Assortative Matching with Large Firms[*]

Jan Eeckhout[†]and Philipp Kircher[‡]

September 1, 2017

Abstract

Two cornerstones of empirical and policy analysis of firms in industrial organization, macro and labor are the determinants of the firm size distribution, and the determinants of sorting between workers and firms. 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, and allows us to characterize the allocation of the quality and quantity of labor to firms of different productivity. We show that our model nests a large number of widely used existing models. We also augment the model to incorporate labor market frictions in the presence of sorting with large firms.

Keywords. Sorting. Large Firms. Span of Control. Firm Size. Complementarities. Supermodularity.

[*] 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. Kircher gratefully acknowledges support from the ERC, grant 184119.

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

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

1 Introduction

Two cornerstones of empirical and policy analysis of firms in industrial organization, macro and labor, are the determinants of the firm size distribution, and the determinants of sorting between workers and firms. Firm size reflects an important aspect of a firm's productivity. Firms that own or invent more productive technologies tend to exploit those advantages by producing and selling more. To that end, they also hire more workers. At the same time, the skill composition of a firm's workforce is crucial for output. Firms' competition for skilled workers leads to sorting of workers of different skills into jobs of different productivity. The literature has treated these two determinants – firm size and sorting – largely independently.[1] What's missing is a tractable framework that allows for the standard firm size choice but also allows thought about sorting heterogeneous workers into such large firms. In this paper we take a first step towards proposing such a framework that connects the two and ask what the key forces are that determine sorting when firm size matters, and what the key forces are that determine the size distribution when sorting matters.

We introduce a model of the firm where the span of control – the number of workers under the control of management within that 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, which has important implications for the size of firms (Lucas (1978); Jovanovic (1982); 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 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

[1] 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.

this framework, we pin down the features of the equilibrium allocation: the sorting pattern, the 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[2] along four margins: (1) *type complementarity* captures the interaction between firm and worker types. Clearly, if better firms receive an 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 standard models with homogeneous labor and resources (or capital). 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, characterizing the efficient equilibrium outcome and measuring 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 between span of control over more workers versus the quality workers. A system of three differential equations governs the equilibrium allocation of types and quantities entirely. In particular, this gives a prediction for the firm's span of control, and therefore, for the firm size distribution and also determines the equilibrium allocation of skills and the wage distribution. This system also makes explicit how firm size interacts with the skill premium, providing a simple mapping between assortative matching and the size distribution of firms. Our conditions tell us when there is Positive Assortative Matching (PAM), when worker types are increasing or decreasing in firm size, and when there is Negative Assortative Matching (NAM), and again, when worker types are increasing or decreasing in firm size.

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.

The major appeal of our model is that it nests a large number of well-established models in the literature. 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 introduce our model, allowing us to combine the discussion of related papers with some simple analytical arguments about

[2] 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).

how these can be represented within our framework. Most notably, we show that Becker's one-to-one matching model is the limit case of a multiplicatively separable version of ours wherein 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 influential papers Witten 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, such as the misallocation debate.

Theoretically, the main contribution of our results is to solve and characterize a model of sorting with many-to-one matches. Moreover, our theory is cast in a framework that is amenable to its application in standard macro, labor and industrial organization models. Testament to the generality this simple model nests a large number of well-known models currently used, yet, we obtain the general formulation while retaining assumptions that render the model tractable and allow for clear insights on sorting and firm size. Conceptually this is novel; we are not aware of anyone having analyzed it. The many-to-one matching model has long been known in the social choice literature, but has typically been analyzed exclusively to show existence in a small economy with finite agents (in particular Kelso and Crawford (1982) and their celebrated gross-substitutes condition). The characterization of the solution of our proposed model offers a set of conditions hitherto unknown. Technically, the derivation of the necessary condition for equilibrium aligns with existing models. What is much more challenging is the sufficiency of the condition. In the standard one-to-one matching model, this is easily satisfied. In this many-to-one matching model with endogenous size, our local sorting condition does not easily integrate to a global condition, which makes the problem substantially more complex.

We also discuss our model in the context of applications to topics of economic relevance, such as mismatch of inputs of production, the skill premium, and economic geography. Existing theories currently discuss heterogeneous inputs only in passing and rely on homogeneity in their formal modeling. We illustrate how heterogeneity can be incorporated and illustrate its importance. Building on the mismatch model of Adamopoulos and Restuccia (2014), we illustrate the role of skill heterogeneity in quantifying the misallocation of resources. And we show how the seminal framework to study the evolution of the skill premium of Krusell, Ohanian, Ríos-Rull, and Violante (2000) can be extended to allow for heterogeneity even within the high and the low skill sector.

Our theory is also amenable to quantitative analysis. In an earlier version of this paper, we quantitatively assessed the theory to analyze the evolution of technology along two dimensions: we parametrized the model to capture "traditional" Skill Biased Technological Change, as well as Quantity Biased Technological Change that allows firms to adjust their size. Using matched employer-employee data from Germany we found that technological change along the second dimension has been important over the last two decades in driving the size distribution. It has also been a mitigating force that limits the impact of skill-biased technological change on wage inequality. These exercises highlight the important equilibrium interplay of sorting and firm size.

Finally, we 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 discuss in Section 4 the special cases that are nested in the model, we review the related literature, and we present extensions that outline how the model might be used in applied settings. In Section 5 we analyze search frictions in the context of sorting with large firms. Section 6 concludes.

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. Heterogeneity exists 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 who spends time 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. In line with the matching literature, we assume that x and y are observable and focus our interest on the equilibrium allocation problem in the presence of complete information. The incomplete information problem remains of interest and importance, but it is beyond the scope of the current paper.

Preferences and Production. A firm of type y that hires l workers of identical type x produces output according to non-negative function $f(x,y,l)$, which is strictly increasing and strictly concave in l. This nests heterogeneous y firms that make intensive margin choices about firm size l familiar from Lucas (1978), Jovanovic (1982), and Hopenhayn and Rogerson (1993), and heterogeneous y firms that make extensive margin choices about worker type x familiar from Becker (1973), enabling us to study their interaction. For the casual reader, without substantial loss of insight, this can serve as a primitive for the subsequent analysis. Nevertheless, we proceed by providing a more micro-founded production structure that in principle allows for multiple worker types within the same firm. This allows us to closely link our study to existing work on worker supervision, to highlight the underlying assumptions that give rise to a uniform workforce in equilibrium, and to eventually represent our results in an even more concise way.

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. To impose discipline on the problem and as is standard in the literature, we assume the technology is common to all firms. What differs are the inputs. We can thus interpret the firm productivity (or TFP) as y. 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 must 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 reasons we assume that F is strictly increasing and strictly concave in each quantity variable in the interior of the type space, no output is produced without resources, and standard Inada conditions apply.[3]

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 imply that the output in (1) can be expressed as the product of the amount of resource r and the output per

[3] 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 $\lim_{l\to 0} F_l(x, y, l, r) = \infty$ for given $x, y, r > 0$, and $\lim_{l\to\infty} F_l(x, y, l, r) = 0$. Similar conditions can be placed on resources.

unit of resource:[4]

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

where $\theta = l/r$ represents the amount of workers per unit of resource, which we call the *intensity.* In line with the interpretation in Lucas (1978), we interpret the intensity or span of control of the manager as the size of the firm, as if we were explicitly talking about the span of control of the CEO. Therefore, in what follows, we use intensity θ and firm size interchangeably. As mentioned earlier, the function $f(x, y, \theta)$ represents the production of a firm that only hires one type of worker, in which case it trivially spends all its resources on this type ($r = 1$). Moreover, when $r > 1$ output $F(x, y, l, r) = rf(x, y, l/r)$ can be viewed as the optimal output of r firms of type y that optimally employ l workers of type y. Because of the tight link between f and F in (2), either can be used as the primitive of the model.

One can interpret the resources r as the time available to the manager. This interpretation follows the notion of span of control in Lucas (1978) and further developed in the literature on knowledge hierarchies (Garicano (2000)). Resources are thus not the equivalent of capital that can be bought on the market. The managerial resources are fixed, and the only choice for management is how to assign those time resources to different skill levels x.[5] We do make clear the distinction between resources r and other generic capital that can be purchased on the market in section 4. There we also discuss an endogenous choice of resources.

The action of a firm y is to choose two distributions, the number of workers of each type and the 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 $\theta^y(x)$ over all resources: $\mathcal{L}^y(x) = \int_{\underline{x}}^{x} \theta(\tilde{x}) d\mathcal{R}^y(\tilde{x})$.[6] We can therefore interchangeably use $(\mathcal{L}^y(\cdot)$,$\mathcal{R}^y(\cdot))$ and $(\theta^y(\cdot), \mathcal{R}^y(\cdot))$ to represent the firm's choices.

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

[5] Because of constant returns to scale (CRS) in quantity variables, an alternative interpretation of the model is one where anyone can combine l units of workers x with r units of managers y to provide output $F(x, y, l, r)$, and workers get paid the equilibrium wage and managers their equilibrium salaries. The operator of F obtains no profits in equilibrium because of CRS and free entry, and all returns go to the scarce resources of worker and managerial time. Our setup can then be interpreted as each manager operating her own firm, in which case the firm profits coincide with the return to management time. While we fixed the time endowment per manager at unity, this could be chosen endogenously. Our sorting conditions would not change as they rely on aggregate amounts of resources in equilibrium, but the discussion of firm size would have to be rescaled by the endogenous choice of resources that each manager acquires.

[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: $\mathcal{L}^y(x) = \int\limits_{(x,\theta):x\leq\tilde{x}} \theta(\tilde{x})d\mathcal{R}^y(\tilde{x})$.

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 allow 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 resources devoted to workers of type x at intensity θ, the output is $f(x, y, \theta)$ but the firm must 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 the density of 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 $l(y) = \theta^y(x)$.

Feasibility of the total allocation of resources requires that firms attempt to hire no more workers than there are in the population. Consider any 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 the implied aggregate demand does not exceed the economy's endowment with such worker types, for all x', x:

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

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*

[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.

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, 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 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 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 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 must 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 the right hand side sums up all workers above x and the left hand side sums up 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 variable firm size $\theta(x)$, this now means that we are matching one firm to $\theta(x)$ 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. The objective is double. We aim to obtain conditions on assortative matching and whether larger firms tend to hire more productive workers. This will give us a simple taxonomy when there is PAM and worker types are increasing or decreasing in firm size, and when there is NAM and again worker types are increasing or decreasing in firm size. We summarize our results in Table 1. Our second objective is to provide a system of differential equations that solves for the equilibrium allocation μ, the firm size distribution θ and the wage schedule w.

3.1 Assortative Matching

Our main result on sorting provides a necessary and sufficient condition that applies to arbitrary type distributions, and places no restrictions on how types influence output. To build up intuition, we 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, $w(x)$ is increasing as better types necessarily earn higher wages. Monotone functions are differentiable 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 must maximize their optimization problem (3). The next Lemma – which also holds for arbitrary distributions and production functions – establishes that we can focus on a simplified problem.

[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) and Eeckhout and Jovanovic (2011).

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. In Appendix. ■

This Lemma states that firms do not choose worker type and intensity unless the combination maximizes the return per unit of resource, which implies that firms with a unique optimizer for (6) hire only one worker type, and therefore endogenously brings our setup in line with one that assumes a uniform workforce from the start. To be clear, this is a result and not a restriction that we impose. The main reason why firms pick only one and not multiple worker types x is because output is additively separable across different x. Therefore, given wages $w(x)$, if a firm generates the highest possible profits from matching with type x, then it would do worse from also hiring worker types $x' \neq x$. Of course, this differentiation results because we lack any direct complementarity between worker types.

