firm heterogeneity – is still be a valuable exercise. In the light of this restriction on the environment, the focus of our attention is how the relation between firm size and *average* within-firm skill and pay has evolved over that 15 year period. As we will see, the model will give us insights well beyond the standard Becker one-to-one-matching model or the canonical Lucas span-of-control model.

Data. We use linked Employer-Employee data from IAB, the Institute for Employment Research at the German Federal Employment Agency. In particular, our data is from the LIAB, the Integrated Establishment and Individual Data, with observations for 1996 and 2010, the first and the last year for which we have consistent data. We focus attention on firms with 5 or more employees. All prices are expressed in 2010 Euros, using the HCPI to deflate the 1996 prices. For 1996 we have 2,984 firms in 150 weighted bins and for 2010 we have 5,083 firms in 255 weighted bins.[24] The weights are applied to make the sample representative of the population of German firms. The variables we use are the number of full-time employees, the median wage in the firm, and the profits.

Technology and distribution. We assume CES in (x, y) with elasticity of substitution σ and weights ω_x, ω_y and CRTS Cobb-Douglas in (l, r) with the expenditure share on l equal to ω_B, where for ease of exposition we represent the production already for a firm that uses all its resources on one type:[25]

[23] See Card, Heining, and Kline (2013) for Germany, Song, Price, Guvenen, Bloom, and von Wachter (2015) and Barth, Bryson, Davis, and Freeman (2014) for the US, Benguria (2015) for Brazil, and Vlachos, Lindqvist, and Hakanson (2015) for Sweden.

[24] Due to privacy concerns, data can only be accessed on the IAB servers and only aggregated information can be exported, hence the binning of the firm distribution.

[25] Recall that the general function for arbitrary r can be recovered simply via $F(x, y, l, r) = rF(x, y, l/r, 1)$. Also, observe

$$F(x,y,l,1) = \left(\omega_x x^{\frac{\sigma-1}{\sigma}} + \omega_y y^{\frac{\sigma-1}{\sigma}}\right)^{\frac{\sigma}{\sigma-1}} l^{\omega_l}.$$

We assume that the distribution of types x and y is log-normal, where the mean and variance are estimated together with parameters of the production function to match the 3 moment conditions as detailed below.

Estimation Procedure. Given the technology and using the three differential equations from Proposition 2 with appropriate end point conditions, we solve the equilibrium allocation $\mu(x)$, the equilibrium firm size $\theta(x)$ and the wage $w(x)$. The parameters $\omega_x, \omega_y, \sigma, \omega_B$ are chosen to minimize the sum of squared residuals between the size-wage and the size-profits relations, as well as the firm size distribution in the data and in the model.[26] The estimation is performed independently for each year. We have also done several robustness checks, for example estimating 2010 fixing the estimated 1996 distribution parameters. The results are qualitatively and quantitatively very similar.

Table 1: Estimated Parameters

	1996	2010	% change
Technology			
ω_x	0.026	0.060	131.6%
ω_y	0.974	0.964	-1.1%
ω_l	0.123	0.217	76.1%
σ	0.998	0.982	-1.6%
Distributions			
x	$\mathcal{LN}(2.49, 1.35)$	$\mathcal{LN}(2.69, 1.35)$	
y	$\mathcal{LN}(0.08, 1.57)$	$\mathcal{LN}(0.03, 1.54)$	

Results. We find the estimates presented in Table 1.[27] With these estimates, Figure 3 shows the model fit with the targeted moments for both years: the relations wage-firm size, and profits-firm size, as well as the firm size distribution.

The distributions estimated to fit a log-normal distribution are plotted in Figure 2. Observe that the distribution of firm types hardly changes. Instead, the distribution of worker types shows a first order dominance shift to the right. While skill distributions in the US have remained fairly constant

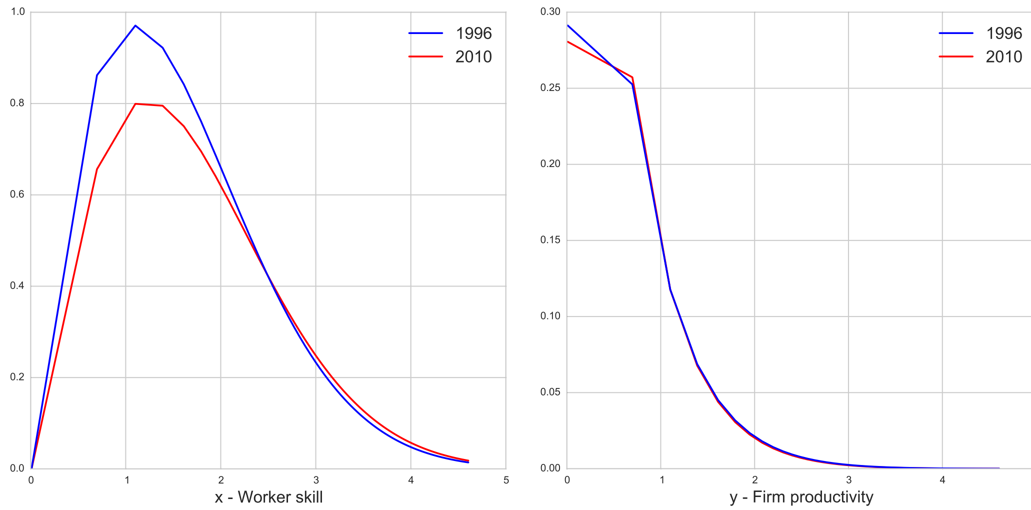
that with a simple transformation of variables, this technology could be written as

$$F(x,y,l,1) = A\left(\omega_A x^{\frac{\sigma-1}{\sigma}} + (1-\omega_A) y^{\frac{\sigma-1}{\sigma}}\right)^{\frac{\sigma}{\sigma-1}} l^{\omega_l}$$

where $\omega_x = \omega_A A^{\frac{\sigma-1}{\sigma}}$ and $\omega_y = (1-\omega_A) A^{\frac{\sigma-1}{\sigma}}$.

[26] We first pin down the parameter space – the support for the distributions of x and y is normalized to [1,100] – and then use the L-BFGS-B algorithm for the minimization routine. To find the global minimum we use the stochastic "Basin Hopping" algorithm.

[27] Because of the time intensity of the solution to the differential equations, each minimization round takes several hours. We can therefore not provide any confidence bounds for the precision of the parameters, since bootstrapping methods are not feasible.



(a) Worker Types x (log scale) (b) Firm Types y (log scale)

Figure 2: The Distributions of Worker Types x and Firm Types y.

since the 1970s, it is well known that in most European countries skills have increased.[28] An immediate implication is that firms can hire better workers, which is picked up in the theory (see the changed assignment in Figure 5b below).

Now we turn to the estimated technology. First, observe that $\sigma < 1$ is consistent with PAM. This follows immediately from verifying condition (9) in Proposition 1. PAM is necessary to rationalize that both profits and wages are high in large firms, implying that high types from both sides match. Second, given that $\sigma \approx 1$, the CES term in the inputs (x, y) can be approximated by the Cobb-Douglas technology.[29] The approximate technology can therefore be written as:

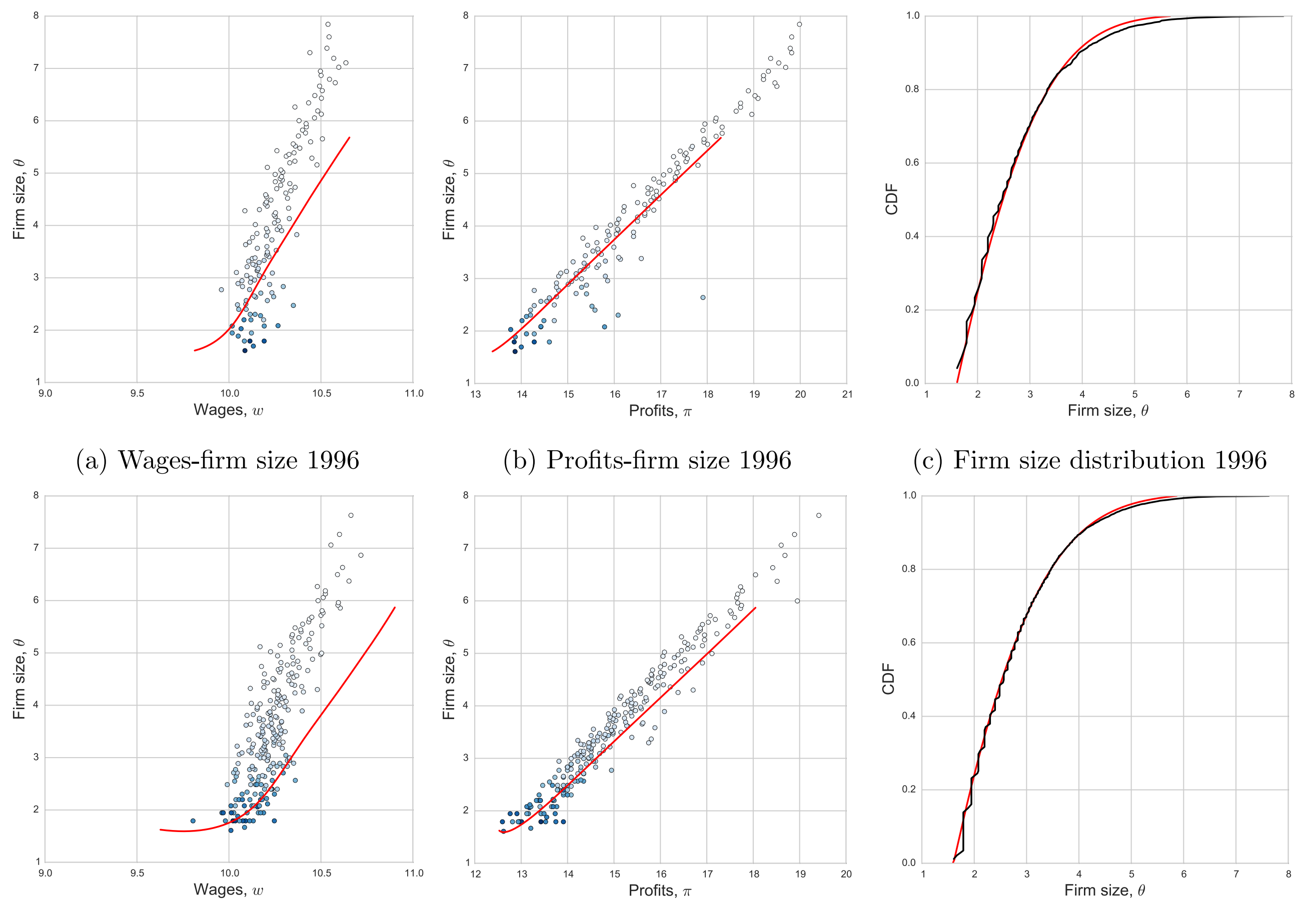
$$F(x, y, l, 1) \approx x^{\omega_x} y^{\omega_y} l^{\omega_l}.$$

Quantity-Biased Technological Change. We now turn to the evolution of the technology between 1996 and 2010. Eyeballing the data, the main difference between these years is a changed wage profile and the presence of larger firms. We discuss how this is reflected in the evolution of the technological parameters in Table 1. In addition, in Figure 4 we decompose the estimated technology by calculating the complementarity terms F_{xy}, F_{lr}, F_{yl} and F_{xr}.