Given we can focus on the optimization problem (6) with firms choosing only one type x to match with, our solution now must find *which* x and how many of those, θ. Our objective is to establish positive or negative sorting, where the allocation $\mu(x)$ is single valued, but for now there is no restriction on μ. 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:

$$f_{\theta\theta} \left[f_{xx} - \theta w''(x) \right] - \left[f_{x\theta} - w'(x) \right]^2 \geq 0, \tag{9}$$

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):

$$\mu'(x) \left[f_{xy} - \frac{f_{y\theta} \left(f_{x\theta} - \frac{f_x}{\theta} \right)}{f_{\theta\theta}} \right] \geq 0. \tag{10}$$

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 obtain higher value from being being large but higher worker types are more productive in small firms, then this counteracts the familiar force. This can be seen even more easily when using (2) to express this in terms of the original production function F. The next proposition makes this point, states this as a necessary and sufficient condition, dispenses with assumptions on the type distribution, and does not require 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:*

$$F_{xy} \geq \frac{F_{yl}F_{xr}}{F_{lr}} \tag{11}$$

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 worker type in equilibrium. More importantly, the argument above only shows that the derivative of the matching function $\mu(x)$ has to be positive when equation (11) holds wherever this derivative is defined. In a positive assortative equilibrium this derivative is defined almost everywhere and equation (11) is necessary at these points. To ensure this under all type distributions, we show in the Appendix that equation (11) is necessary everywhere. Equation (11) does not immediately guarantee 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 (11) 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 (11) 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 be integrated from x_l to x_r and from y_l to y_r to yield the global implication that $F(x_h, y_h, 1, 1) + F(x_l, y_l, 1, 1) > F(x_h, y_l, 1, 1) + F(x_l, y_h, 1, 1)$ 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 (11) is more involved and cannot simply be integrated, requiring a substantially more involved argument despite the similarity in spirit.

As in the standard Becker (1973) model with one-to-one matching, equation (11) in Proposition 1 is a functional equation, and the inequality may not be satisfied everywhere on the support of F. That may still lead to PAM (or NAM) for a given distribution, but for an arbitrary distribution there will not always be a monotonic sorting pattern. The conditions we derive below apply when the sorting pattern is monotone, and as such, the characterization that we obtain below in Proposition 2 and Corollary 1 is complete only under monotonicity.

Interpretation: Condition (11) 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 derivative 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 learning by high types is faster.[9] Notably, if better firms particularly value a large size but better workers who particularly excel with plentiful resources (implying few co-workers and, therefore, smaller 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 this term is tightly linked to the concavity in labor holding resources fixed,

[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.

it captures the extent to which additional labor decreases the value of output.[10] The overall condition (11) 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]

The condition for positive assortative matching then compares the within-complementarities with the across-complementarities. In the absence of a quantity dimension (e.g., as in Becker (1973)) 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 negatively affected when additional workers are added. If size is important and better workers lose more in productivity when they receive fewer 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.

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 (47) is still necessary for differential positive assortative matching under increasing returns to scale, only the steps that require homogeneity do not apply.

Finally, our condition (11) depends on four cross-partials, but given four inputs, there are six different possible cross-partials. Why are these four in the condition and why are the other two – namely F_{xl} and F_{yr} – not in the condition? The derivation of the condition exploits the constant returns assumption of F in l, r, which renders it tractable. However, the constant returns assumption also links the missing derivatives to those found in the condition. To see this, observe that constant returns imply we can write $F = lF_l + rF_r$ from Eulers Theorem for homogeneous functions. Taking the derivative with respect to x and rearranging, we obtain $F_{xr} = \frac{F_x - lF_{xl}}{r}$. Similarly, when taking the derivative with respect to y, we obtain: $F_{yl} = \frac{F_y - rF_{yr}}{l}$. Now condition (11) is equivalent to:

$$F_{xy} \geq \frac{(F_y - rF_{yr})(F_x - lF_{xl})}{lrF_{lr}} \tag{12}$$

[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-Mirrlees 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 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.

or any combination of cross-partials, for example:

$$F_{xy} \geq \frac{F_{yl}(F_x - lF_{xl})}{lrF_{lr}}. \tag{13}$$

While this gives another dimension to interpretation of the mechanism, it should be stressed that these expressions are identical because we use the identity that follows from the constant returns assumption.[12]

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 pattern 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 could be 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 differential equations evaluated along the equilibrium allocation at almost all types :*

$$\text{PAM:} \qquad \theta'(x) = \frac{\mathcal{H}(x)F_{yl} - F_{xr}}{F_{lr}}; \quad \mu'(x) = \frac{\mathcal{H}(x)}{\theta(x)}; \quad w'(x) = \frac{F_x}{\theta(x)}, \tag{14}$$

$$\text{NAM:} \qquad \theta'(x) = -\frac{\mathcal{H}(x)F_{yl} + F_{xr}}{F_{lr}}; \quad \mu'(x) = -\frac{\mathcal{H}(x)}{\theta(x)}; \quad w'(x) = \frac{F_x}{\theta(x)}, \tag{15}$$

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

Proof. In Appendix. ■

This first order differential equation system in μ and θ together with appropriate boundary conditions can be used to compute an equilibrium.[13] Proposition 2 is stated from the point of view of workers,

[12] At face value, condition (11) can be interpreted as the geometric average of the "own cross partials" (i.e. within quantity and quality) to be at least as large as the geometric average of the "across cross partials" (i.e. across quantity and quality).

[13] For PAM, one boundary condition is $\mu(\bar{x}) = \bar{y}$. For a guess of $\theta(\bar{x})$, an equilibrium allocation must 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 $\theta(\bar{x})$ must be such that at the lowest active

and $\theta(x)$ is the size of the firm in which worker type x is employed. From the firm's 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 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 interpreted as a normalization.[14] 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 these firms also hire better workers. If these workers have a higher marginal value from obtaining many resources of the firm (F_{xr} large), then the firm will tend to be smaller. 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 patterns 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, 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 mom-and-pop stores, NAM seems to prevail. From condition (11) 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

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 $\mu(\hat{x}) = \underline{y}$.

[14] 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 her rank in the type distribution: $\hat{x}(x) = H^w(x)$. For firms use similarly $\hat{y}(y) = H^f(y)$. Now production is $\hat{F}(\hat{x}, \hat{y}, l, r) = F((H^w)^{-1}(\hat{x}), \left(H^f\right)^{-1}(\hat{y}), l, r)$, where $(H^w)^{-1}$ and $\left(H^f\right)^{-1}$ 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.

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 ($F_{yl} > 0$), they also benefit from spending 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, the distributions H^w, H^f are uniform, 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 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 $\theta(y)$ may change even if we hold the production function and the distribution of firm 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. 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.

We can summarize our results on assortative matching and the size distribution in Table 1.

Table 1: Summary of the Theoretical Results

	PAM	NAM
Worker types increase with firm size: $\theta'(x) > 0$	$F_{xy} > \frac{F_{yl}F_{xr}}{F_{lr}}$ $\mathcal{H}F_{yl} > F_{xr}$	$F_{xy} < \frac{F_{yl}F_{xr}}{F_{lr}}$ $\mathcal{H}F_{yl} < -F_{xr}$
Worker types decrease with firm size: $\theta'(x) < 0$	$F_{xy} > \frac{F_{yl}F_{xr}}{F_{lr}}$ $\mathcal{H}F_{yl} < F_{xr}$	$F_{xy} < \frac{F_{yl}F_{xr}}{F_{lr}}$ $\mathcal{H}F_{yl} > -F_{xr}$

Examples. We now illustrate for a specific technology how the equilibrium allocation changes. The technology is multiplicatively separable in the qualitative inputs and the quantitative inputs, with the quantitative inputs CES and the qualitative inputs Cobb-Douglas.

$$f(x,y,\theta)=\left(\omega_A x^{\frac{1-\sigma_A}{\sigma_A}}+(1-\omega_A)y^{\frac{1-\sigma_A}{\sigma_A}}\right)^{\frac{\sigma_A}{1-\sigma_A}}\theta^{\omega_\theta} \tag{16}$$

In this setting, condition (11) that governs PAM versus NAM is particularly straightforward: there is PAM if $\sigma_A < 1$ and NAM if $\sigma_A > 1$. We illustrate three cases: 1. variation of ω_A under PAM; 2. variation of σ_A under PAM; and 3. variation of ω_θ under NAM.[15]

Figure 1 illustrates that as the share parameter governing the weight on worker quality in production decreases (ω_A goes down), productive firms (high $y = \mu(x)$, and under PAM, also higher x) become larger. The equilibrium allocation $\mu(x)$ therefore shifts upwards as the more productive firms use up more skilled workers. The logic is that as the marginal product of labor decreases, firms substitute quality for quantity and hire more workers. The general equilibrium effect is that the more productive firms become larger and the less productive workers become smaller.

Note that at $\omega_A = 0.5$ the size is constant across firms, $\theta' = 0$. There is no particular relevance to a share parameter of $\omega_A = 0.5$, except that in this particular specification with a lot of symmetry built in already (with identical, uniform distributions for x and y and $\omega_\theta = \frac{1}{2}$), it renders the technology fully symmetric: the production function F remains unchanged if the roles of workers and firms are reversed. Under such symmetry, the tradeoff between better and more workers exactly balances and firm size remains constant for all types. This can straightforwardly (though tediously) be derived. If $\omega_A < 0.5$ then the size effect dominates, and the opposite holds for $\omega_A > 0.5$. By changing any of the other conditions to break the symmetry, it is nevertheless possible to have firm size increasing in firm type even if $\omega_A > 0.5$.

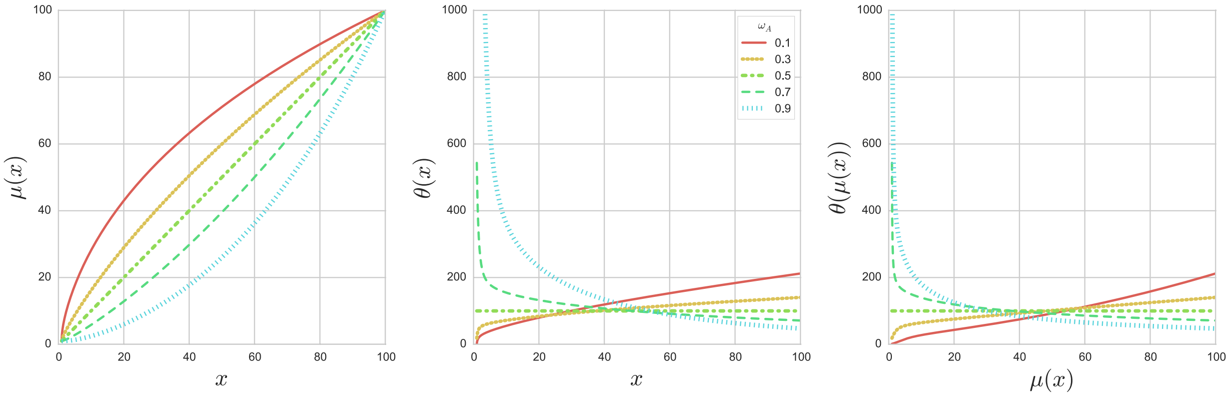

Figure 1: For different values of ω_A, PAM Allocation $\mu(x)$ and intensity/size θ. Simulation with both H_f, H_w uniform on $[0,1]$, $\omega_\theta = 0.5$, and $\sigma_A = 0.9$.

[15] The figures that analyze variation of ω_A under NAM, variation of σ_A under NAM and variation of ω_θ under PAM are similar and available in the online appendix.

Figure 2 depicts the effect on the equilibrium allocation of a decrease in the degree of complementarity between workers skills x and firm productivity y. As the degree of complementarity increases, i.e. σ_A decreases, more productive firms hire more workers. Due to higher complementarity, the demand for skilled labor has gone up and firms hire more. Of course, given limited supply, size cannot increase for all firms. Mediated by higher wages, lower productivity firms in equilibrium hire fewer workers and they are of lower skill x.