We find that the marginal product of the skill coefficient ω_x has increased by 3.4 percentage points and has more than doubled. In contrast, the marginal product of firm technology has decreased by

[28]It may appear that the measures of both distributions in Figure 2a are not the same, but that is merely due to the log scale. In the Appendix we report the CDF in levels.

[29]It is important that the estimate is strictly less than zero because with σ exactly equal to 1, the sorting condition holds with equality and the allocation $\mu(x)$ is indeterminate, i.e., any allocation is consistent with equilibrium.



(a) Wages-firm size 1996 (b) Profits-firm size 1996 (c) Firm size distribution 1996

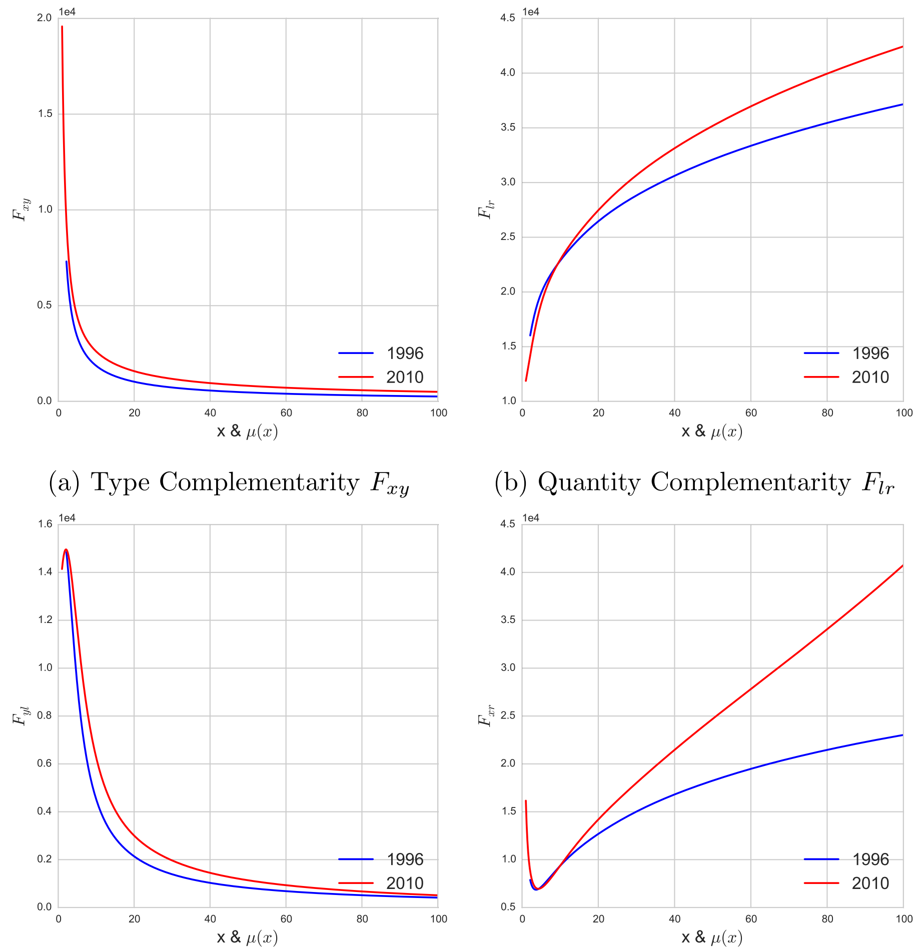
(d) Wages-firm size 2010 (e) Profits-firm size 2010 (f) Firm size distribution 2010

Figure 3: The fit of the targeted moments in 1996 and 2010. The shading of the data points in figures (a) and (b) reflects the sample weights used to normalize the sample to the German size distribution.

1 percentage point, though the relative change is small given the high level. This reflects the fact that technological progress has been predominantly in labor productivity, and allows to rationalize an increased skill premium that has long been associated with Skill-Biased Technological Change.

The complementarity between worker and firm quality x and y continues to be close to Cobb-Douglas, but since the elasticity drops by more than 1.5 percentage points, the complementarity between x and y has increased (Figure 4a), contributing further to Skill-Biased Technological Change since higher skills are now more valued at better firms.

Most importantly, the biggest technological change is the increase in the coefficient ω_l measuring the marginal product of the quantity of labor. It has gone up by 71% from 0.123 to 0.217. This reflects that there has been Quantity-Biased Technological Change (QBTC). This can be observed graphically in different ways. The direct effect is that the quantity complementarity F_{lr} has gone up (Figure 4b). But also the span-of-control complementarity has increased (Figure 4c), and so has the managerial resource complementarity F_{xr}, especially for high types (Figure 4d).



(a) Type Complementarity F_{xy} (b) Quantity Complementarity F_{lr}

(c) Span-of-Control Compl. F_{yl} (d) Managerial Resource Compl. F_{xr}

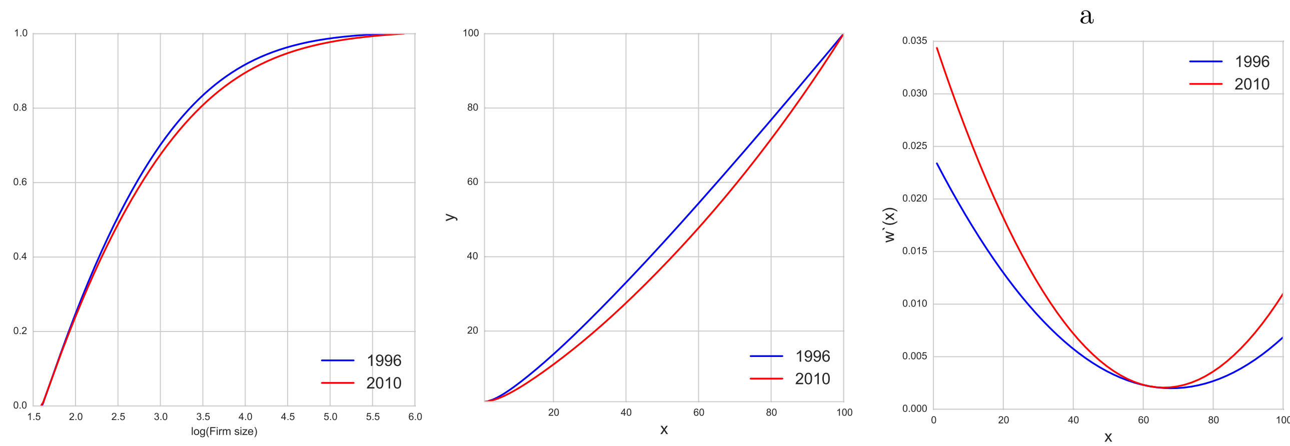
Figure 4: Quantity-Biased Technological Change: Complementarities.

The quantity-bias in the technological change rationalizes the overall increase in firm sizes, as is evident from the First-Order Stochastic Dominance shift in the distribution of firms between 1996 and 2010 (Figure 5a). In general this pushes more workers into top firms, but this is counteracted by the increased supply of high skilled workers (Figure 5b).

The wage premium as measured by $w'(x)$ is clearly positive and has gone up substantially since 1996 (Figure 5c).[30]

For those working in the lowest skilled firms, the wage premium has gone up nearly 40% from 0.023 to 0.032, as well as for those working in firms with the highest skilled workers, where it has gone up 35% from 0.07 to 0.095. In the middle of the distribution, the increase in the wage premium is close to zero. This is consistent with the finding in the literature on job polarization that middle skill jobs

[30]The skill premium is often expressed as the elasticity of the wage $xw'(x)/w(x)$ which can readily be interpreted as the coefficient of a regression of log wages on log skills. In our model, the elasticity has a simple analytic expression – $\varepsilon_w = xF_x/(\theta F_\theta)$ – and a transparent interpretation with a Cobb-Douglas technology: if θ goes down, then θF_θ goes down.



(a) Size Distribution (b) Equilibrium Allocation $\mu(x)$ (c) Skill Premium $w'(x)$

Figure 5: Firm Size, Allocation, Skill Premium: 1996 vs. 2010.

have relatively become less productive (see Goos and Manning (2007), Katz, Kearney, et al. (2006), and Autor and Dorn (2009)).[31]

From this exercise we conclude that the magnitude of quantity biased technological change has been substantial. The technology has changed favoring larger firms. But at the same time, the skill premium has gone up substantially. Of course, these QBTC and SBTC interact: as the skill premium rises, part of the benefits from hiring more workers is offset by higher wages (and in addition there has been a supply response: the distribution of skills shifts outwards). The observed effect on the size distribution therefore masks some of the technological change. From the differential equation that pins down the wages $\theta w'(x) = F_x$, we immediately observe that an increase in firm size is mitigated by an increase in the skill premium. In other words, even if the effect on the size distribution is relatively small (See Figure 5a), underlying there is a huge change in both the marginal product of skilled labor ω_x and the marginal product of size ω_l, with these effects offsetting each other in the equilibrium size distribution and also in the magnitude of the skill premium. Without the QBTC, the skill premium would have been even higher.

Counterfactual Exercises. Finally, we run some simple counterfactuals that decompose the effects of the individual changes in parameters and illustrate the point in the previous paragraph. We fix the economy at the 1996 estimated parameters and then substitute one-at-a-time a parameter for their estimated counter-part in 2010. In Table 2 we report the change in the median firm size and the average skill premium for all parameters, and we then graphically evaluate the impact on the size distribution, on the equilibrium allocation and on the skill premium in Figure 6. For the figure we focus on the 2010 values for ω_x, ω_l and the distribution parameters, respectively, and report the figures for

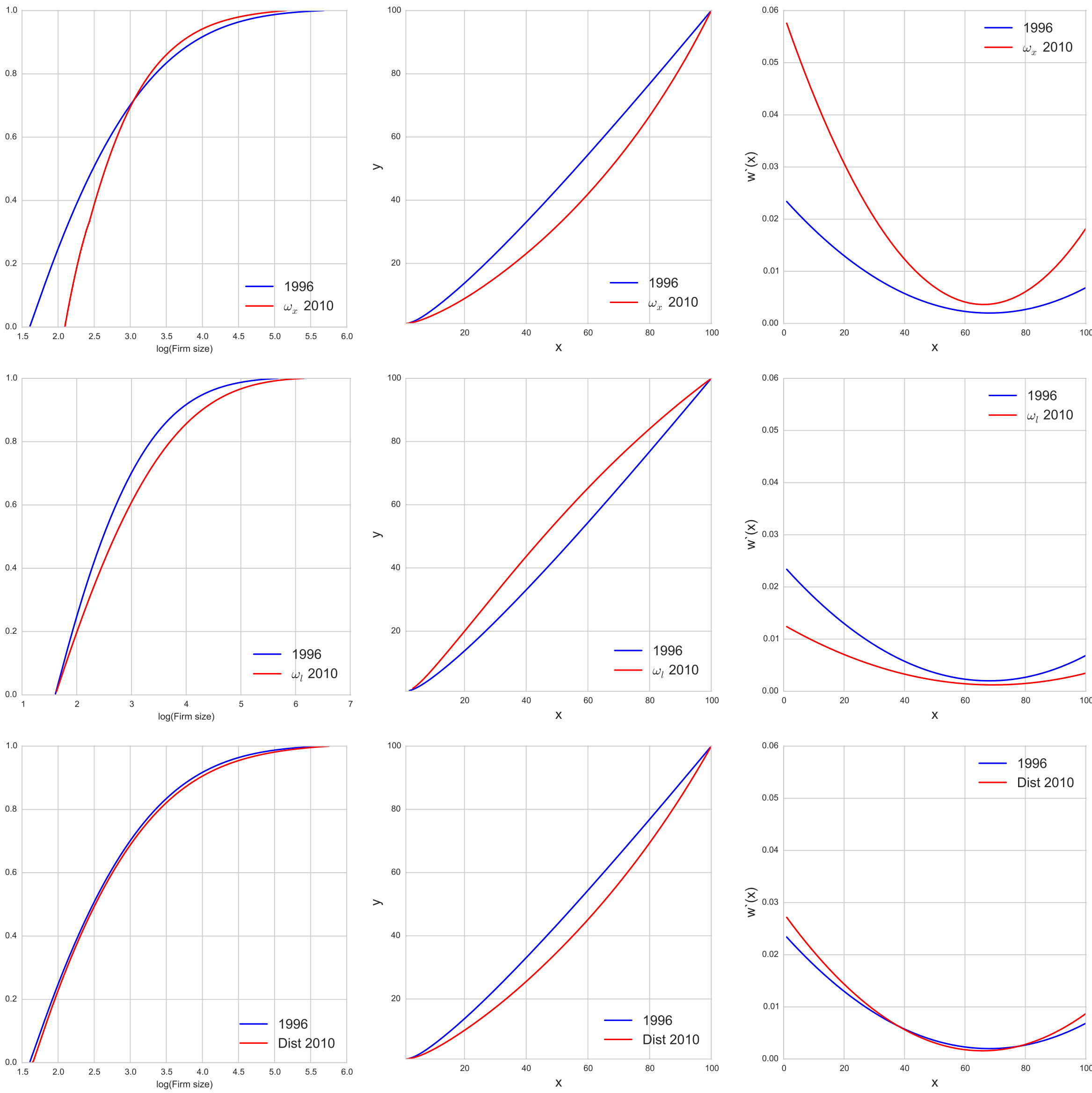
[31] Usually, when people talk about job polarization, they compare wage change $w_{2010}(x) - w_{1996}(x)$ across different x. Then the relative change between x and x' $(w_{2010}(x)-w_{1996}(x))-(w_{2010}(x')-w_{1996}(x'))$ equals $\int_{x'}^{x} [w'_{2010}(\tilde{x}) - w'_{1996}(\tilde{x})]\, d\tilde{x}$, and is therefore captured by the skill premium.

Table 2: Counterfactuals: Change in Firm Size and Skill Premium

	Median Firm Size	% change 1996	Average $w'(x)$	% change 1996
1996	11.98		0.019	
2010	12.53	4.60 %	0.027	44.06%
2010 ω_x	14.21	18.66%	0.049	156.90%
2010 ω_y	11.95	-0.21%	0.019	1.90%
2010 ω_l	14.81	23.65%	0.009	-52.04%
2010 σ	12.01	0.24%	0.022	13.68%
2010 Distributions	12.36	3.20%	0.022	13.68%

ω_y and σ in the Appendix as there are virtually no effects for them.

The Table and the Figure confirm that importance of both ω_x and ω_l for the determination of the firm size as well as the wage premium in 2010 compared to 1996. But what is most striking is the fact that the individual contribution of each of these technological components is much bigger than what we observe in equilibrium. The reason is precisely that some of the effects offset each other in equilibrium. The contribution of the increase in the productivity of skilled labor ω_x to the skill premium is enormous, and more than 3 times the equilibrium increase. The reason that the equilibrium effect is damped is due to the increase in the productivity of the quantity of labor ω_l, which has a substantial negative impact on the skill premium. The skill premium $w'(x) = F_x/\theta$ is increased by the presence of ω_x which raises the numerator, but the effect of ω_l ends up raising the denominator to counteract this effect. At the same time, it appears that ω_x also increases firm size, but this effect is very different at different points in the distribution. The top of Figure 6a shows that the impact on the size of firms that hire high skilled workers is actually negative. In contrast, an increase in the size parameters ω_l not only increases the median firm size but also increases firm sizes throughout the distribution and ends up substantially reducing the skill premium as mentioned earlier. Also the distributional parameters have some impact, though much more modest than that of ω_x and ω_l. Finally, there is hardly any effect of the change in the firm productivity ω_l or the elasticity of substitution σ between x and y.



(a) Size Distribution (b) Equilibrium Allocation $\mu(x)$ (c) Skill Premium $w'(x)$

Figure 6: Counterfactuals. The 1996 economy is simulated with one parameter changed to the estimated value for the 2010 economy. Row 1: ω_x from 2010; Row 2. ω_l from 2010; Row 3. Distributions for x and y from 2010. In the first column we report the firm size distribution, in the second column the equilibrium allocation and in the third column the skill premium.

6 Concluding Remarks

Assortative Matching is prevalent across firms of different sizes. We propose a tractable theory of the labor market where firms choose both the quality of the work force and the quantity. This allows us to study sorting and firm size simultaneously. Whether assortative matching is positive now depends on a tradeoff of complementarities between types, between quantities, and across types and quantities. The equilibrium allocation is completely characterized by a system of differential equations that pins down the allocation, the firm size distribution and the wage distribution.

Our model provides a unified approach to a number of existing models in the macro and labor literatures. It is sufficiently rich to incorporate the most relevant features of heterogeneity, in particular worker skill, firm productivity, firm size and wage inequality. Yet, it is remarkably simple to analyze and can readily be used to plug into a larger model of the economy. For example, we establish that equilibrium unemployment can be incorporated, thus proposing the first analysis of frictional unemployment in the presence of both sorting of heterogeneous workers and firms, and of matching in large firms.

To further illustrate the applicability of the theory, we have analyzed the role of technological change in the presence of both sorting and endogenous firm size. We find that Quantity-Biased Technological Change is sizable, and much bigger than what we would impute from merely inspecting the change in the size distribution. Likewise, Skill-Biased Technological Change is bigger than what already transpires from observing the skill premium in isolation. The reason that these big technological determinants of QBTC and SBTC appear muted is the interplay between the two in equilibrium, where the effects of these technological determinants partially offset each other. This has important implications for our understanding of technological change and wage inequality. Technological change has been even more pronounced than what was thought, and the mechanism we propose shows that the firm size distribution has important implications for labor market outcomes.

7 Appendix

7.1 Convex Preferences, Existence, and Welfare Theorems

One can interpret our economy in terms of a classical exchange economy: "Consumers" in the classical model are our firms $y \in \mathcal{Y}$. They consume a bundles $(\mathcal{L}^y, n^y)$ where $\mathcal{L}^y$ denotes the amounts of labor of various types employed by firm y and n^y is the amount of numeraire it consumes. To make this an endowment economy, assume that each firm is initially endowed with some of the workers and a sufficiently high level of the numeraire. The exact endowment of workers to firms does not matter because of the presence of the numeraire, so endowing each firm with the average distribution of workers would suffice. Firm preferences are represented by utility function $u(\mathcal{L}^y, n^y|y) = n^y + \max_{\mathcal{R}^y} \int f(x, y, \theta^y(x))d\mathcal{R}^y(x)$ such that $\theta^y(x) = d\mathcal{L}^y/d\mathcal{R}^y$ and $\int d\mathcal{R}^y < 1$ where the first term captures the numeraire and the second optimal production. If these preferences are convex, we can apply Ostroy (1984) or Khan and Yannelis (1991) for existence and the former for core equivalence.

Lemma 2 *Firm preferences are convex.*

Proof. Consider three bundles $(\mathcal{L}^y, n^y)$, $(\mathcal{L}'^y, n'^y)$ and $(\mathcal{L}''^y, n''^y)$ such that $u(\mathcal{L}^y, n^y|y) \geq u(\mathcal{L}''^y, n''^y|y)$ and $u(\mathcal{L}'^y, n'^y|y) \geq u(\mathcal{L}''^y, n''^y|y)$. We then establish that $u(\alpha\mathcal{L}^y + (1-\alpha)\mathcal{L}'^y, \alpha n^y + (1-\alpha)n'^y|y) \geq u(\mathcal{L}''^y, n''^y|y)$ for any $\alpha \in (0,1)$ since the firm can simply assign a fraction α of its internal resources to workers with distribution $\mathcal{L}^y$ and the remainder to the other workers, that is

$$
\begin{aligned}
& u(\alpha\mathcal{L}^y + (1-\alpha)\mathcal{L}'^y, \alpha n^y + (1-\alpha)n'^y|y) \\
= \; & \alpha n^y + (1-\alpha)n'^y + \max_{\substack{\mathcal{R}^y \text{ s.t.} \\ \theta^y(x) = d(\alpha\mathcal{L}^y + (1-\alpha)\mathcal{L}'^y)/d\mathcal{R}^y \\ \int d\mathcal{R}^y < 1}} \int f(x, y, \theta^y(x))d\mathcal{R}^y(x) \\
\geq \; & \alpha n^y + (1-\alpha)n'^y + \max_{\substack{\mathcal{R}^y \text{ s.t.} \\ \theta^y(x) = \alpha d\mathcal{L}^y/d\mathcal{R}^y \\ \int d\mathcal{R}^y < \alpha}} \int f(x, y, \theta^y(x))d\mathcal{R}^y(x) + \max_{\substack{\mathcal{R}^y \text{ s.t.} \\ \theta^y(x) = (1-\alpha)d\mathcal{L}'^y/d\mathcal{R}^y \\ \int d\mathcal{R}^y < 1-\alpha}} \int f(x, y, \theta^y(x))d\mathcal{R}^y(x) \\
= \; & \alpha u(\mathcal{L}^y, n^y|y) + (1-\alpha)u(\mathcal{L}^y, n^y|y) \\
\geq \; & u(\mathcal{L}''^y, n''^y|y).
\end{aligned}
$$

■

7.2 Sorting with monotone production and smooth distributions

Lemma 3 *If output F is strictly increasing in x and y and the type distributions have non-zero continuous densities, then almost all active firm types y hire exactly one worker type $\nu(y)$ and reach unique size $l(y)$ in an assortative equilibrium.*

Proof. First, note that optimality requires that for any given firm type y almost all its choices $x \in \text{supp}\mathcal{L}^y$ and $\theta^y(x)$ solve problem (6) by Lemma 1. Next, observe that if x is in the support of the labor demand $\mathcal{L}^y$ for any firm y, then all $x' > x$ have to be active. If not, the wage for the higher type x' is $w(x') = 0$ and therefore weakly below $w(x)$, but x' produces more output than x, which violates that x is optimal for firm y (formally: it violates that x is in the support of $\mathcal{L}^y$ as all types in a small enough neighborhood around x such that their type is below x' fail to maximize (6)).

Next, we show that an assortative equilibrium requires that almost all active firm types hire only one worker type. We proceed by contradiction. Assume this were not true, i.e., for any type y in a set of active firm types $\mathcal{Y}'$ with strictly positive measure it holds that the support of its labor demand $\mathcal{L}^y$ contains more then one element. Assortativeness still means that for almost all active firm types with $y > y'$ it holds that x is in the support of $\mathcal{L}^y$ and x' is in the support of $\mathcal{L}^{y'}$ only if $x \geq x'$. In particular, this applies also to almost all types in $\mathcal{Y}'$. So for a non-generic $y \in \mathcal{Y}'$ with x and $\tilde{x} > x$ in the support of its labor demand $\mathcal{L}^y$ the following has to hold: almost all firms with higher types have labor demands that only place support on types above $\tilde{x}$, while almost all firm types below have labor demands that only place support on types below x. Therefore, labor demand for all worker types in interval $(x, \tilde{x})$ has measure zero. But the supply of workers in this set has strictly positive mass since the type distributions have non-zero densities. Market clearing then implies that wages are almost everywhere zero for the types in $(x, \tilde{x})$, meaning that these types are inactive which violates the previous paragraph.