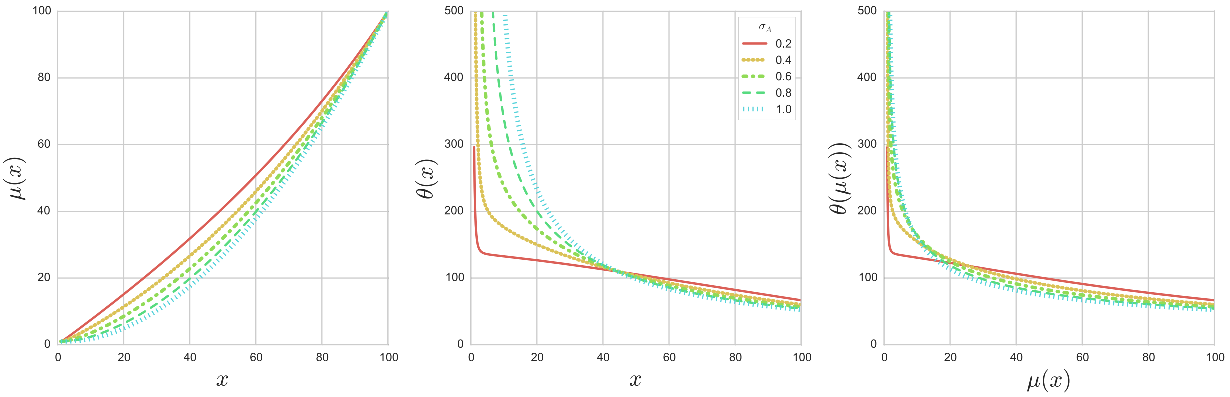

Figure 2: For different values of σ_A, PAM Allocation $\mu(x)$ and intensity/size θ. Simulation with both H_f, H_w uniform on $[0, 1]$, $\omega_A = 0.3$, and $\omega_\theta = 0.6$.

And as ω_θ increases, the penalty for large firm size is reduced and overall labor demand is raised. In the limit, this eliminates decreasing returns and eventually allocates all labor to the best firms. For our parameter constellations with PAM this is monotone: any increase in ω_θ shifts the allocation in this direction in the sense that good firms become larger and workers are matched to better firms (the corresponding figure is omitted to save space).

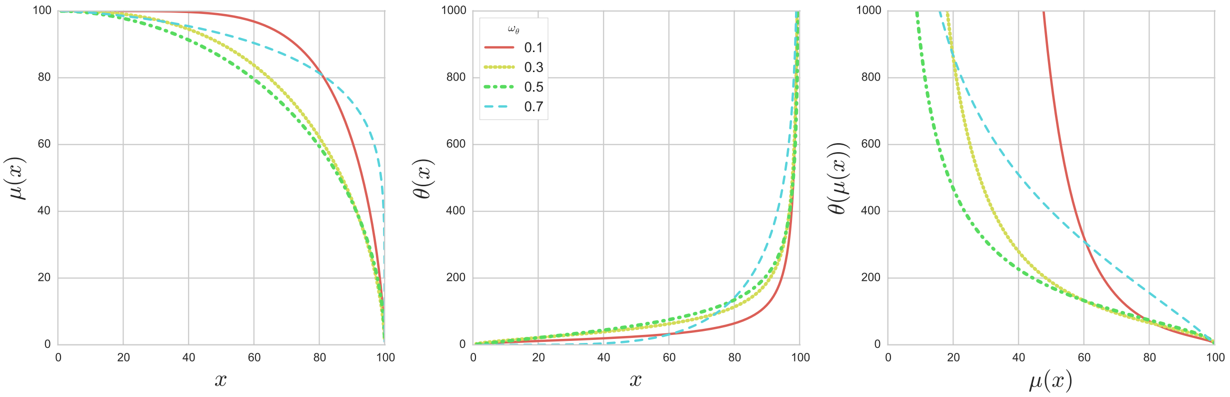

Figure 3: For different values of ω_θ, NAM Allocation $\mu(x)$ and intensity/size θ. Simulation with both H_f, H_w uniform on $[0, 1]$, $\omega_A = 0.5$, and $\sigma_A = 1.1$.

When types are substitutes ($\sigma_A > 1$) the allocation is NAM and the effect of increasing ω_θ becomes more subtle. The size distribution is shaped by market clearing and the fact that under negative sorting the best firms hire the worst workers. This is evident for a given parameter configuration in Figure 3, where the allocation and size distribution do not follow a monotone pattern in ω_θ. As ω_θ increases, more mass is shifted towards high productivity (small) firms, away from low productivity (large) firms. In the right panel we see more at the top for higher μ firms and less a the bottom as ω_θ increases. For mid-level firms the shift is not necessarily monotone: it depends on how many workers from lower type firms are shifted up to the middle firms and how many are moved from the middle further up. Eventually, as ω_θ approaches unity, the more productive firms will attract all the workers, and the slope of the size distribution $\theta'(\mu)$ will change from negative to positive.

The next two sections are devoted to showing that the general setup that we studied, our main theory, is flexible enough to encompass a large range of environments. The subsequent section 4 highlights that our model captures many existing environments as special cases and allows the introduction of two-sided heterogeneity into many competitive economies that have so far been analyzed without them. The most elaborate extension that incorporates search frictions is presented in Section 5.

4 Discussion: Special Cases and Relation to the Literature

In this section we discuss the relation to the existing literature, we present simple extensions, and we highlight the applicability of our framework to economic issues such as mismatch of factor inputs, the measurement of the skill premium, and economic geography. 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 extensively used in the literature. It also highlights that our model can capture new settings that have not been analyzed before. The introduction of search frictions and unemployment is, because of its theoretical and economic significance, the most substantial application, separately presented in Section 5.

Efficiency units of labor have been long-standing instruments 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. In such formulations a firm of type y that hires several worker types x_i at quantities l_i produces output according to some function $\tilde{F}(y, \sum x_i l_i)$. That means 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. Our setup captures this case when $F(x, y, l, r) = r\tilde{F}(y, xl/r)$. Clearly, if only one worker type is hired, the unit amount of internal resources is concentrated on them and we replicate the output $\tilde{F}(y, xl)$. With multiple workers types x_i at quantities l_i, the firm devotes its resources optimally between them, and it can be shown that the resulting output is indeed $\tilde{F}(y, \sum x_i l_i)$ (see Appendix A.11).[16] It can easily be

[16] A generalized version of this setup is a production function $\tilde{f}(\bar{x}, y, L)$ that takes as inputs the average type $\bar{x} =$

verified that in this case our sorting condition is satisfied exactly with equality ($F_{xy}F_{lr} = F_{yl}F_{xr}$), which captures the well-known fact that sorting is 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.

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.[17] They restrict attention to settings where agents must be matched in pairs, which limits insights into the size of the firm and its capital intensity.[18] 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 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 (11) for positive assortative matching is also multiplicatively separable and can be written as $[AA_{xy}/(A_xA_y)][BB_{lr}/(B_lB_r)] \geq 1$.[19]

Constant elasticity of substitution in quantities for the multiplicatively separable case arises if $B(l, r)$'s elasticity is constant and equal to ε; the sorting condition then 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 $B = l^\gamma r^{1-\gamma}$, which 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 $FF_{xy}/(F_xF_y) \geq 1$,

$\sum x_i l_i / \sum l_i$ and the total labor $L = \sum l_i$, whereas standard efficiency units only feature the product $\bar{x}L$. (We thank an anonymous referee for pointing this out.) While our framework does not replicate the generalized case when multiple types are hired, our sorting condition remains valid as long as $\tilde{f}$ is convex in its first argument, as discussed in the Appendix.

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

[18] 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.

[19] In directed search with two-sided heterogeneity, a similar separable formulation arises where B represents the matching technology (see Eeckhout and Kircher (2010)).

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,[20] sorting condition (11) 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 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.[21] 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}. Eeckhout and Jovanovic (2011) extend the Lucas model to allow for worker heterogeneity but worker skills are perfectly substitutable, as in models of efficiency units of labor, so there is no sorting of managers and workers. Rosen (1982)'s supervision model can be interpreted as introducing heterogeneous worker types into this framework through his production function $F = yl\varphi(g(y)r/l, x)$.[22] 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

[20] 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 $\varepsilon \to 0$.

[21] Here we interpret span of control in sense of Lucas (1978), i.e., the number of workers managed. This also encompasses the interpretation by Simon (1957) and Lydall (1959) who refer the to number of managers at a lower management level who are supervised by a manager in the next higher level. In that early literature span of control is related to the shape of a hierarchy and the Pareto coefficient describing the wage distribution in a hierarchy.

[22] 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.

never analyzed his general version, but additionally assumed efficiency units of labor. Within our framework we can apply our sorting condition (11) 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 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 or between countries of different levels of development. We briefly touch on the latter after discussing the role of generic capital in the model.

The Skill Premium with Generic Capital Investment. Additional generic capital inputs can be easily introduced as other factors of production into our framework. 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 $\hat{F}(x, y, l, r, k)$. In this formulation k is allowed to be a vector in case there is more than one capital good. Optimal use of resources requires $F(x, y, l, r) = \max_k \left[\hat{F}(x, y, l, r, k) - i'k\right]$, where i is the vector of unit prices which the individual firm takes as given and is either fixed to world market levels in a small open economy or otherwise determined by an additional equilibrium condition that equates the capital demand across all firms to the aggregate capital stock. Note that F is constant returns in the last two arguments if $\hat{F}$ is constant returns in the last three arguments, and all the machinery in this paper can be applied. Recall the alternative interpretation of our model (see footnote 5) in which any entrepreneur can hire input 1 at type x at quantity l and input 2 of type r at quantity y to maximize $\hat{F}(x, y, l, r, k) - lw_x(x) - rw_y(y) - i'k$ where w_x and w_y are the hedonic price schedules for each of the inputs (equal to the wages and profits in our derivation). Labelling input 1 as "high-skilled" workers and input 2 as "low-skilled" workers, this exactly captures the within-period interaction in the seminal work on the skill premium following Krusell, Ohanian, Ríos-Rull, and Violante (2000), only there it is assumed that individuals within each group are homogeneous. Our setup allows for within-group heterogeneity, when the skill premium $w_y(y)/w_x(x)$ becomes type-specific. Our sorting condition can be expressed in terms of F as defined above. If there is only one generic capital good and k is simply a number, then rewriting the cross-margin-complementarity condition (11) in terms of the new primitive yields the following condition for positive assortative matching: $\hat{F}_{xy}\hat{F}_{lr}\hat{F}_{kk} - \hat{F}_{xy}\hat{F}_{lk}\hat{F}_{rk} - \hat{F}_{xk}\hat{F}_{yk}\hat{F}_{lr} \geq \hat{F}_{xr}\hat{F}_{yl}\hat{F}_{kk} - \hat{F}_{xr}\hat{F}_{yk}\hat{F}_{lk} - \hat{F}_{xk}\hat{F}_{yl}\hat{F}_{rk}$.

This is obtained from applying equation (11) to F together with the fact that F solves the maximization problem $\max_k \left[\hat{F}(x,y,l,r,k) - i'k\right]$. 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.

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 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 i.[23] 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 in line with discussion in the preceding paragraph allows us to use the results from the main body that did not explicitly incorporate generic capital; 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.[24] Such dispersion would be an efficient outcome given the supply of good land, rather than a distortion.

Figure 4 uses the parameters for the developing countries and shows how firms of different types 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

[23] 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.

[24] 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)).

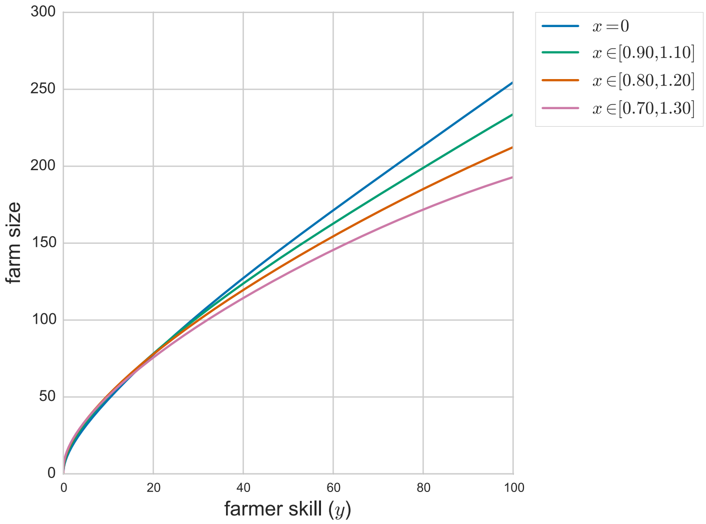

Figure 4: Firm size distribution for different dispersion in x (x is log-normally distributed $\mathcal{LN}(0, 0.2)$, i.e. $\log x$ is normally distributed with mean 0 and variance 0.2, truncated at the bounds indicated in the legend, with the measure of the truncated distribution normalized to one).

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.[25] 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 (11) is

[25] 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).

satisfied if $v(\cdot)$ is concave, so that individuals with high incomes locate centrally and those with low incomes in the periphery.[26]

Relation to the General Equilibrium Literature. 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 sorting or wage patterns. The classic example in the combinatorial matching literature by Kelso and Crawford (1982) proposes 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 grows 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.

Multiple Skill Inputs. The prediction of the model that each firm hires exactly one skill type is obviously counterfactual. Firms do hire workers of different ability. And while there is mounting evidence that most of the increase inequality in worker earnings is due to the increase between firms, not within firms,[27] the within firm inequality continues to be substantial and important. The reason why in our model only one skill type is hired is because we assumed, despite the multiple levels of complementarities between different inputs, there are no complementarities between different x's. Not only does this ensure that the gross substitutes condition for existence discussed in the previous paragraph is satisfied, this assumption has allowed us to characterize the sorting and size patterns despite that

[26] 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.

[27] 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, 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.

fact that that the general many-to-many matching model is notoriously hard to analyze (Kelso and Crawford (1982)).

To address the issue of the multiple skill inputs and in order to draw closer to the empirically observed heterogeneity, we have generalized the model to a setting in which there are n distinct categories of skilled workers, each with a distribution of skills $x_i, i = 1, ..., n$. We think of these categories as education. Output produced then needs inputs from the different education categories, and the firm chooses the skill level x_i for each of these categories i.

Consider therefore a technology with firm type y as before and n skilled inputs, x_i for $i = 1, ..., n$. We assume that these skill types are from disjoint sets and that there is no substitution between these skills. Let the measure of these skills be $H^{w,i}$. Then the technology can be written as

$$f(x_1, \theta_1, , ..., x_n, \theta_n, y) = f(\mathbf{x}, \boldsymbol{\theta}, y), \tag{17}$$

where $\mathbf{x} = x_1, ..., x_n, \theta_i = \frac{l_i}{r_i}$ and $\boldsymbol{\theta} = \theta_1, ..., \theta_n$. Now an allocation are matchings: $y = \mu_i(x_i)$ for all i. In the case of PAM the market clearing (or feasibility) condition is satisfied for each category i:

$$\int_{\mu_i(x_i)}^{\overline{y}} \theta_i(s) h^f(s) ds = \int_x^{\overline{x}} h^{w,i}(s) ds. \tag{18}$$

Unfortunately, this is a complex problem for which we are not able to derive general conditions for assortative equilibrium as in the $n = 1$ case.[28] The problem is challenging even with two skill inputs because of the complementarities between different skills and the associated sizes. Moreover, we know from Kelso and Crawford (1982) that existence is not guaranteed, except under a condition of gross substitutes. Finally, we can also not simply rewrite the conditions for sorting in terms of increasing differences (as in the one-to-one matching model) because now it is not simply a matter of rearranging types, there are also quantities. However, here we show that under some assumptions, we can make progress in solving this allocation problem.

First, we assume that each category i produces output exactly as in our baseline model, that is, we write output as an aggregation g of the output of each category i:

$$f(\mathbf{x}, \boldsymbol{\theta}, y) = g(f^1(x_1, y, \theta_1), ..., f^n(x_n, y, \theta_n)). \tag{19}$$

We require output to be concave in θ_i, i.e., $g_{ii} f^i_\theta + g_i f^i_{\theta\theta} < 0$ where g_i and g_{ii} denotes the partial first and second derivative with respect to the $i'th$ argument. Clearly this condition is met if g is concave in each argument. We also require $g_i > 0$. Again define $F^i(x, y, l, r) = r f^i(x, y, l/r)$ as under $n = 1$. This functional form in equation (19) will allow us to express the conditions for sorting in terms of the

[28] Even with $n = 2$, the Hessian from which we derive the sorting condition is now a 4×4 matrix that needs to be negative definite. After substituting for the differentiated first order conditions, the determinant of the Hessian becomes an expression with many different interactions that are complicated to handle. The derivation is reported in the online appendix.

each individual function f^i and associated F^i as in the case of $n = 1$, in combination with properties on g. Clearly, additive separability where $g(f^1, ..., f^n) = \sum_i f^i$ is a special case in which we immediately obtain the results from Propositions 1 and 2 extend for each skill input: there is PAM between y and x_i provided $F^i_{xy} \geq \frac{F^i_{yl} F^i_{xr}}{F^i_{lr}}$ (NAM under the opposite sign) and the differential equations (14) and (15) fully characterize the equilibrium allocation $\mu_i(x_i)$ for each i. Separability implies that the allocation problem of each skill input x_i is solved in isolation and can easily be verified that the conditions for negative definiteness of Hessian of the full problem coincide with the negative definiteness of each allocation problem of F^i in isolation. Only now different worker types associated with the same firm y are bundled together for observational purposes. The subsequent analysis aims to extend this to more meaningful aggregators.

Second, assume that each category i makes its own hiring decisions, aiming to maximize the overall profitability of the firm. We will refer to this as “independent hiring”. That is, we assume each unit $i = 1, ..., n$ solves the problem

$$\max_{x_i, \theta_i} g(\mathbf{f}(\mathbf{x}, \boldsymbol{\theta}, y)) - \sum_i \theta_i w_i(x_i). \tag{20}$$

This is a simpler problem than the full problem of a single firm that maximizes over all $x_1, ..., x_n$ and $\theta_1, ..., \theta_n$ simultaneously:

$$\max_{\mathbf{x}, \boldsymbol{\theta}} g(\mathbf{f}(\mathbf{x}, \boldsymbol{\theta}, y)) - \sum_i \theta_i w_i(x_i). \tag{21}$$

Problem (20) might be reasonable in some larger organizations where coordination between units is difficult and each unit optimizes its allocation independently. More importantly, this also has the advantage to yield a simpler pass at the difficult overall problem (21). A solution to the general problem (21) is also a solution to the problem (20). One way to see this is to notice that the $2n$ first-order conditions of both sets of problems are identical. The main differences are the second-order conditions, which are more demanding in the general problem than in the simplified one. Nevertheless, if we can prove conditions that apply to *all* equilibria of the simplified setting (for example, conditions for positive assortative matching that apply PAM for all such equilibria), we have also shown that these apply in the general problem.

With these two restrictions we now derive a set of sufficient conditions on the technology such that there is assortative matching. The reason why we need additional restrictions is that along the equilibrium allocation $f^i(\nu_i(y), y, \theta_i(\nu_i(y)))$ may not be increasing in y, where $\theta_i(x)$ is the size of the workforce of type x and $\nu_i(y)$ is the worker type i hired by firm y. Because the function g is monotonic in each argument, then a decreasing f^i may make the solution not positively assortative in another dimension j, precisely because it is increasing in its arguments and supermodular. To that end, we calculate the total derivative of f^i with respect to y along its equilibrium allocation, which we denote by the total derivative dF^j/dy. We obtain the following result:

Proposition 3 *Consider the model with multiple skill inputs* $f(\mathbf{x}, \boldsymbol{\theta}, y) = g(f^1(x_i, y, \theta_1), ..., f^n(x_n, y, \theta_n))$

and independent hiring; a necessary condition for PAM requires along the equilibrium path for all $i \in 1, ..., n$:

$$(g_i)^2 \left[F^i_{x_i y} F^i_{l_i r_i} - F^i_{y l_i} F^i_{x_i r_i}\right] + g_i g_{ii} \left[-\frac{F^i}{r_i^2} + \frac{F^i_{r_i}}{r_i}\right] \left[F^i_{x_i y} F^i_{l_i} - F^i_{y l_i} F^i_{x_i}\right]$$

$$+ g_i \left[g_{ii} \frac{F^i_y}{r_i} + \sum_{j \neq i} g_{ij} \frac{dF^j}{dy}\right] \left[F^i_{l_i r_i} F^i_{x_i} - F^i_{x_i r_i} F^i_{l_i}\right] \geq 0. \tag{22}$$

If this holds for all $(\mathbf{x}, \mathbf{l}, \mathbf{r}, y)$ in a candidate equilibrium satisfying the first order conditions to (20) with given $\frac{dF^j}{dy}$, then this allocation is an equilibrium with PAM. If this holds for all $(\mathbf{x}, \mathbf{l}, \mathbf{r}, y)$ and all possible $\frac{dF^j}{dy}$, then any equilibrium has to be PAM.

Proof. In Appendix. ■

The next proposition establishes that if the production function is convex and supermodular, our condition for sorting together with the condition that more productive firms are larger under any type densities, coupled with a new condition on the interplay with size, ensures even in the multi-unit setting an equilibrium that is PAM and more productive firms are larger:

Proposition 4 *Consider the model with multiple skill inputs $f(\mathbf{x}, \boldsymbol{\theta}, y) = g(f^1(x_i, y, \theta_1), ..., f^n(x_n, y, \theta_n))$ and independent hiring, and assume that higher type firms y produce more output. A sufficient condition for an equilibrium under PAM in which more productive firms are larger and produce more (i.e., $dF^j/dy \geq 0$) is: $g_{ij} \geq 0$ for all $(i, j) \in \{1, 2, .., n\}^2$ with $i \neq j$, and for all i and all $(\mathbf{x}, \mathbf{l}, \mathbf{r}, y)$ it holds that*

$$F^i_{x_i y} F^i_{l_i r_i} - F^i_{y l_i} F^i_{x_i r_i} \geq 0, \tag{23}$$

$$F^i_{y l_i} \geq 0, F^i_{x_i r_i} \leq 0 \tag{24}$$

$$g_{ii} \left(\left[F^i_{l_i r_i} F^i_{x_i} - F^i_{x_i r_i} F^i_{l}\right] F^i_y - \left[F^i_{x_i y} F^i_{l} - F^i_{y l_i} F^i_{x_i}\right] F^i_{l} \frac{l_i}{r_i} \right) \geq 0. \tag{25}$$

Proof. In Appendix. ■

If g is concave in each argument (i.e., $g_{ii} < 0$) then the last condition highlights naturally that a large $F^i_{x_i y}$ is a force towards positive sorting while $F_{y l_i}$ large is a force against it. This is in spirit similar but in functional form different from our main condition (23) that carries over from the model with $n = 1$.