Finally, given a unique choice $x = v(y)$, there exists a unique choice of intensity (or firm size): Since optimality requires solving (6), and this problem is strictly concave in intensity, there is a uniquely optimal intensity choice. ■

7.3 Derivations for Assortative Matching Omitted in the Text

Here we lay out the derivations that follow from the firm's maximization problem (6) in Lemma 1 to the sorting conditions that are built up towards Proposition 1.

Maximization (6) gives rise to first order conditions (7) and (8). The second order condition for optimality requires the Hessian $\mathbf{H}$ to be negative definite, where:

$$\mathbf{H} = \begin{pmatrix} f_{\theta\theta} & f_{x\theta} - w'(x) \\ f_{x\theta} - w'(x) & f_{xx} - \theta w''(x) \end{pmatrix}.$$

This requires $f_{\theta\theta}$ to be negative and the determinant $|\mathbf{H}|$ to be positive, or

$$f_{\theta\theta}[f_{xx} - \theta w''(x)] - (f_{x\theta} - w'(x))^2 \geq 0. \tag{12}$$

We can differentiate (7) and (8) with respect to the worker type to get

$$\begin{aligned} f_{x\theta} - w'(x) &= -\mu'(x) f_{y\theta} - \theta'(x) f_{\theta\theta} && (13)\\ f_{xx} - \theta(x) w''(x) &= -\mu'(x) f_{xy} - \theta'(x)\left[f_{x\theta} - w'(x)\right]. && (14) \end{aligned}$$

In the following three lines we successively substitute (13), (14) and then (8) into condition (12):

$$\begin{aligned} -\mu'(x) f_{\theta\theta} f_{xy} - \left[\theta'(x) f_{\theta\theta} + f_{x\theta} - w'(x)\right]\left[f_{x\theta} - w'(x)\right] &> 0 \\ -\mu'(x) f_{\theta\theta} f_{xy} + \mu'(x) f_{y\theta}\left[f_{x\theta} - w'(x)\right] &> 0 \\ -\mu'(x)[f_{\theta\theta} f_{xy} - f_{y\theta} f_{x\theta} + f_{y\theta} f_x/\theta] &> 0 \end{aligned}$$

Since $f_{\theta\theta} < 0$ we can devide by $-f_{\theta\theta}$ to get the condition reported in the main body. For strictly positive assortative matching ($\mu'(x) > 0$) it has to hold that the term in square brackets in the last line is negative, for strictly negative assortative matching the term in square brackets in the last line needs to be positive. Focussing on positive assortative matching, and using the relationship in (8), we obtain the condition:

$$f_{\theta\theta} f_{xy} - f_{y\theta} f_{x\theta} + f_{y\theta} f_x/\theta \leq 0. \tag{15}$$

This condition can be summarized more conveniently in terms of the original function $F(x, y, r, s)$, for which we know that $F(x, y, \theta, 1) = f(x, y, \theta)$. The following relationships will also prove useful. Homogeneity of degree one of F in l and r implies that $-F_{lr} = \theta F_{ll}$. Since F is constant returns, so is F_x.[32] A standard implication of constant returns it then $F_x(x, y, \theta, 1) = \theta F_{xl} + F_{xr}$. We can now rewrite (15) in terms of $F(x, y, \theta, 1)$ and rearrange to obtain the following cross-margin-complementarity condition:

$$
\begin{aligned}
& F_{ll}F_{xy} - F_{yl}\left[F_{xl} - F_x/\theta\right] \leq 0 && (16)\\
\Leftrightarrow\quad & F_{ll}F_{xy} + F_{yl}F_{xr}/\theta \leq 0 &&\\
\Leftrightarrow\quad & F_{xy}F_{lr} \geq F_{yl}F_{xr}. && (17)
\end{aligned}
$$

Since $F_{lr} > 0$ we can divide through to obtain inequality (9) in Proposition 1. This derivation provides a necessary condition for assortative matching for the specific conditions with increasing output and non-zero type densities. It does not deal with distributions that have mass points, nor does it provide sufficient conditions to rule out the existence of other equilibria, for example those where the bottom half of workers matches positively assortatively with the top half of firms and the top half of workers matches positively assortatively with the bottom half of firms. The following proof of Proposition 1 accounts for these cases.

7.4 Proof of Proposition 1

Proof. Part I: sufficiency. Focus on sufficiency for positive assortative matching. (The same logic applies to negative assortative matching.) We need to prove that condition (9) is sufficient to rule out any equilibria that are not positive assortative. This part of the proof relies on the first welfare theorem. Since we have quasi-linear utility, Pareto optimality requires output maximization. A feasible collection of labor demands $\mathcal{L} = \{\mathcal{L}^y\}_{y\in\mathcal{Y}}$ and resource allocations $\mathcal{R} = \{\mathcal{R}^y\}_{y\in\mathcal{Y}}$ for all firm types yields aggregate output

$$
S(\mathcal{L}, \mathcal{R}) = \int_{y\in\mathcal{Y}} \int_{x\in\mathcal{X}} F(x, y, \theta^y(x), 1) d\mathcal{R}^y dH^f, \tag{18}
$$

where $\theta^y = d\mathcal{L}^y/d\mathcal{R}^y$. The first welfare theorem implies that the equilibrium $(\mathcal{L}^*, \mathcal{R}^*)$ combination yields a weakly higher aggregate output than any other feasible $(\mathcal{L}, \mathcal{R})$ combination. In the following we will show that if (9) holds strictly, then any allocation $(\mathcal{L}, \mathcal{R})$ that is not positive assortative can be improved upon by some (positive assortative) reallocation of workers that improves aggregate output, and therefore $(\mathcal{L}, \mathcal{R})$ cannot be an equilibrium.

Assume that $F_{xy} > F_{xr}F_{yl}F_{lr}$ for all $(x, y, l, r) \in R^4_{++}$ but an equilibrium allocation $(\mathcal{L}, \mathcal{R})$ is not positive assortative. The lack of positive sorting implies that there exist two combinations (x_1, y_1, θ_1) and (x_2, y_2, θ_2) with $x_1 > x_2$ but $y_1 < y_2$, that have strictly positive probability: i.e., for any ε there is a strictly positive measure of firm types in any $\varepsilon-$neighborhood around y_i with labor demands whose support includes worker types in an $\varepsilon-$neighborhood around x_i that get resources at intensity in an $\varepsilon-$neighborhood around θ_i. These firms are active as otherwise the support of their labor demand would be empty, and for active firms optimality requires a strictly positive intensity, so we can focus on combinations with $\theta_i > 0$. The rest of the proof will proceed by assuming that a mass of workers of type x_i is employed by firms y_i and receives resources with intensity θ_i, and we will show that assigning some of the low type workers to the low type firms while assigning some of the high type worker to the high type firms *strictly* increases output, yielding a contradiction to the First Welfare Theorem. If this is the case, then the same argument holds if the mass of workers is not at (x_i, y_i, θ_i) but in its neighborhood, since for a small enough neighborhood the output is arbitrarily close to the output that arises if all mass were concentrated only on the exact point, by continuity of F. We proceed in two steps. Step 1 has the key insight.

[32] It holds that $F(x, y, l, r) = rF(x, y, l/r, 1)$, so differentiation implies that $F_x(x, y, l, r) = rF_x(x, y, l/r, 1)$

1. Establish the marginal benefit from assigning additional workers to some resource type: Consider some (x, y, θ) such that a total measure r of resources are deployed in this match (where r is the product of the number of firms and their internal resources deployed to x workers at intensity θ). To achieve this intensity, they are obviously paired with the appropriate number of workers (of measure θr). As a preliminary step to the variational argument that follows, we are interested in the marginal benefit of pairing an additional measure r' of resources of type y' firms with workers of type x. The optimal output is generated by withdrawing some optimal measure $\theta' r'$ of the workers that were supposed to be working with resources of type y and reassigning them to work with resources of type y'. The joint output at (x, y) and (x, y') in (18) is given by

$$rf(x, y, \theta - \theta' r'/r) \; + \; r' f(x, y', \theta'). \tag{19}$$

Optimality of θ' requires, according to the first order condition, that $f_\theta(x, y, \theta - \theta' r'/r) = f_\theta(x, y', \theta')$, which shows that the optimal θ' is itself a function of r'. Denote $\beta(y'; x, y, \theta)$ the marginal increase of (19) from increasing r', evaluated at $r' = 0$. It is given by

$$\beta(y'; x, y, \theta) = f(x, y', \theta') - \theta' f_\theta(x, y', \theta') \tag{20}$$

$$\text{where } \theta^{'} \text{ is determined by } \; f_\theta(x, y', \theta') = f_\theta(x, y, \theta). \tag{21}$$

The **constraint** (21) reiterates the optimality of θ'. Sometimes we will write $\theta'(y'; x, y, \theta)$ to highlight that θ' is a function of y', y, x and θ. The cross-partial β_{xy} of the marginal benefit in (20) with respect to x and y' is strictly positive, evaluated at $y' = y$, iff

$$f_{xy} > -\left[\theta f_{y\theta} f_{x\theta} + f_{y\theta} f_x\right] / \left[\theta f_{\theta\theta}\right],$$

i.e., exactly when our cross-margin condition holds (see (15)). Therefore, it is optimal to assign higher buyers to higher sellers locally around (x, y). This is at the heart of the argument. The next step simply extends this logic to a global argument where y' might be far away from y.

2. Not PAM has strictly positive marginal benefits from matching the high types: We started under the assumption that matching is not assortative, so that x_1 is matched to y_1 at intensity θ_1 and x_2 to y_2 at intensity θ_2, but $x_1 > x_2$ while $y_1 < y_2$. For this to be efficient, it must be more efficient to pair the last unit of resources of type $y' = y_1$ to workers with combination (x_1, y_1, θ_1) than we workers that are otherwise matched at (x_2, y_2, θ_2) :

$$\beta(y_1; x_2, y_2, \theta_2) \leq \beta(y_1; x_1, y_1, \theta_1), \tag{22}$$

where $\beta(\cdot; \cdot, \cdot, \cdot)$ was defined in (19). Similarly, the marginal gains from pairing the last unit of resources of type $y' = y_2$ to workers otherwise matched at (x_2, y_2, θ_2) than to workers matched at (x_1, y_1, θ_1) :

$$\beta(y_2; x_2, y_2, \theta_2) \geq \beta(y_2; x_1, y_1, \theta_1). \tag{23}$$

We will show that if (22) holds, then (23) cannot hold, which yields the desired contradiction. We will show this by proving that the benefit $\beta(y'; x_1, y_1, \theta_1)$ on the right hand side of (22) and (23) always remains above the benefit $\beta(y'; x_2, y_2, \theta_2)$ on the left hand side, for all y'. By (22) this has to be true at $y' = y_1$, and we will show that it remains true when we move to higher y'. The marginal increase of β with respect to its first argument y' is given by

$$\beta_1(y'; x, y, \theta) = f(x, y', \theta'), \tag{24}$$

where θ' is again determined as in (21). Assume there is some $y'' \geq y_1$ such that marginal benefits are equalized, i.e., $\beta(y''; x_2, y_2, \theta_2) = \beta(y''; x_1, y_1, \theta_1)$. We obtain a contradiction if we can show that at such a point $\beta_1(y''; x_2, y_2, \theta_2) < \beta_1(y''; x_1, y_1, \theta_1)$, since this implies that whenever the marginal benefits are nearly equalized the right hand side rises faster than the left hand side.

The inequality $\beta_1(y''; x_2, y_2, \theta_2) < \beta_1(y''; x_1, y_1, \theta_1)$ is by (24) equivalent to $f(x_2, y'', \theta_2') < f(x_1, y'', \theta_1')$, where $\theta_1' = \theta'(y''; x_1, y_1, \theta_1)$ and $\theta_2' = \theta'(y''; x_2, y_2, \theta_2)$ as in (21). To show this, define $\xi(x)$ for all x in resemblance of (20) by the following equality

$$f(x, y'', \xi(x)) - \xi(x) f_3(x, y'', \xi(x)) = \beta(y''; x_2, y_2, \theta_2),$$

which implies $\xi(x_2) = \theta_2'$ and $\xi(x_1) = \theta_1'$ by equality of the marginal benefits at y'', i.e., by $\beta(y''; x_2, y_2, \theta_2) = \beta(y''; x_1, y_1, \theta_1)$. Differentiating $f(x, y'', \xi(x))$ with respect to x reveals that it is strictly increasing exactly under our strict inequality $f_{\theta\theta} f_{xy} - f_{y\theta} f_{x\theta} + f_{y\theta} f_x / \theta < 0$. This in turn implies $f(x_2, y'', \theta_2') < f(x_1, y'', \theta_1')$. This establishes that output can be improved by pairing types positively assortatively, which proofs sufficiency.

Part II: necessity. We need to show that (9) is necessary to have PAM under any distribution of types. That is, if it is not true that (9) holds for all (x, y, l, r), then there will be a type distribution for which PAM will not be an equilibrium. Assume that (9) fails at some (x', y', l'', r''). By continuity it also fails at some (x', y', l', r') with $l' > 0$ and $r' > 0$ sufficiently close to (x', y', l'', r''). Then it also fails at $(x', y', \theta', 1)$ for $\theta' = l'/r'$. By continuity, this means that $F_{xy} F_{lr} < F_{yl} F_{xr}$ for all $(x, y, \theta, 1) \in \mathcal{N}$, where $\mathcal{N}$ is a small enough open neighborhood of $(x', y', \theta', 1)$. If we can restrict the equilibrium allocation to lie in $\mathcal{N}$, then by the analogy of the preceding section for negative assortative matching we know that matching can only be negative assortative, and therefore (9) cannot fail if we want to obtain positive assortative matching. Since we want to ensure positive assortative matching for all type distributions, we can choose the support of x and y within this neighborhood. But since θ is endogenous, this requires slightly more work. Assume that $X = [x', x' + \varepsilon]$ and $Y = [y', y' + \varepsilon]$, and uniform type distributions with mass $H_w^\varepsilon(x' + \varepsilon) = \theta'$ and $H_f^\varepsilon(y' + \varepsilon) = 1$. For small enough ε', firms make nearly identical profits. Since they can only match with nearly identical types, identical profits require them to have nearly identical factor ratios $\theta(x)$. These have to be close to the average ratio in the population. Therefore, for ε small enough all matches lie in $\mathcal{N}$, which rules out that matching can be positive assortative for all type distributions if (9) fails. ■

7.5 The Non-Homogeneous Production Technology

Let output of the firm be $F(x, y, r, s)$, and the firm of type y chooses the worker type and the labor intensity l. As before, let the capital intensity r be given, but we no longer require constant returns to scale in the quantity dimensions. Then the problem of a firm that chooses exactly one type x is

$$\max_{\tilde{x}, \tilde{l}} F(\tilde{x}, y, \tilde{l}, r) - \tilde{l} w(\tilde{x}) - r v(y).$$

The first order conditions for optimality are

$$\begin{aligned} F_x(x, \mu(x), l, r) - l w'(x) &= 0 \\ F_l(x, \mu(x), l, r) - w(x) &= 0 \end{aligned}$$

where $\mu(x)$ and l are the equilibrium values. The second order condition of this problem requires the Hessian $\mathbf{H}$ to be negative definite:

$$\mathbf{H} = \begin{pmatrix} F_{xx} - lw'' & F_{xl} - w' \\ F_{xl} - w' & F_{ll} \end{pmatrix}$$

which requires that all the eigenvalues are negative or equivalently, $F_{xx} - lw'' < 0$ (which follows from concavity in all the arguments (x, y, l, r)), and

$$\begin{vmatrix} F_{xx} - lw'' & F_{xl} - w' \\ F_{xl} - w' & F_{ll} \end{vmatrix} > 0.$$

After differentiating the two FOCs along the equilibrium allocation to substitute for $F_{xx} - lw'' = -F_{xy}\mu'$ and $F_{xl} - w' = -F_{yl}\mu'$ and also using the first FOC to rewrite $w' = F_x/l$ we get

$$\begin{vmatrix} -F_{xy}\mu' & -F_{yl}\mu' \\ F_{xl} - w' & F_{ll} \end{vmatrix} > 0$$

or $-F_{xy}F_{ll}\mu' + (F_{xl} - F_x/l)F_{yl}\mu' > 0$ and thus PAM requires (knowing that $F_{ll} < 0$)

$$F_{xy} > \frac{(F_x/l - F_{xl})F_{yl}}{|F_{ll}|}. \tag{25}$$

Observe that this condition is similar to the one we obtained for the homogeneous case, only that now it depends on the marginal product F_x and the concavity of F in l, F_{ll}.[33] Finally, it is easy to show that the firm finds it strictly optimal to indeed concentrate all its resources on one worker type if F has increasing returns to scale in l and r. With decreasing returns to scale this is not true, and one has to additionally impose the restriction that firms can only hire one worker type to use our methodology.

7.6 Misallocation Debate

This section supplies additional material for the discussion on misallocation in Section 4. Our illustration is based on Adamopoulos and Restuccia (2014), who use the following production function $\tilde{f}(x, y, l, k) = a(\eta(xk)^\rho + (1-\eta)(yl)^\rho)^{\frac{\gamma}{\rho}}$ with $x = 1$ and parameters η, ρ, γ and productivity $a = Ap_a\kappa$, where A is aggregate productivity, p_a is the output price, and κ is a scalar that can distinguish developing and developed countries. They determine the rental rate R of generic capital within a larger multi-sector model, we take it as given for our illustration that only focusses on the agricultural sector. We use their values for their case with different aggregate factors but without distortions, as summarized in Table 3.

The distribution of farmer skill is assumed to be a lognormal with the parameters specified above. Average land per capita plays a role as a scaling factor for the distributions of x and y. Our only adjustment is to round the elasticity of substitution to $\rho = 0.25$ from their original value of 0.24, which allows us to use third degree polynomials to calculate the optimal capital. Figure 7 shows that this does not matter much for our ability to replicate their firm size distribution in developed and developing countries within our matching setup with nearly identical workers ($x \approx 1$).

Defining $\tilde{F}(x, y, l, r, k) = r\tilde{f}(x, y, l/r, k)$ and $F(x, y, l, r) = \max_k \tilde{F}(x, y, l, r, k) - Rkr$, we can either use (9) directly to see whether sorting is PAM. Alternatively we can use the envelop condition developed

[33] The condition for sorting here depends on F_{xl} which is not the case in condition (9). Of course, there are transformations of (9) that include different derivatives (e.g. F_{xl}), obviously with a less concise and intuitive interpretation.

Table 3: Parameters

Parameter	US Baseline (Rich)	Less A (Poor)	Source
TFP (A)	1.0	0.3987	Normalization
Price of agricultural good (p_a)	0.3159	0.5209	Calibration
Rental price of Capital (R)	0.13099	0.3958	Equilibrium
Average and per capita (L/N)	169.249	19.595	Data
κ	1.0	1.0	Normalization
ρ	1/4	1/4	Calibration
η	0.89	0.89	Calibration
Mean farmer skill (μ_y)	-1.8316	-1.8316	Calibration
Std farmer skill (σ_y)	4.6553	4.6553	Calibration

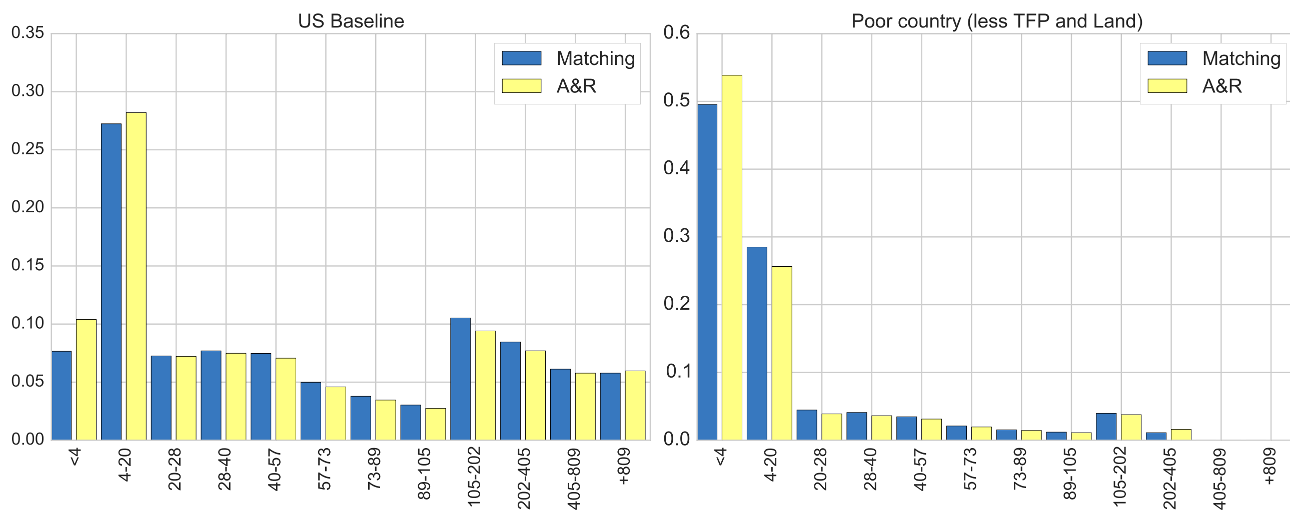


Figure 7: Replication of the firm size distribution in Adamopoulos and Restuccia (2014) with our computational algorithm with negligible spread in x and $\rho = .25$.

in Extension II below in this Appendix and PAM arises if

$$\begin{aligned} \{F_{xy}F_{lr}F_{kk} - F_{xy}F_{lk}F_{rk} - F_{xk}F_{yk}F_{lr}\} - \{F_{xr}F_{yl}F_{kk} - F_{xr}F_{yk}F_{lk} - F_{xk}F_{yl}F_{rk}\} &> 0 \\ \Leftrightarrow \frac{0.00272284375 A^3 p_a^3 r^2 (kx)^{0.5} \left(\frac{ly}{r}\right)^{0.25}}{k^2 lxy} \left(0.89\,(kx)^{0.25} + 0.11 \left(\frac{ly}{r}\right)^{0.25}\right)^{3.0} &> 0, \end{aligned}$$

so that it holds for the chosen parameters in both the rich and the poor country as long as x, y, l, r, k are positive. We find the equilibrium by first finding the optimal generic capital level given x, y, l, r and for computational ease we approximate it by a high-order polynomial. Substituting this out to obtain $F(x, y, l, r)$, we can use the first two equations in (10) as a differential equation system in $\mu(x)$ and $\theta(x)$. We use truncated distributions on both sides, and know that the top agents are matched. The other end-point condition is that the lowest types are matched if all agents can get positive payoffs, otherwise the cutoff type at the side that is not fully matched makes zero payoffs. We use a shooting algorithm to hit the end-point conditions along the equilibrium path.