Finally, we provide a short illustration that we obtain very simple conditions for sorting for a specific technology. Consider the setting where output is given by $f = \Pi_i f^i(x_i, y, l_i)$. In this case the aggregator g is multiplicative in f^i, and we can prove that the condition for PAM in the benchmark model with $n = 1$ extends to each of the technologies F^i for each category i.

Corollary 2 *Let $F^i = A^i(x_i, y)B^i(l_i, r_i)$ where $B^i = l_i^{\phi_i} r_i^{1-\phi_i}$ $(\phi_i \in (0,1))$, and let $g_{ii} \leq 0$. A sufficient condition for positive assortative matching in any equilibrium with independent hiring is that for all i:*

$$F^i_{x_i y} F^i_{l_i r_i} - F^i_{y l_i} F^i_{x_i r_i} > 0. \tag{26}$$

Therefore, any equilibrium where a single firm jointly makes all hiring decisions also displays positive assortative matching.

Proof. In Appendix. ■

This condition to be satisfied for each skill group i is the same as the condition with a single skill input (11). Observe that this condition does not require that output of any given skill category j is increasing along the equilibrium allocation. Any effect a decreasing output of input j could have on the optimal choice of skill i is neutralized by the linearity of g and the Cobb-Douglas functional form of $B(l, r)$. Note also that the condition in the Corollary is sufficient, not necessary. For example, g_{ii} may be positive as long as ϕ is not too large.

This can be applied to the technology that we used to derive the example in section 3 in the benchmark case, but now let $n = 2$. Let g be multiplicative so that $g(f^1, f^2) = f^1 \cdot f^2$ with associated aggregates $F^i = A(x_i, y)B(l_i, r_i)$ where A is CES and B is Cobb-Douglas. Then firm y's output $g(f^1(x_1, y, \theta_1), f^2(x_2, y, \theta_2))$ can be written as:

$$g = \left(\omega_1 x_1^{\frac{\sigma_1 - 1}{\sigma_1}} + (1-\omega_1) y_1^{\frac{\sigma_1 - 1}{\sigma_1}}\right)^{\frac{\sigma_1}{\sigma_1 - 1}} \theta_1^{\phi_1} \cdot \left(\omega_2 x_2^{\frac{\sigma_2 - 1}{\sigma_2}} + (1-\omega_2) y_2^{\frac{\sigma_2 - 1}{\sigma_2}}\right)^{\frac{\sigma_2}{\sigma_2 - 1}} \theta_2^{\phi_2}. \tag{27}$$

For this technology, there will be PAM for all $\phi_i \in (0, 1)$ provided that $\sigma_i < 1$ for all i, where the latter condition follows from equation (26).

Extensions that allow for monopolistic competition at the output level within a Dixit-Stiglitz setup are presented in the Appendix, where an extended condition for sorting is derived. There we link our model to the optimal transportation literatureand also introduce endogenous type distributions.

5 Sorting with Large Firms and Search frictions

Questions about sorting between workers and firms have attracted interest not only for competitive markets, but increasingly also for markets with matching frictions. Incorporating matching frictions allows us to study not only matching and wage patterns but also the determinants of unemployment across heterogeneously skilled agents (see, e.g., Shimer and Smith (2000) and Atakan (2006) in the presence of random search and Shi (2001), Shimer (2005) and Eeckhout and Kircher (2010) under directed search). At the same time, multi-worker matching has recently attracted attention in that literature to analyze how queue lengths vary across firms of different sizes (e.g., Smith (1999) for random and Menzio and Moen (2010), Garibaldi and Moen (2010), Hawkins (2011), Kaas and Kircher (2015), and Schaal (2015) for directed search), yet, little is known about how unemployment varies in the presence of sorting *and* variation in firm size jointly.[29] The contribution of the extension of the model developed here is to enable the joint analysis of firm size, search frictions and sorting.

We now show that the techniques developed in this paper are suitable to analyze search frictions with large firms and sorting; notably 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 technology 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. To illustrate that our model is amenable to this, we embed a costly recruiting and search process in the previous setup in order to capture the hiring behavior of large firms. We first formally lay out the search model, then relate it to the competitive economy we studied thus far, and finally we use this connection to prove theorems about the search economy.

The Directed Search Economy with Two-Sided Heterogeneity and Large Firms. This setup builds on the directed search literature cited above, combining insights from one-to-one sorting and large firms into one framework. As in the previous literature, we assume for simplicity's sake 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. Following Shi (2001), firms post for each vacancy the type of worker they want to hire and the wage. That is, if firm y dedicates a vacancy to worker type x, it specifies wage $\omega^y(x)$. But now, a firm of type y decides also how many vacancies $v^y(x)$ to post per unit of resources devoted to workers of type x that it wants to hire. We assume that all $v^y(x)$ vacancies have the same wage $\omega^y(x)$, but it is easy to show that this assumption is without loss of generality. Posting $v^y(x)$ vacancies has a linear cost $cv^y(x)$. Observing all vacancy postings, workers decide where to search for a job amongst those dedicated to their type. 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 job-finding rate is $m(q^y(x))/q^y(x)$, it is assumed to be strictly decreasing in the number of workers $q^y(x)$ queueing per vacancy.

Workers expect to obtain an expected equilibrium payoff of $w(x)$. Firms can attract workers to their vacancies as long as these workers get in expectation their equilibrium payoff, meaning that $q^y(x)$ adjusts depending on $\omega^y(x)$ to satisfy: $\omega^y(x)m(q^y(x))/q^y(x) = w(x)$. Often $w(x)$ is called the "market

[29] Lentz (2010) and Bagger and Lentz (2016) study sorting when firm size is limited only by search frictions, albeit with linear production and with no interaction of workers within the production process of the firm.

utility" as it represents the return that workers could get by queueing elsewhere for a job. 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

$$\max_{\mathcal{R}^y,\theta^y,\omega^y,v^y} \int \left[f(x,y,\theta^y(x)) - \theta^y(x)\omega^y(x) - v^y(x)c\right] d\mathcal{R}^y(x) \tag{28}$$
$$\text{s.t. } \theta^y(x) = v^y(x)m(q^y(x)); \quad \text{and} \quad \omega^y(x)m(q^y(x))/q^y(x) = w(x)$$

and $\mathcal{R}^y(x)$ integrates to unity. The first line simply takes into account that the firm must pay the vacancy-creation cost. The second line states that the number of hires equals the number of vacancies times the job-filling rate, and links the job-filling rate to the wage offer as explained earlier.

The amount of workers with types below x queueing at firms of type y is now $\tilde{\mathcal{L}}^y(x) = \int_{\underline{x}}^{x} \theta^y(\tilde{x})d\mathcal{R}^y(\tilde{x})$. Clearly, feasibility requires that no more workers can be queueing than there are in the economy, i.e., for all x' and x it has to hold that

$$\int_y \left[\tilde{\mathcal{L}}^y(x) - \tilde{\mathcal{L}}^y(x')\right] dH^f \leq H^w(x) - H^w(x'). \tag{29}$$

Finally, workers' wage expectations $w(x)$ must be consistent with the expected payments that firms are willing to make in equilibrium. In particular, if wage expectations $w(x'')$ for almost all workers x'' in some interval $(x', x]$ are strictly positive, then there must be enough vacancies that actually offer such wages: all these workers must actually be queueing for jobs that offer such wages (as workers who do not queue cannot obtain a strictly positive payoff). Therefore, equation (29) cannot be slack, which leads to the following equilibrium definition for a directed search economy with large firms:

Definition 2 *An equilibrium in a directed search economy is a tuple of functions* $(w, \theta^y, \omega^y, v^y, \mathcal{R}^y, \mathcal{L}^y)$ *consisting of a non-negative market utility* $w(x)$ *as well as intensity functions* $\theta^y(x)$*, vacancy function* $v^y(x)$*, actual pay* $\omega^y(x)$*, and resource allocations* $\mathcal{R}^y(x)$ *with associated feasible* $\tilde{\mathcal{L}}^y(x)$ *such that*

1. *Optimality: For any* y *the combination* $(\theta^y, \omega^y, v^y, \mathcal{R}^y)$ *solves (28).*

2. *Correct Wage Expectations: (29) holds with equality if the market utility is strictly positive a.e. on* $(x', x]$.

Relation to a Competitive Economy. We now show that a relabeling of variables allows us to represent this economy as if it were a competitive market economy. This allows us to use the insights from the previous sections to provide insights even in the case of search frictions. The main idea is the following: in the current setup $q^y(x)$ and $v^y(x)$ determine the number $s^y(x) = q^y(x)v^y(x)$ of workers who search for jobs (per unit of resource). Alternatively, the firm could choose $s^y(x)$ and $v^y(x)$, and then $q^y(x) = s^y(x)/v^y(x)$ follows. Moreover, instead of maximizing over $s^y(x)$ and $v^y(x)$ simultaneously, the firm can first determine $s^y(x)$ and then choose v^y subsequently to maximize the following problem:

$$g(x, y, s) = \max_v \left[f(x, y, vm(s/v)) - vc\right] \tag{30}$$

where we replaced θ in (28) by the constraint $\theta = vm(q) = vm(s/v)$. Problem 30 is strictly convex, and the unique optimum $v^*(x, y, s)$ is given by the first order condition $m(q)[1-qm'(q)/m(q)]f_l(x, y, sm(q)/q) = c$, where $q = s/v^*$. This corresponds to the well-known Hosios condition for efficient search. Simple replacement then transforms (28) and (29) and associated Definition 2 into the following equivalent ones:

$$\max_{\mathcal{R}^y, s^y} \int \left[g(x, y, s^y(x)) - s^y(x)w(x)\right] d\mathcal{R}^y(x) \tag{31}$$

where $\mathcal{R}^y$ integrates to unit; $\mathcal{L}^y(x) = \int_{\underline{x}}^{x} s^y(\tilde{x}) d\mathcal{R}^y(\tilde{x})$ and

$$\int_y \left[\mathcal{L}^y(x) - \mathcal{L}^y(x')\right] dH^f \leq H^w(x) - H^w(x'); \tag{32}$$

and $(w, s^y, \mathcal{R}^y, \mathcal{L}^y)$ constitutes an equilibrium if $(s^y\ \mathcal{R}^y)$ solves (31) and if $\mathcal{L}^y$ satisfies (32), with equality if $w(x'') > 0$ for allmost all types in $(x, x']$. Observe that this is exactly our definition of a competitive equilibrium, only now with fictitious production function g that embeds both the production as well as the matching process. Note that this works precisely because the firm has to pay any "searcher" that queues for a job the expected wage, whether or not it ends up getting the job. Note $G(x, y, s, r) = rg(x, y, s/r)$ provides the aggregate production function, which has constant returns to scale in s and r by construction. We sum up the connection between the directed search economy and the competitive economy in the following proposition:

Proposition 5 *The expected wages (market utility) and matching patterns in a competitive search economy with type distributions H_f and H_w and production function $F(x, y, l, r)$ coincide with those of a competitive economy with same type distributions but fictional production function $G(x, y, s, r) = \max_v \left[F(x, y, rvm(\frac{s}{vr}), r) - rvc\right]$. Moreover, for types that match in equilibrium $v^y(x) = v^*(x, y, s)$, $q^y(x) = s^y(x)/v^*(x, y, s)$, and $\theta^y(x)$ and $\omega^y(x)$ are then determined as in the constraint in (28).*

Results for Competitive Search with Large Firms. The beauty of the previous result is that G is fully determined by the primitives, and we can apply the tools we laid out in Section 2 (where now G replaces F). The firm can be viewed as if it hires "searchers" who must be paid their expected wage. Applying the machinery from the previous section allows us to assess whether sorting is assortative, and surprisingly our cross-partial condition on G reduces to the standard cross-partial condition on F, so no further requirements need to be imposed for sorting:

Proposition 6 *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 (11) holds for all $(x, y, l, r) \in \mathbb{R}^4_{++}$ that satisfy $f_l(x, y, l/r) > c$. If equation (11) holds with strict inequality on this domain, 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. ■

Intuitively, the reason for this result is that vacancies can be adjusted at linear costs to absorb the distortions in the matching function, and ultimately the driving force for sorting remains on the production side.[30] The main reason why we only need our sorting condition for combinations of (x, y, l, r) such that $f_l(x, y, l/r) > c$ is due to the fact that a number of searchers s leads to optimal vacancies v^*, but only to the level where the cost of posting can be covered.

The next question concerns the matching probabilities and the size of the firm. Let $l(x)$ denote again the equilibrium size of the firm in which worker type x is employed. While the sign of sorting is not dependent on the matching function, the condition whether higher-type firms are larger, does. This is suggested already by the definition of firm size:

$$l(x) = s(x)m(q(x))/q(x), \tag{33}$$

where $s(x)$ and $q(x)$ are again the equilibrium values for the number of searchers and the queue lengths, and $m(q)/q$ is the matching probability of the worker. The matching probability depends on the optimal number of vacancies. Rather than studying firm size directly, it is convenient to start with the workers matching probabilities for which we obtain an unambiguous result:

Proposition 7 *Assume higher worker types create more output ($F_x > 0$). In the competitive search equilibrium with large firms, higher skilled workers have lower market tightness ($q(x)$ decreasing) and associated higher matching probability. For types with strictly positive expected wages $w(x) > 0$ these relationships are strict.*

Proof. In appendix. ■

Since better workers obtain higher expected payoff $w(x)$ as determined in Problem 1 (otherwise a firm could hire better workers at equal cost), they face proportionally lower competition for each job

[30] We expect the matching function to matter more if the marginal cost of posting vacancies increase with the number of vacancies that are posted. With convex creation costs we expect results to feature the matching function more prominently, and in extreme cases where some vacancies are costless but after a threshold costs increase to infinity we expect sorting conditions that resemble those of settings where each firm can post one vacancy.

and correspondingly higher job finding probabilities. This outcome 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. Note that this result is independent of the sorting patterns.

The sorting patterns matter when one wants to study the link between firm productivity, the vacancy-filling probability and the vacancy rate v/l, which are often used in empirical work. Since we know how these are linked to worker characteristics, we can link them to firm productivities under sorting. But since size is not necessarily increasing with productivity, there is no direct link to firm size:

Proposition 8 *In a competitive search equilibrium the vacancy rate (v/l) is increasing and the vacancy filling probability is decreasing in firm productivity (y) under PAM, and the reverse under NAM. They are ambiguous in firm size.*

Proof. In Appendix. ■

Finally, conditions for overall firm size are dependent both on the production and the matching function. Consider again the setting with $F_x > 0$. Sufficient conditions for increasing size under PAM is that the number of searchers s is increasing in productivity, i.e., that $\mathcal{H}(x)G_{yl} \geq G_{xr}$. This resembles part 1. of Corollary 1 and ensures a larger number of searchers, $s(x)$ is increasing, which then is reinforced by the higher matching rate that these have (Proposition 7), and leads more productive firms to be larger overall. Under NAM these two counteract, and size conditions depend in an intricate way on the matching and production function.

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, labor and industrial organization literatures. While sufficiently rich to incorporate the most relevant features of heterogeneity, in particular worker skill, firm productivity, firm size and wage inequality our model is remarkably simple to analyze and can readily be used to plug into a larger model of the economy. As an example, we establish that equilibrium unemployment can be incorporated, thus enabling the joint analysis of frictional unemployment in the presence of both sorting of heterogeneous workers and firms, and of assortative matching in large firms.

Appendix A Omitted Proofs and Derivations

A.1 Proof of Lemma 1

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 (max-problem-simple). Let (x^*, θ^*) be an optimizer of (6). 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. ■

A.2 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

$$\begin{aligned}
& u(\alpha\mathcal{L}^y + (1-\alpha)\mathcal{L}'^y, \alpha n^y + (1-\alpha)n'^y|y) && (34)\\
= \; & \alpha n^y + (1-\alpha)n'^y + \max_{\substack{\mathcal{R}^y \text{ s.t.} \\ \theta^y(x)=d(\alpha\mathcal{L}^y+(1-\alpha)\mathcal{L}'^y)/d\mathcal{R}^y \\ \int d\mathcal{R}^y<1}} \int f(x,y,\theta^y(x))d\mathcal{R}^y(x) && (35)\\
\geq \; & \alpha n^y + (1-\alpha)n'^y + \max_{\substack{\mathcal{R}^y \text{ s.t.} \\ \theta^y(x)=\alpha d\mathcal{L}^y/d\mathcal{R}^y \\ \int d\mathcal{R}^y<\alpha}} \int f(x,y,\theta^y(x))d\mathcal{R}^y(x) + \max_{\substack{\mathcal{R}^y \text{ s.t.} \\ \theta^y(x)=(1-\alpha)d\mathcal{L}'^y/d\mathcal{R}^y \\ \int d\mathcal{R}^y<1-\alpha}} \int f(x,y,\theta^y(x))d\mathcal{R}^y(x) && (36)\\
= \; & \alpha u(\mathcal{L}^y, n^y|y) + (1-\alpha)u(\mathcal{L}^y, n^y|y) && (37)\\
\geq \; & u(\mathcal{L}''^y, n''^y|y). && (38)
\end{aligned}$$

■

A.3 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 $x \in$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. ■

A.4 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:

$$\mathbf{H} = \begin{pmatrix} f_{\theta\theta} & f_{x\theta} - w'(x) \\ f_{x\theta} - w'(x) & f_{xx} - \theta w''(x) \end{pmatrix}. \tag{39}$$

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

$$f_{\theta\theta}[f_{xx} - \theta w''(x)] - (f_{x\theta} - w'(x))^2 \geq 0. \tag{40}$$

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

$$f_{x\theta} - w'(x) = -\mu'(x) f_{y\theta} - \theta'(x) f_{\theta\theta} \tag{41}$$

$$f_{xx} - \theta(x) w''(x) = -\mu'(x) f_{xy} - \theta'(x) \left[f_{x\theta} - w'(x)\right]. \tag{42}$$

In the following three lines we successively substitute (42), (41) and then (8) into condition (40):

$$-\mu'(x) f_{\theta\theta} f_{xy} - \left[\theta'(x) f_{\theta\theta} + f_{x\theta} - w'(x)\right] \left[f_{x\theta} - w'(x)\right] > 0 \tag{43}$$

$$-\mu'(x) f_{\theta\theta} f_{xy} + \mu'(x) f_{y\theta} \left[f_{x\theta} - w'(x)\right] > 0 \tag{44}$$

$$-\mu'(x)[f_{\theta\theta} f_{xy} - f_{y\theta} f_{x\theta} + f_{y\theta} f_x/\theta] > 0 \tag{45}$$

Since $f_{\theta\theta} < 0$ we can divide 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. Focusing on positive assortative matching, and using the relationship in (8), we obtain the condition:

$$f_{\theta\theta} f_{xy} - f_{y\theta} f_{x\theta} + f_{y\theta} f_x/\theta \leq 0. \tag{46}$$

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.[31] A standard implication of constant returns is then $F_x(x, y, \theta, 1) = \theta F_{xl} + F_{xr}$. We can now rewrite (46) in terms of $F(x, y, \theta, 1)$ and rearrange to obtain the following cross-margin-complementarity condition:

$$\begin{aligned} & F_{ll} F_{xy} - F_{yl} \left[F_{xl} - F_x/\theta \right] \leq 0 & (47) \\ \Leftrightarrow \quad & F_{ll} F_{xy} + F_{yl} F_{xr}/\theta \leq 0 & \\ \Leftrightarrow \quad & F_{xy} F_{lr} \geq F_{yl} F_{xr}. & (48) \end{aligned}$$

Since $F_{lr} > 0$, we can divide through to obtain inequality (11) 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.

A.5 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 the strict version of inequality (11) 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

$$S(\mathcal{L}, \mathcal{R}) = \int_{y\in\mathcal{Y}} \int_{x\in\mathcal{X}} F(x, y, \theta^y(x), 1) d\mathcal{R}^y dH^f, \tag{49}$$

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 (11) 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 $F_{xy} > \frac{F_{xr}F_{yl}}{F_{lr}}$ 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 two combinations (x_1, y_1, θ_1) and (x_2, y_2, θ_2) exist 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

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

worker types in an $\varepsilon-$neighborhood around x_i that obtain 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.

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 is 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 $\hat{r}$ of resources of type $\hat{y}$ firms with workers of type x. The optimal output is generated by withdrawing some optimal measure $\hat{\theta}\hat{r}$ of the workers that were supposed to be working with resources of type y and reassigning them to work with resources of type $\hat{y}$. The joint output at (x, y) and $(x, \hat{y})$ in (49) is given by

$$r f(x, y, \theta - \hat{\theta}\hat{r}/r) \; + \; \hat{r} f(x, \hat{y}, \hat{\theta}). \tag{50}$$

Optimality of $\hat{\theta}$ requires, according to the first order condition, that $f_\theta(x, y, \theta - \hat{\theta}\hat{r}/r) = f_\theta(x, \hat{y}, \hat{\theta})$, which shows that the optimal $\hat{\theta}$ is itself a function of $\hat{r}$. Denote $\beta(\hat{y}; x, y, \theta)$ the marginal increase of (50) from increasing $\hat{r}$, evaluated at $\hat{r} = 0$, is given by

$$\beta(\hat{y}; x, y, \theta) = f(x, \hat{y}, \hat{\theta}) - \hat{\theta} f_\theta(x, y, \theta) \tag{51}$$

$$\text{where } \hat{\theta} \text{ is determined by } \; f_\theta(x, \hat{y}, \hat{\theta}) = f_\theta(x, y, \theta). \tag{52}$$

The **constraint** (52) reiterates the optimality of $\hat{\theta}$. Sometimes we will write $\hat{\theta}(\hat{y}; x, y, \theta)$ to highlight that $\hat{\theta}$ is a function of $\hat{y}, y, x$ and θ.

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 $\hat{y} = y_1$ to workers with combination (x_1, y_1, θ_1) than with workers that are otherwise matched at (x_2, y_2, θ_2):

$$\beta(y_1; x_2, y_2, \theta_2) \leq \beta(y_1; x_1, y_1, \theta_1), \tag{53}$$

where $\beta(\cdot; \cdot, \cdot, \cdot)$ was defined in (51). Similarly, the marginal gains from pairing the last unit of resources of type $\hat{y} = y_2$ to workers otherwise matched at (x_2, y_2, θ_2) than to workers matched at (x_1, y_1, θ_1):

$$\beta(y_2; x_2, y_2, \theta_2) \geq \beta(y_2; x_1, y_1, \theta_1). \tag{54}$$

We will show that if (53) holds, then (54) cannot hold, which yields the desired contradiction. Start by assuming that (53) is true. Now consider any $\check{y} \geq y_1$ for which

$$\beta(\check{y}; x_2, y_2, \theta_2) = \beta(\check{y}; x_1, y_1, \theta_1). \tag{55}$$

Observe that such a $\check{y}$ exists, by continuity of $\beta(y', ., ., .)$ in y'. Assuming (53), if (54) also holds then there must be such a $\check{y}$ by the Intermediate value theorem (if there is none, then we already know that (54) cannot hold).

We will show that at this $\check{y}$ the marginal increase

$$\beta_1(\check{y}; x_2, y_2, \theta_2) < \beta_1(\check{y}; x_1, y_1, \theta_1). \tag{56}$$

Given (53), this immediately implies that $\beta(\hat{y}; x_2, y_2, \theta_2) < \beta(\hat{y}; x_1, y_1, \theta_1)$ for all $\hat{y} > y_1$ (because around any point of equality the right hand side grows strictly faster then the left hand side in $\hat{y}$). Therefore the strict inequality applies in particular also for $\hat{y} = y_2$ and contradicts (54), which completes this part of the proof.