When we spread the type distribution for land, we use a truncated lognormal distribution with $\mu_x = 1$ and $\sigma_x = 0.2$. That is, the mean of land quality is still 1. We increase the spread by increasing the distance between the truncation points. The actual algorithm is solved on a grid. Supplementary material with construction and code are available.

7.7 Extension I. Frictions and Involuntary Unemployment

Frictional unemployment is an important aspect in the study of labor markets. Moreover, in recent years substantially more has been understood about both the determinants of unemployment across heterogeneously skilled agents in the presence of sorting (amongst others Shimer and Smith (2000), Eeckhout and Kircher (2010)) and about how unemployment varies across firms of different sizes (Smith (1999), Hawkins (2011), Kaas and Kircher (2015), Menzio and Moen (2010); Garibaldi and Moen (2010)). Yet, little is known about how unemployment varies in the presence of sorting *and* variation in firm size jointly.

The sorting framework that we laid out in the previous section is well-suited to capture multi-worker firms with decreasing returns in production. In this section we embed a a costly recruiting and search process in the previous setup in order to capture the hiring behavior of large firms. This setup builds on the directed search literature (e.g., Peters (1991); Acemoglu and Shimer (1999); Burdett, Shi, and Wright (2001); Shi (2001); Shimer (2005); Guerrieri, Shimer, and Wright (2010)), now with sorting of heterogeneous agents and large firms. As in the previous literature, we assume for simplicity that workers and firms are risk-neutral.

Consider a situation where the workers are unemployed and can only be hired by firms via a frictional hiring process. As part of this process, each firm of type y decides how many vacancies $v^y(x)$ to post for each unit of resources devoted to workers of type x that it wants to hire. Posting $v^y(x)$ vacancies has a linear cost $cv^y(x)$. It also decides to post wage $\omega^y(x)$ for this worker type. Observing all vacancy postings, workers decide where to search for a job. Let $q^y(x)$ denote the "queue" of workers searching for a particular wage offer, defined as the number of workers per vacancy. Frictions in the hiring process make it impossible to fill a position for sure. Rather, the probability of filling a vacancy is a function of the number of workers queueing for this vacancy, denoted by $m\left(q^y(x)\right)$, which is assumed to be strictly increasing and strictly concave. Since there are $q^y(x)$ workers queueing per vacancy, the workers' job-finding rate for these workers is $m(q^y(x))/q^y(x)$. The job finding rate is assumed to be strictly decreasing in the number of workers $q^y(x)$ queueing per vacancy. Firms can attract workers to their vacancies as long as these workers get in expectation their equilibrium utility, meaning that $q^y(x)$ adjusts depending on $\omega^y(x)$ to satisfy: $\omega^y(x)m(q^y(x))/q^y(x) = w(x)$. Note the difference between the wage $\omega^y(x)$ which is paid when a worker is actually hired, and the expected wage $w(x)$ of a queueing worker who does not yet know whether he will be hired or not. In equilibrium the firm takes the latter as given because this is the utility that workers can ensure themselves by searching for a job at other firms, while the former is the firm's choice variable with which it can affect how many workers will queue for its jobs. Therefore, a firm maximizes instead of (3) the new problem

$$\max_{\mathcal{R}^y,\theta^y,\omega^y} \int \left[F(x,y,\theta^y,1) - \theta^y\omega^y(x) - v^y(x)c\right] d\mathcal{R}^y(x) \tag{26}$$
$$\text{s.t. } \theta^y(x) = v^y(x)m(q^y(x)); \quad \text{and} \quad \omega^y(x)m(q^y(x))/q^y(x) = w(x)$$

and $\mathcal{R}^y(x)$ integrates to unity. The first line simply takes into account that the firm has to pay the vacancy-creation cost, and that the number of hires depends on the amount of hiring per vacancy which is in turn related to the wage that it offers. There are two equivalent representations of this problem that substantially simplify the analysis. It can easily be verified that problem (26) is mathematically equivalent to both of the following two-step problems:

1. Let $G(x, y, s, r) = \max_v [F(x, y, rvm(s/vr), r) - rvc]$, and solve $\max_{s^y(x), \mathcal{R}^y(x)} \int [G(x, y, s(x), 1) - w(x)s(x)]d\mathcal{R}^y(x)$ where $\mathcal{R}^y(x)$ integrates to unity.
2. Let $C(l, x) = \min_{v,q}[cv + vqw(x)]$ s.t. $l = vm(q)$, and solve $\max_{\theta(x), \mathcal{R}^y(x)} \int [F(x, y, \theta(x), 1) - C(\theta(x), x)]d\mathcal{R}^y(x)$ where $\mathcal{R}^y(x)$ integrates to unity.

In the first equivalent formulation, the firm attracts "searchers" $s^y(x)$, who queue up to get jobs at this firm. To these s searchers it allocates vr vacancies, so the queue length that is relevant in the matching function is $s/(vr)$. In order to entice searchers to come to the firm, it has to offer in expectation a wage $w(x)$ to them, whether or not they actually get hired. The definition of G then reflects the fact that the firm can still decide how many vacancies to create for these workers. If the firm creates more vacancies, searchers have an easier time finding a vacancy suitable to them, and this increases the amount of actual labor that is employed within the firm. In the second formulation the firm maximizes the output minus the costs of hiring the desired amount of labor. The costs include both the vacancy-creation costs as well as the wage costs, where again the expected wage has to be paid to all workers that are queueing for the jobs. Writing the problem in terms of G and C, respectively, has two direct consequences:

Problem 1 has the beauty that G is fully determined by the primitives, and can be directly integrated into the framework we laid out in Section 2 (where now G replaces F). The firm can be viewed as if it hires "searchers" who have to be paid their expected wage. Applying the machinery from the previous section allows us to assess whether sorting is assortative, and what the expected wages $w(x)$ are that are paid in equilibrium. We take this formulation embedded in the equilibrium definition of the previous section as the definition of a competitive search equilibrium with large firms. This allows the application of our sorting condition (9) to G:

Proposition 3 *Even in the framework with directed search frictions, a necessary condition to have equilibria with positive assortative matching under any arbitrary distribution of types is that inequality (9) holds for all $(x, y, l, r) \in \mathbb{R}^4_{++}$. With a strict inequality, it is also sufficient to ensure that any equilibrium entails positive assortative matching. The opposite inequality provides a necessary and sufficient condition for negative assortative matching.*

Proof. Let $v^*(x, y, s, r)$ maximize $F(x, y, rvm(s/vr), r) - rvc$. Also, define $V^*(x, y, s, r)$ as the maximizer of $F(x, y, Vm(s/V), r) - Vc$ with respect to V. Given our assumptions on the production and matching function, it is easy to show that V^* is unique and determined by the appropriate first order condition, which in turn implies that it is differentiable by the implicit function theorem. Clearly $V^*(x, y, s, r) = rv^*(x, y, s, r)$ and we can write $G(x, y, s, r) = F(x, y, V^*(x, y, s, r)m(s/V^*(x, y, s, r)), r) - V^*(x, y, s, r)c$. We obtain the first order conditions

$$
\begin{aligned}
G_y &= F_y(x, y, V^*(x, y, s, r)m(s/V^*(x, y, s, r)), r) \\
G_r &= F_r(x, y, V^*(x, y, s, r)m(s/V^*(x, y, s, r)), r)
\end{aligned}
$$

where we drop arguments from the equation whenever there is no possibility of confusion. The arguments related to $\partial V^*/\partial y$ and $\partial V^*/\partial r$ do not appear because of the envelop condition. Cross-partial derivatives are

$$
\begin{aligned}
G_{xy} &= F_{xy} + F_{yl}\frac{\partial(V^*m(s/V^*))}{\partial V^*}\frac{\partial V^*}{\partial x} \\
G_{sr} &= F_{lr}\left[\frac{\partial(V^*m(s/V^*))}{\partial s} + \frac{\partial(V^*m(s/V^*))}{\partial V^*}\frac{\partial V^*}{\partial s}\right] \\
G_{ys} &= F_{yl}\left[\frac{\partial(V^*m(s/V^*))}{\partial s} + \frac{\partial(V^*m(s/V^*))}{\partial V^*}\frac{\partial V^*}{\partial s}\right] \\
G_{xr} &= F_{xr} + F_{rl}\frac{\partial(V^*m(s/V^*))}{\partial V^*}\frac{\partial V^*}{\partial x}.
\end{aligned}
$$

Now the sorting condition $G_{xy}G_{sr} \geq G_{ys}G_{xr}$ is equivalent to the condition on output $F_{xy}F_{rl} \geq F_{yl}F_{xr}$. ■

Problem 2 then relates the expected wages $w(x)$ that were determined in the previous problem to job finding probabilities of the searchers. Substituting the constraint in Problem 2 into the objective function and taking the first order condition with respect to the queue length yields the main characterization of this section. It can best be expressed by writing the elasticity of the matching probability as $\eta(q) := qm'(q)/m(q)$ and by denoting the queue length that solves the minimization problem by $q(x)$. We then obtain

$$
w(x)q(x) = \frac{\eta(q(x))}{1-\eta(q(x))}c \tag{27}
$$

The right hand side is related to the well-known Hosios condition (Hosios (1990)), which showed that efficient vacancy creation is related to the elasticity of the matching function. The condition becomes particularly tractable in commonly used settings in which the elasticity is constant. In this case the queue length that different workers face is inverse proportional to the expected utility that they obtain in equilibrium. Since better workers obtain higher expected utility $w(x)$ as determined in Problem 1 (otherwise a firm could higher better workers at equal cost), they face proportionally lower competition for each job and correspondingly higher job finding probabilities. This arises because the opportunity costs of having high skilled workers unsuccessfully queue for employment is higher, and therefore firms are more willing to create enough vacancies to enable most of these applicants to actually get hired for the job. The logic applies even if the elasticity is not constant:

Proposition 4 *Assume higher worker types create more output ($F_x > 0$). In the competitive search equilibrium with large firms, higher skilled workers have lower unemployment rates.*

Proof. The term $\eta(q)/[q(1-\eta(q))] = m'(q)/[m(q) - qm'(q)]$. This term is strictly decreasing in q, since the numerator is strictly decreasing and the denominator is strictly increasing in q. Since output at any firm is increasing in skill ($F_x > 0$) it follows immediately that in any equilibrium $w(x)$ is increasing in x. Implicit differentiation of (27) implies that $q(x)$ is decreasing, which in turn implies that the chances of finding employment are increasing in x. ■

The reason for this result is that the opportunity cost of an unfilled vacancy is linked to the cost of creating another vacancy, and this cost is identical for all firms. This differs from settings with one-to-one matching (e.g., Shi (2001), Shimer (2005), Eeckhout and Kircher (2010)) where the opportunity cost of not filling the vacancy means loss of production, which is type-dependent and can reverse this insight.