To show equation (56), observe first that the marginal increase of β with respect to $\hat{y}$ in (51) is given by

$$\beta_1(\hat{y}; x, y, \theta) = f_y(x, \hat{y}, \hat{\theta}), \tag{57}$$

where $\hat{\theta}$ is again determined as in (52). So (56) is equivalent to

$$f_y(x_2, \check{y}, \hat{\theta}_2) < f_y(x_1, \check{y}, \hat{\theta}_1) \tag{58}$$

where $\hat{\theta}_1 = \hat{\theta}(\check{y}; x_1, y_1, \theta_1)$ and $\hat{\theta}_2 = \hat{\theta}(\check{y}; x_2, y_2, \theta_2)$ as in (52). To show this, define $\xi(x)$ for all x in resemblance of (51) by the following equality

$$f(x, \check{y}, \xi(x)) - \xi(x) f_\theta(x, \check{y}, \xi(x)) = \beta(\check{y}; x_2, y_2, \theta_2). \tag{59}$$

Substituting (52) into (51) reveals immediately that $\xi(x_2) = \hat{\theta}_2$. By (55) we can replace the right hand side in (59) by $\beta(\check{y}; x_1, y_1, \theta_1)$, and then the same argument establishes that $\xi(x_1) = \hat{\theta}_1$. Implicit differentiation yields $\xi'(x) = \frac{f_x}{\xi(x) f_{\theta\theta}} - \frac{f_{x\theta}}{f_{\theta\theta}}$ where we suppressed the argument $(x, \check{y}, \xi(x))$. Since we want to show (56) or its equivalent (58), and $x_1 > x_2$ by assumption, we have to show that $f_y(x, \check{y}, \xi(x))$ is strictly increasing in x. It is strictly increasing in x if $f_{xy} + f_{y\theta}\xi'(x) > 0$ where we again suppressed the argument $(x, \check{y}, \xi(x))$, which is then equivalent to

$$f_{\theta\theta} f_{xy} - f_{y\theta} f_{x\theta} + f_{y\theta} f_x / \xi(x) < 0.$$

This formula coincides with the strict inequality (46) and therefore the strict inequality (48), i.e., the strict version of inequality (11) which we assumed to hold. This condition establishes that output can be strictly improved by pairing types positively assortatively, which proofs sufficiency.

Part II: necessity. We need to show that (11) is necessary to have PAM under any distribution of types. That is, if it is not true that (11) holds for all (x, y, l, r), then there will be a type distribution for which PAM will not be an equilibrium. Assume that (11) 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 (11) 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 must 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 (11) fails. ■

A.6 Proof of Proposition 2

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 (14). 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

$$f_{x\theta} + f_{y\theta}\mathcal{H}(x)/\theta(x) + f_{\theta\theta}\theta'(x) - f_x/\theta(x) = 0 \tag{60}$$

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 (46) and (47)) 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 (14). ■

A.7 The Non-Homogeneous Production Technology

Let output of the firm be $F(x, y, l, r)$, 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

$$\max_{\tilde{x}, \tilde{l}} F(\tilde{x}, y, \tilde{l}, r) - \tilde{l}w(\tilde{x}) - rv(y). \tag{61}$$

The first order conditions for optimality are

$$\begin{aligned} F_x(x, \mu(x), l, r) - lw'(x) &= 0 && (62)\\ F_l(x, \mu(x), l, r) - w(x) &= 0 && (63) \end{aligned}$$

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:

$$\mathbf{H} = \begin{pmatrix} F_{xx} - lw'' & F_{xl} - w' \\ F_{xl} - w' & F_{ll} \end{pmatrix} \tag{64}$$

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

$$\begin{vmatrix} F_{xx} - lw'' & F_{xl} - w' \\ F_{xl} - w' & F_{ll} \end{vmatrix} > 0. \tag{65}$$

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

$$\begin{vmatrix} -F_{xy}\mu' & -F_{yl}\mu' \\ F_{xl} - w' & F_{ll} \end{vmatrix} > 0 \tag{66}$$

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

$$F_{xy} > \frac{(F_x/l - F_{xl})F_{yl}}{|F_{ll}|}. \tag{67}$$

Observe that this condition is similar to the one we obtained for the homogeneous case, only that now the condition depends on the marginal product F_x and the concavity of F in l, F_{ll}.[32] 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.

A.8 Proof of Proposition 6

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 $G(x, y, s, r) = F(x, y, V^*(x, y, s, r)m(s/V^*(x, y, s, r)), r) - V^*(x, y, s, r)c$. The sorting condition on G has to hold for all (x, y, s, r). Note that any quantity of labor $l = V^*(x, y, s, r)m(s/V^*(x, y, s, r))$ can be sustained by an appropriate choice of (x, y, s, r) as long as $f_l(x, y, l) > c$, which obtains directly from the first order condition for v^*. So conditions on F are restricted to this domain.

For G, we obtain the first order conditions

$$G_y = F_y(x, y, V^*(x, y, s, r)m(s/V^*(x, y, s, r)), r) \tag{68}$$

$$G_r = F_r(x, y, V^*(x, y, s, r)m(s/V^*(x, y, s, r)), r) \tag{69}$$

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

[32] The condition for sorting here depends on F_{xl} which is not the case in condition (11). Of course, there are transformations of (11) that include different derivatives (e.g. F_{xl}), obviously with a less concise and intuitive interpretation.

$$
\begin{aligned}
G_{xy} &= F_{xy} + F_{yl}\frac{\partial(V^*m(s/V^*))}{\partial V^*}\frac{\partial V^*}{\partial x} && (70)\\
G_{sr} &= F_{lr}\left[\frac{\partial(V^*m(s/V^*))}{\partial s} + \frac{\partial(V^*m(s/V^*))}{\partial V^*}\frac{\partial V^*}{\partial s}\right] && (71)\\
G_{ys} &= F_{yl}\left[\frac{\partial(V^*m(s/V^*))}{\partial s} + \frac{\partial(V^*m(s/V^*))}{\partial V^*}\frac{\partial V^*}{\partial s}\right] && (72)\\
G_{xr} &= F_{xr} + F_{rl}\frac{\partial(V^*m(s/V^*))}{\partial V^*}\frac{\partial V^*}{\partial x}. && (73)
\end{aligned}
$$

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

A.9 Proof of Proposition 7

Proof. We can rewrite the firms' original maximization problem (28) as $\max_{\theta(x),\mathcal{R}^y(x)} \int [F(x, y, \theta(x), 1) - C(\theta(x), x)] d\mathcal{R}^y(x)$ where $\mathcal{R}^y(x)$ integrates to unity, with $C(l, x) = \min_{v,q}[cv + vqw(x)]$ s.t. $l = vm(q)$. To hire a given number l of workers of type x, the firm has two choice variables: its queue length or its number of vacancies. The firm chooses them optimally to minimize the costs of achieving l hires, and $C(l, x)$ represents the minimum cost. 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 obtain

$$
w(x)q(x) = \frac{\eta(q(x))}{1 - \eta(q(x))}c. \tag{74}
$$

For commonly used matching functions for which the elasticity is constant, this immediately implies that $q(x)$ is decreasing since $w(x)$ is increasing, and strictly so if the workers are actually hired (otherwise firms could attract better workers at lower costs, violating optimality). In general, the term $\eta(q)/[q(1 - \eta(q))] = m'(q)/[m(q) - qm'(q)]$ is decreasing in q, since the numerator is strictly decreasing and the denominator is strictly increasing in q. Implicit differentiation of (74) implies that $q(x)$ is decreasing. ■

A.10 Proof of Proposition 8

Proof. Consider PAM (likewise for NAM). The vacancy filling probability $m(q)$ is decreasing in x, and under PAM also in y. The vacancy rate $(v/l = 1/m(q))$ is then increasing. However, from Proposition 2, firm size ambiguous is in y. In particular, size is increasing if $\mathcal{H}(x)G_{yl} \geq G_{xr}$ and decreasing if the inequality is reversed. ■

A.11 Efficiency Units of Labor

A special case of our framework is the setting with efficiency units of labor, where firm y hires types $x_1, .., x_n$ at quantities $l_1, .., l_n$ produces output $\tilde{F}(y, \sum_{i=1}^{n} x_i l_i)$ that is strictly increasing and strictly concave in the second argument with appropriate Inada conditions. We claim that the same level of output is replicated in our setting when $F(x, y, l, r) = r\tilde{F}(y, xl/r)$. This is trivially true when there is only one worker type, as all resources are concentrated on this type. Now consider several worker types $x_1, .., x_n$ at quantities $l_1, .., l_n$. The firm allocates its unit amount of resources optimally between them, i.e., solves

$$\max_{r_1,\ldots,r_n\in[0,1]^n}\sum_{i=1}^{n}F(x_i,y,l_i,r_i)\ \text{ s.t.}\sum_{i=1}^{n}r_i=1 \tag{75}$$

$$\Leftrightarrow\quad\max_{r_1,\ldots,r_n\in[0,1]^n}\sum_{i=1}^{n}r_i\tilde{F}(y,x_il_i/r_i)\ \text{ s.t.}\sum_{i=1}^{n}r_i=1. \tag{76}$$

Writing $r_1=1-\sum_{i>1}r_i$ and taking first order conditions to the problem are for all $i=2,...,n$ yields

$$\tilde{F}\left(y,\frac{x_1l_1}{r_1}\right)-\frac{x_1l_1}{r_1}\tilde{F}\left(y,\frac{x_1l_1}{r_1}\right)-\tilde{F}'\left(y,\frac{x_il_i}{r_i}\right)+\frac{x_il_i}{r_i}\tilde{F}'\left(y,\frac{x_il_i}{r_i}\right)=0. \tag{77}$$

An obvious (and also unique) solution to this is to set r_i^* such that factor intensities are equalized, i.e., $\frac{x_1l_1}{r_1^*}=\frac{x_il_i}{r_i^*}$. That is, $r_i^*=\frac{x_il_i}{\sum_{j=1}^n x_jl_j}$. Then the total output that the firm achieves is

$$\begin{aligned}\sum_{i=1}^{n}F(x_i,y,l_i,r_i^*) &= \sum_{i=1}^{n}r_i^*\tilde{F}(y,x_il_i/r_i^*) &(78)\\ &= \sum_{i=1}^{n}\frac{x_il_i}{\sum_{j=1}^n x_jl_j}\tilde{F}(y,\sum_{j=1}^{n}x_jl_j) &(79)\\ &= \tilde{F}(y,\sum_{j=1}^{n}x_jl_j), &(80)\end{aligned}$$

which was to be proven.