Interestingly, the finding in Proposition 4 implies that under positive assortative matching the firm-size can be increasing in firm type even though the number of workers that apply for jobs is decreasing.

This can be seen mathematically as follows. The amount of labor that is actually hired, $l(x)$, relates to the actual number of searchers and their queue per vacancy as $l(x) = s(x)m(q(x))/q(x)$, where we omitted the superscript for the firm type ($y = \nu(x)$) that hires this worker type. This relationship implies:

$$l'(x) = s'\frac{m}{q} + s\frac{m'q - m}{q^2}q'.$$

The change in the number of searchers (s') is determined by (10) under appropriate change of variables (θ and f replaced by s and g). Even if the number of workers that search for employment at better firms is not increasing, the number of hires might still be increasing because the second term is strictly positive. This is due to the fact that high productivity firms put more resources into creating jobs for their high-skilled applicants. If a firm tries to attract workers for whom their time-constraints make it very costly to apply, it will invest resources to make sure that the applicants perceive a sufficiently high probability that they will find a suitable appointment in the hiring process. In this model this is captured through creating a sufficient number of different vacancies.

In contrast to the finding of monotonicity for the hiring probability across different workers, the vacancy rate across firms of different sizes is ambiguous.

Proposition 5 *The vacancy rate (v/l) is increasing in firm productivity (y) und PAM and decreasing under NAM. It ambiguous in firm size.*

Proof. Consider PAM (likewise for NAM). The vacancy rate (v/l=$1/m(q)$) is increasing in x, and under PAM then also in y. However, from Proposition 1, firm size ambiguous in y. In particular, it is increasing if $G_{yl} \geq G_{xr}$ and decreasing if $-G_{yl} \leq G_{xr}$. ■

This result immediately stems from the fact that firm size in general is ambiguous in firm type y.

7.8 Extension II. Capital Investment

As in the discussion on the misallocation debate, consider a production process that not only takes as inputs the amount of labor and of proprietary firm resources, but also some amount k of a generic capital good, and creates output $\hat{F}(x, y, l, r, k)$. The generic capital can be bought on the world market at price i per unit. Optimal use of resources requires $F(x, y, l, r) = \max_k \left[\hat{F}(x, y, l, r, k) - ik\right]$, where F is constant returns in its last two arguments if $\hat{F}$ is constant returns in its last three arguments. Rewriting the cross-margin-complementarity condition (9) in terms of the new primitive yields the following condition for positive assortative matching: $\hat{F}_{xy}\hat{F}_{lr}\hat{F}_{kk} - \hat{F}_{xy}\hat{F}_{lk}\hat{F}_{rk} - \hat{F}_{xk}\hat{F}_{yk}\hat{F}_{lr} \geq \hat{F}_{xr}\hat{F}_{yl}\hat{F}_{kk} - \hat{F}_{xr}\hat{F}_{yk}\hat{F}_{lk} - \hat{F}_{xk}\hat{F}_{yl}\hat{F}_{rk}$. We expect that particular functional form assumptions for the way that generic capital affects the production process will simplify this condition and make it more amenable for interpretation in specific cases.

7.9 Extension III. Monopolistic Competition

In the previous sections, we analyzed the case where the firm's output is converted one-for-one into agents utility. Therefore, there are no consequences of output on its price, which is normalized to one. An often used assumption in the industrial organization and the trade literature concerns consumer preferences pioneered by Dixit and Stiglitz, which are CES with elasticity of substitution $\rho \in (0, 1)$ among the goods produced by different firms. For these preferences it is well-known that a firm that produces output $\tilde{f}$ achieves sales revenues $\chi\tilde{f}^{\rho}$, where χ is an equilibrium outcome that is viewed as constant from the perspective of the individual firm.[34] The difficulty in this setup is that, despite the fact that output is constant returns to scale in employment and firm resources, the revenue of the firm has decreasing returns to scale. Therefore, we cannot directly apply (9). But if there is assortative matching the firm employs only one worker type, in which case revenues are $f(x,y,l) = \chi\tilde{f}(x,y,l)^{\rho}$, and we can apply (15) directly. If $\tilde{f}(x,y,l)$ is multiplicatively separable and linear in l so that we can write $\tilde{f}(x,y,l) = g(x,y)l$, then our sorting condition reduces to the requirement of log-supermodularity of g, which is a known condition in the trade literature. Our condition also implies insights into the non-separable and non-linear case. Rearranging and using $\tilde{F}(x,y,l,r) = r\tilde{f}(x,y,l/r)$ we get the condition for positive assortative matching

$$\left[\rho\tilde{F}_{xy} + (1-\rho)(\tilde{F})\frac{\partial^2 \ln \tilde{F}}{\partial x \partial y}\right]\left[\rho\tilde{F}_{lr} - (1-\rho)l\tilde{F}\frac{\partial^2 \ln \tilde{F}}{\partial l^2}\right]$$
$$\geq \left[\rho\tilde{F}_{yl} + (1-\rho)\tilde{F}\frac{\partial^2 \ln \tilde{F}}{\partial y \partial l}\right]\left[\rho\tilde{F}_{xr} + (1-\rho)\left(l\tilde{F}_{xl} - l\tilde{F}\frac{\partial^2 \ln \tilde{F}}{\partial x \partial r}\right)\right].$$

Several points are note-worthy. First, the condition is independent of χ, and therefore can be checked before this term is computed as an outcome of the market interaction. Furthermore, for elastic preferences ($\delta \to 1$) the condition reduces to our original condition (9). Otherwise, log-supermodularity also appears in the condition.

7.10 Extension IV. Optimal transportation

Assume it costs $-r \cdot c(x,y)$ to move r units of waste from production site x into destination storage y, and if one attempts to move more units r into any given amount l of storage then there is some probability of damage $d(r/l)$ that each unit that is stored gets destroyed. This leads to function $F(x,y,l,r) = -rc(x,y) - \alpha rd(r/l)$, where α represents the lost revenue because of destruction. Unlike in the standard Monge-Kantorovich transportation problem, storage sites do not have a fixed capacity (except if $d(r/l)$ is zero when r/l is below unity and a very large number if it is above). Rather, more or less can be stored in a given location, but at increasing costs.

7.11 Extension V. Endogenous type distributions, technology choice, team-work

One way to endogenize the type distribution is to assume that there is free entry of firms (free entry of resources in the model), but entry with type y costs $c(y)$. If output increases in y, i.e., $F_2 > 0$, then it is crucial for a meaningful entry decision that $c(y)$ is strictly increasing. If c is strictly increasing and differentiable, and our sorting condition is satisfied everywhere, it is not difficult to construct an equilibrium where profits equal the entry cost $c(y)$ for all active firms. In fact, this formulation is easier to construct: We know that the highest types match, so that $\mu(\bar{x}) = \bar{y}$. The problem is usually how to determine at which ratio they match, i.e., to find $\theta(\bar{x})$. But here it is given simply by the requirement that the profits of the highest firm equals the entry costs. Substituting the first order condition (7) into the objective function yields profit $f(\bar{x},\mu(\bar{x}),\theta(\bar{x})) - \theta(\bar{x})f_\theta(\bar{x},\mu(\bar{x}),\theta(\bar{x}))$, which have to equal $c(\mu(\bar{x}))$. This can be then used together with the first order conditions and the differential equations in (10) to construct the type distribution after entry at all lower types.

[34]The underlying form for the utility function is $U = x_0^{1-\mu}\left(\int c(y)^{\rho}dy\right)^{\mu/\rho}$, where x_0 is a numeraire good and $c(y)$ is the amount of consumption of the good of producer y. Then one obtains $\chi = (\mu Y)^{1-\rho}P^{\rho}$ where Y is the aggregatve income, p_y denotes the price achieved by firm y through its equilibrium quantity, and $P = \left(\int p_y^{\rho/(1-\rho)}\right)^{\rho/(1-\rho)}$ represents the aggregate price index.

More complicated is the analysis when one considers a common pool of workers, some of whom choose to be managers while others choose to remain workers. This is then a teamwork problem, where one team becomes the $y's$ and the other the $x's$. While interesting, we leave this analysis for further work.

7.12 Application QBTC: Additional Results

CDF of the worker skill Distribution

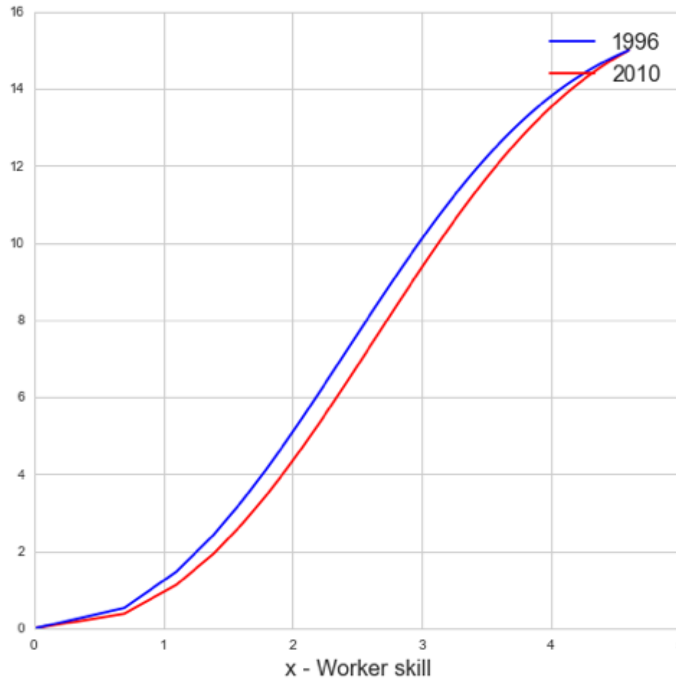
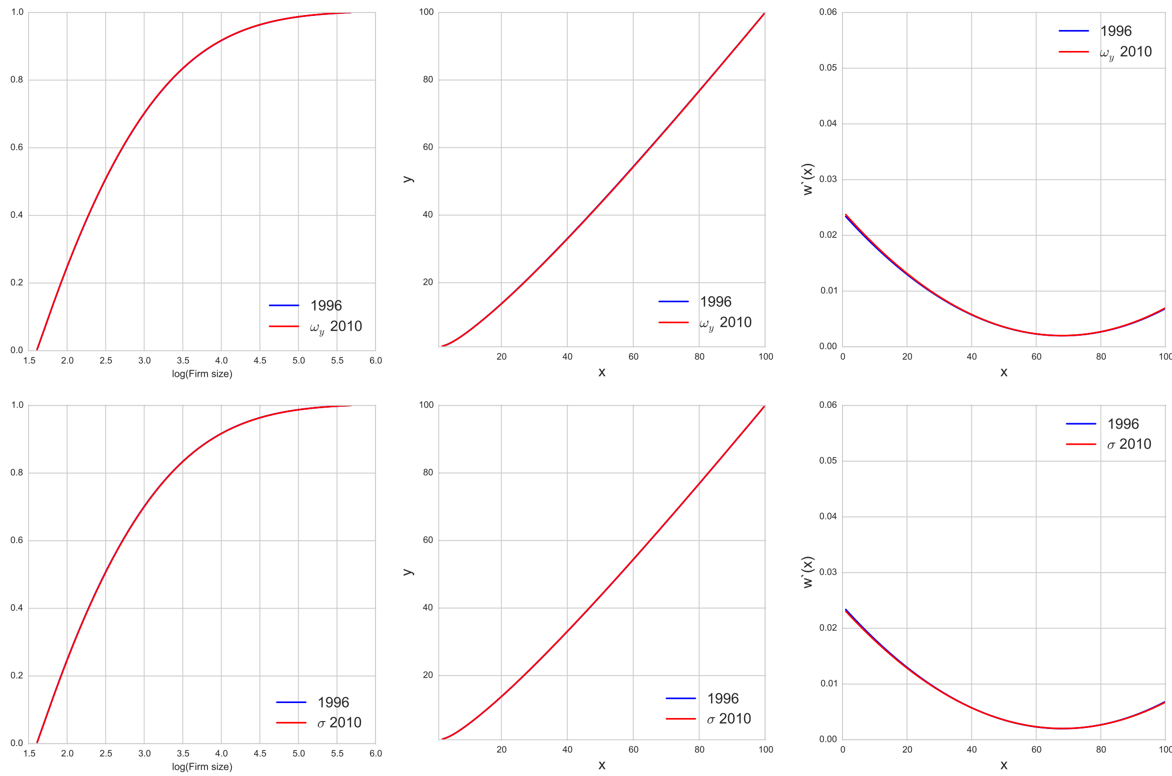


Figure 8: Worker Types x Distribution (CDF in levels).

Counterfactuals: 2010 values for ω_y and σ



(a) Size Distribution (b) Equilibrium Allocation $\mu(x)$ (c) Skill Premium $w'(x)$

Figure 9: Counterfactuals: 1996 economy with one parameter changed to 2010. For each row, we compare the 1996 economy with the 1996 where one parameter has been changed. Row 1: ω_y from 2010; Row 2. σ from 2010. In the first column we report the firm size distribution, in the second column the equilibrium allocation and in the third column the skill premium.

References

Acemoglu, D. (1998): "Why do new technologies complement skills? Directed technical change and wage inequality," *Quarterly Journal of Economics*, pp. 1055–1089.

——— (2002): "Technical Change, Inequality, and the Labor Market," *Journal of Economic Literature*, 40, 7–72.

Acemoglu, D., and R. Shimer (1999): "Holdups and Efficiency with Search Frictions," *International Economic Review*, 40(4), 827–49.

Adamopoulos, T., and D. Restuccia (2014): "The size distribution of farms and international productivity differences," *The American Economic Review*, 104(6), 1667–1697.

Antràs, P., L. Garicano, and E. Rossi-Hansberg (2006): "Offshoring in a Knowledge Economy," *Quarterly Journal of Economics*, 121(1).

Atakan, A. (2006): "Assortative Matching with Explicit Search Costs," *Econometrica*, 74, 667–680.

Autor, David, H., and D. Dorn (2009): "The growth of low skill service jobs and the polarization of the US labor market," Discussion paper, National Bureau of Economic Research.

Bagger, J., and R. Lentz (2016): "An Equilibrium Model of Wage Dispersion and Sorting," Discussion paper, Wisconsin mimeo.

Barth, E., A. Bryson, J. C. Davis, and R. Freeman (2014): "It's where you work: Increases in earnings dispersion across establishments and individuals in the US," Discussion paper, National Bureau of Economic Research.

Becker, G. (1973): "A Theory of Marriage I," *Journal of Political Economy*, 81, 813–846.

Benguria, F. (2015): "Inequality Between and Within Firms: Evidence from Brazil," *Available at SSRN 2694693*.

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

Card, D., J. Heining, and P. Kline (2013): "Workplace Heterogeneity and the Rise of West German Wage Inequality," *The Quarterly Journal of Economics*, 128(3), 967–1015.

Chade, H., J. Eeckhout, and L. Smith (2016): "Sorting Through Search and Matching Models in Economics," *Journal of Economic Literature*, forthcoming.

Costinot, A. (2009): "An Elementary Theory of Comparative Advantage," *Econometrica*, 77(4), 1165–1192.

Eeckhout, J., and P. Kircher (2010): "Sorting and Decentralized Price Competition," *Econometrica*, 78, 539–574.

Eeckhout, J., R. Pinheiro, and K. Schmidheiny (2014): "Spatial Sorting," *Journal of Political Economy*, 122(3), 554–620.

Gabaix, X., and A. Landier (2008): "Why Has CEO Pay Increased So Much?," *Quarterly Journal of Economics*, 123, 49–100.

Garibaldi, P., and E. R. Moen (2010): "Job to job movements in a simple search model," *The American Economic Review P&P*, 100(2), 343–347.

Garicano, L. (2000): "Hierarchies and the Organization of Knowledge in Production," *Journal of Political Economy*, 108(5), 874–904.

Garicano, L., and E. Rossi-Hansberg (2006): "Organization and inequality in a knowledge economy," *Quarterly Journal of Economics*, 121(4), 1383–1435.

Goos, M., and A. Manning (2007): "Lousy and lovely jobs: The rising polarization of work in Britain," *The review of economics and statistics*, 89(1), 118–133.

Grossman, G. M., E. Helpman, and P. Kircher (2016): "Matching and sorting in a global economy," Discussion paper, National Bureau of Economic Research.

Guerrieri, V., R. Shimer, and R. Wright (2010): "Adverse Selection in Competitive Search Equilibrium," *Econometrica*, 78, 1823–1862.

Gul, F., and E. Stacchetti (1999): "Walrasian equilibrium with gross substitutes," *Journal of Economic theory*, 87(1), 95–124.

Guner, N., G. Ventura, and Y. Xu (2008): "Macroeconomic implications of size-dependent policies," *Review of Economic Dynamics*, 11(4), 721–744.

Hatfield, J. W., and P. R. Milgrom (2005): "Matching with Contracts," *American Economic Review*, 95(4), 913–935.

Hawkins, W. (2011): "Competitive Search, Efficiency, and Multi-Workers Firms," mimeo, University of Rochester.

Heckman, J. J., and B. E. Honore (1990): "The Empirical Content of the Roy Model," *Econometrica*, 58(5), 1121–1149.

Hopenhayn, H., and R. Rogerson (1993): "Job Turnover and Policy Evaluation: A General Equilibrium Analysis," *Journal of Political Economy*, 101(5), 915–38.

Hosios, A. J. (1990): "On the efficiency of matching and related models of search and unemployment," *The Review of Economic Studies*, 57(2), 279–298.

Hsieh, C.-T., and P. J. Klenow (2009): "Misallocation and Manufacturing TFP in China and India," *The Quarterly Journal of Economics*, 124(4), 1403–1448.

Kaas, L., and P. Kircher (2015): "Efficient Firm Dynamics in a Frictional Labor Market," *The American Economic Review*, 105(10), 3030–3060.

Kantorovich, L. V. (1942): "On the translocation of masses," in *Dokl. Akad. Nauk SSSR*, vol. 37, pp. 199–201.

Katz, L. F., M. S. Kearney, et al. (2006): "The Polarization of the US Labor Market," *American Economic Review*, 96(2), 189–194.

Kelso, Alexander S, J., and V. P. Crawford (1982): "Job Matching, Coalition Formation, and Gross Substitutes," *Econometrica*, 50(6), 1483–1504.

Khan, M. A., and N. C. Yannelis (1991): "Equilibria in markets with a continuum of agents and commodities," in *Equilibrium theory in infinite dimensional spaces*, pp. 233–248. Springer.

Koopmans, T. C., and M. Beckmann (1957): "Assignment problems and the location of economic activities," *Econometrica: journal of the Econometric Society*, pp. 53–76.

Krusell, P., L. E. Ohanian, J.-V. Ríos-Rull, and G. L. Violante (2000): "Capital-skill complementarity and inequality: A macroeconomic analysis," *Econometrica*, 68(5), 1029–1053.

Lentz, R. (2010): "Sorting by Search Intensity," *Journal of Economic Theory*, 145(4), 1436–1452.

Lindenlaub, I. (2016): "Sorting Multidimensional Types: Theory and Application," Yale mimeo.

Lucas, R. E. (1978): "On the Size Distribution of Business Firms," *Bell Journal of Economics*, 9(2), 508–523.

Lucas, R. E., and E. Rossi-Hansberg (2002): "On the internal structure of cities," *Econometrica*, 70(4), 1445–1476.

Menzio, G., and E. R. Moen (2010): "Worker replacement," *Journal of Monetary Economics*, 57(6), 623–636.

Michalopoulos, S. (2012): "The origins of ethnolinguistic diversity," *The American Economic Review*, 102(4), 1508.

Michalopoulos, S., and E. Papaioannou (2012): "National institutions and subnational development in Africa," Discussion paper, National Bureau of Economic Research.

Ostroy, J. M. (1984): "On the existence of Walrasian equilibrium in large-square economies," *Journal of Mathematical Economics*, 13(2), 143–163.

Peters, M. (1991): "Ex Ante Price Offers in Matching Games Non-Steady States," *Econometrica*, 59, 1425–1454.

Restuccia, D., and R. Rogerson (2008): "Policy Distortions and Aggregate Productivity with Heterogeneous Plants," *Review of Economic Dynamics*, 11(4), 707–720.

Rosen, S. (1982): "Authority, control, and the distribution of earnings," *The Bell Journal of Economics*, pp. 311–323.

Roy, A. D. (1951): "Some thoughts on the distribution of earnings," *Oxford economic papers*, 3(2), 135–146.

Sattinger, M. (1975): "Comparative advantage and the distributions of earnings and abilities," *Econometrica: Journal of the Econometric Society*, pp. 455–468.

——— (1993): "Assignment Models and the Distribution of Earnings," *Journal of Economic Literature*, 31(2), 831–880.

Schaal, E. (2015): "Uncertainty and Unemployment," Discussion paper, NYU Working paper.

Shapley, L., and M. Shubik (1972): "The Assignment Game I: The Core," *International Journal of Game Theory*, pp. 111–130.

Shi, S. (2001): "Frictional assignment. i. efficiency," *Journal of Economic Theory*, 98(2), 232–260.

Shimer, R. (2005): "The Assignment of Workers to Jobs in an Economy with Coordination Frictions," *Journal of Political Economy*, 113(5), 996–1025.

Shimer, R., and L. Smith (2000): "Assortative Matching and Search," *Econometrica*, 68(2), 343–369.

Smith, E. (1999): "Search, concave production, and optimal firm size," *Review of Economic Dynamics*, 2(2), 456–471.

Song, J., D. J. Price, F. Guvenen, N. Bloom, and T. von Wachter (2015): "Firming up inequality," Discussion paper, National Bureau of Economic Research.

Stigler, G. (1961): "The Economics of Information," *Journal of Political Economy*, 69, 213–225.

Tervio, M. (2008): "The Difference that CEOs Make: An Assignment Model Approach," *American Economic Review*, 98(3), 642–668.

Thomas, V., Y. Wang, and X. Fan (2001): *Measuring education inequality: Gini coefficients of education*, vol. 2525. World Bank Publications.

Van Nieuwerburgh, S., and P.-O. Weill (2010): "Why has house price dispersion gone up?," *The Review of Economic Studies*, 77(4), 1567–1606.

Vlachos, J., E. Lindqvist, and C. Hakanson (2015): "Firms and skills: the evolution of worker sorting!," Discussion paper, Stockholm.