A.12 Generalized Efficiency Units of Labor

Consider firm y that hires different worker types $x_1,x_2,...,x_n$ at quantities $l_1,l_2,...,l_n$. Generalized efficiency units are captured by a production function $\tilde{f}(\bar{x},y,L)$ which takes as input the average type $\bar{x}=\sum l_ix_i/\sum l_i$ and the sum of labor $cL=\sum l_i$. In contrast, in our setting there is a production function $f(x,y,l)$ if there is only one worker type, but otherwise output is given by $\max_{r_1,r_2,...,r_n}\sum F(x_i,y,l_i,r_i)$ with $\sum r_i=1$ and $F(x_i,y,l_i,r_i):=r_if(x_i,y,l_i/r_i)$. Clearly, for our production function to satisfy the generalized efficiency unit formulation, it must do so even if only one worker type x_1 is present. So $f(x,y,l)=\tilde{f}(x,y,l)$. This defines f completely, and therefore F is also completely defined. When $\tilde{f}$ is strictly convex in its first argument, our production function does not capture its production under multiple types. In fact, our production function always produces more: $\tilde{f}(\bar{x},y,L)\le\sum_{i=1}^n\frac{l_i}{\sum_{j=1}^n l_j}\tilde{f}(x_i,y,\sum_{j=1}^n l_j)\le\max_{r_1,r_2,...,r_n}\sum r_i\tilde{f}(x_i,y,l_i/r_i)$ with $\sum r_i=1$. The first inequality follows from convexity, and the second would be fulfilled with equality if $r_i=\frac{l_i}{\sum_{j=1}^n l_j}$, but in general will be strict because the maximum is higher. But note that $\max_{r_1,r_2,...,r_n}\sum r_i\tilde{f}(x_i,y,l_i/r_i)$ with $\sum r_i=1$ is the output that a firm in our setup would make (since $F\left(x_i,y,l_i,r_i\right)=r_i\tilde{f}(x_i,y,l_i/r_i)$). Under our production functions the welfare theorems apply, so output must be maximized. If our sorting condition applies, the firm objective is maximized when firms do not hire multiple workers in our setting. With generalized efficiency units the same output can be produced if only one type is hired, but strictly less if multiple types are hired by one firm, so also in that case, firms do not want to mix multiple types and our sorting condition remains sufficient. The condition also remains necessary since its violation leads to breakdown of PAM but not to break-down of pure matching, which can still be replicated even under generalized efficiency units.

A.13 Proof of Proposition 3

Proof. To establish PAM between x_i and y, we can fix all other variables and focus on this dimension. Let $\hat{f}^i(x_i, y, l_i) = g(f^i(x_i, y, l_i), \mathbf{f}^{-i}(\mathbf{x}_{-i}, \mathbf{l}_{-i}, y))$ Then the associated $\hat{F}^i$ is by (2) defined as $\hat{F}^i(x_i, y, l_i, r_i) = r_i \hat{f}^i(x_i, y, l_i/r_i)$ or equivalently $\hat{F}^i(x, y, l, r) = r_i g(f^i(x_i, y, l_i/r_i), \mathbf{f}^{-i}(\mathbf{x}_{-i}, \mathbf{l}_{-i}, y)) = r_i g(F^i(x_i, y, l_i, r_i)/r_i, \mathbf{f}^{-i}(\mathbf{x}_{-i}, \mathbf{l}_{-i}, y))$. Successively doing this for all i leads to output function

$$F = (\Pi_{i=1,\ldots,n} r_i)\, g(F^1(x_1, y, l_1, r_1)/r_1, \ldots, F^n(x_n, y, l_n, r_n)/r_n). \tag{81}$$

For there to be PAM in x_i, y, we require

$$F_{x_i y} F_{l_i r_i} \geq F_{y l_i} F_{x_i r_i} \tag{82}$$

where

$$F_{x_i} = (\Pi_{j=1,\ldots,n, j\neq i} r_j) g_i F^i_{x_i} \tag{83}$$

$$F_{l_i} = g_i F^i_{l_i} \tag{84}$$

$$F_{x_i y} = g_i F^i_{x_i y} + F^i_{x_i} \left[g_{ii} \frac{F^i_y}{r_i} + \sum_{j \neq i} g_{ij} \frac{dF^j}{dy} \right] \tag{85}$$

$$F_{l_i r_i} = g_i F^i_{l_i r_i} + g_{ii} F^i_l \left[-\frac{F^i}{r_i^2} + \frac{F^i_{r_i}}{r_i} \right] \tag{86}$$

$$F_{y l_i} = g_i F^i_{y l_i} + F^i_l \left[g_{ii} \frac{F^i_y}{r_i} + \sum_{j \neq i} g_{ij} \frac{dF^j}{dy} \right] \tag{87}$$

$$F_{x_i r_i} = g_i F^i_{x_i r_i} + g_{ii} F^i_{x_i} \left[-\frac{F^i}{r_i^2} + \frac{F^i_{r_i}}{r_i} \right] \tag{88}$$

and g_i denotes the partial derivative of g with respect to its $i'th$ argument. Then we can write the condition for PAM in x_i, y as:

$$\left[g_i F^i_{x_i y} + F^i_{x_i} \left[g_{ii} \frac{F^i_y}{r_i} + \sum_{j \neq i} g_{ij} \frac{dF^j}{dy} \right] \right] \left[g_i F^i_{l_i r_i} + g_{ii} F^i_l \left[-\frac{F^i}{r_i^2} + \frac{F^i_{r_i}}{r_i} \right] \right] \tag{89}$$

$$\geq \left[g_i F^i_{y l_i} + F^i_l \left[h_{ii} \frac{F^i_y}{r_i} + \sum_{j \neq i} g_{ij} \frac{dF^j}{dy} \right] \right] \left[g_i F^i_{x_i r_i} + g_{ii} F^i_{x_i} \left[-\frac{F^i}{r_i^2} + \frac{F^i_{r_i}}{r_i} \right] \right] \tag{90}$$

or

$$(g_i)^2 \left[F^i_{x_i y} F^i_{l_i r_i} - F^i_{y l_i} F^i_{x_i r_i} \right] + g_i g_{ii} \left[-\frac{F^i}{r_i^2} + \frac{F^i_{r_i}}{r_i} \right] \left[F^i_{x_i y} F^i_l - F^i_{y l_i} F^i_{x_i} \right] \tag{91}$$

$$+ g_i \left[g_{ii} \frac{F^i_y}{r_i} + \sum_{j \neq i} g_{ij} \frac{dF^j}{dy} \right] \left[F^i_{l_i r_i} F^i_{x_i} - F^i_{x_i r_i} F^i_l \right] \geq 0. \tag{92}$$

This condition has to hold along the equilibrium path, otherwise the second-order condition for optimality is violated. If the condition holds everywhere, we have proven earlier that this is sufficient for optimality for the given subunit. ■

A.14 Proof of Proposition 4

Proof. Consider a subunit i. If $dF^j/dy \geq 0$ for all $j \neq i$, then the conditions on g together with (23) and (25) insure that (22) holds for all $(\mathbf{x}, \mathbf{l}, \mathbf{r}, y)$. This is easiest seen by noticing that CRTS implies that $F^i = l_i F^i_l + r_i F^i_r$ so that $F^i_l \frac{l_i}{r_i} = F^i / r_i - F^i_{r_i}$, which can be substituted into (25). (22) then follows immediately. Since we can do this for all subunits i, this implies PAM everywhere as long as each subunits believes that the other subunits expand output with productivity.

So we simply must check that under PAM it is indeed optimal for each i to increase output with productivity. Note that output changes according to

$$dF^i/dy = F^i_x \nu'_i + F^i_y + F^i_l l'_i \nu' \tag{93}$$

$$\geq F^i_x \nu'^i + F^i_y + F^i_l \frac{\mathcal{H} F^i_{yl} - F^i_{xr}}{F^i_{lr}} \nu', \tag{94}$$

where the last step follows because size $l'_i(x)$ in equilibrium can be determined in similar ways as $\theta^{\acute{i}\prime}(x)$ in condition (60). The only difference is that when “totally differentiating (7)” additional terms arise that have to do with dF^j/dy in the other subunits. Under the beliefs that these are increasing, this leads to additional forces that also point towards larger firms, leading to an even larger response than in the basic model. Clearly (24) ensures that $dF^i/dy \geq 0$ no matter what the ratio of types $\mathcal{H}$. ■

A.15 Proof of Corollary 2

Proof. Without loss, we drop the superscript i. B being Cobb-Douglas can then be interpreted as CES with elasticity of substitution $\varepsilon_B = 1$. Then

$$\frac{B B_{lr}}{B_l B_r} = \frac{1}{\varepsilon_B} = 1. \tag{95}$$

Therefore $F^i_{l_i r_i} F^i_{x_i} - F^i_{x_i r_i} F^i_l = A B_{lr} A_x B - A_x B_r A B_l = A A_x (B_{lr} B - B_r B_l) = 0$, and the third term in the main condition (22) becomes zero. Then the main condition (22) becomes:

$$(g_i)^2 \left[A_{xy} A B B_{lr} - A_x A_y B_l B_r \right] - g_i g_{ii} \phi A \frac{B}{r^2} B B_l \left[A_{xy} A - A_x A_y \right] > 0. \tag{96}$$

The term $A_{xy} A B B_{lr} - A_x A_y B_l B_r > 0$ is implied by (26). In the second term, the expression in brackets $A_{xy} A - A_x A_y > 0$ follows from (26) and (95). With $g_{ii} < 0$, the second term is positive, a sufficient condition for the whole expression (96) to be positive. ■

Appendix B 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 $a = A p_a \kappa$, 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. Instead, we take the rental rate 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 2.

Table 2: 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

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 5 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 ($x \approx 1$).

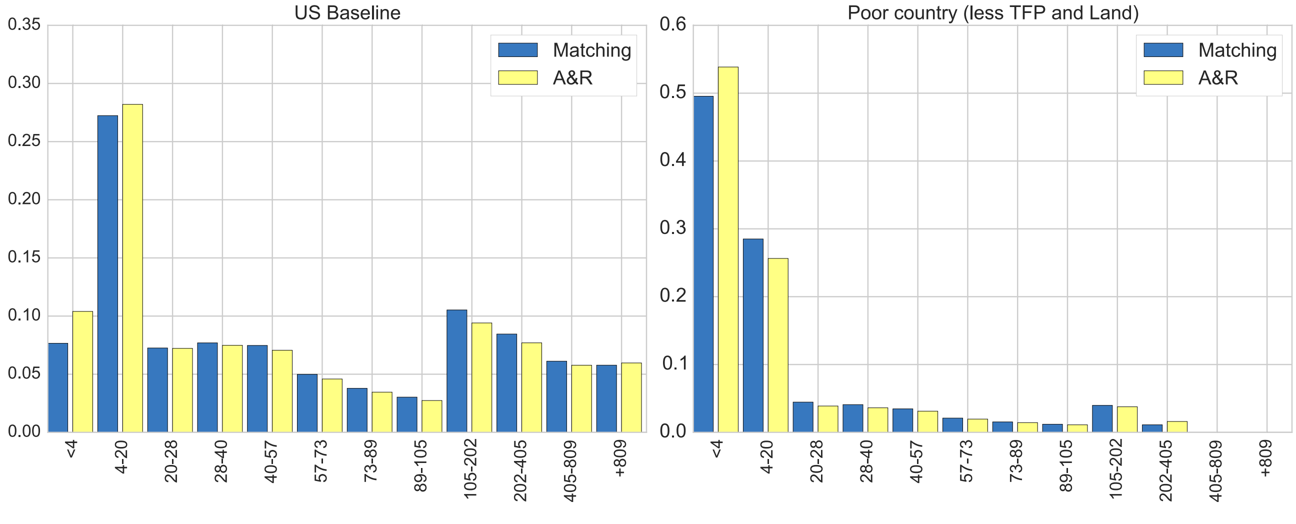

Figure 5: Replication of the firm size distribution in Adamopoulos and Restuccia (2014) with our computational algorithm with negligible spread in x and $\rho = .25$.

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 (11) directly to see whether sorting is PAM; alternatively we can use the envelope condition developed in Extension II below in this Appendix and PAM arises if

$$\{F_{xy}F_{lr}F_{kk} - F_{xy}F_{lk}F_{rk} - F_{xk}F_{yk}F_{lr}\} - \{F_{xr}F_{yl}F_{kk} - F_{xr}F_{yk}F_{lk} - F_{xk}F_{yl}F_{rk}\} \quad > \quad 0 \tag{97}$$

$$\Leftrightarrow \frac{0.00272284375A^3p_a^3r^2\,(kx)^{0.5}\left(\frac{ly}{r}\right)^{0.25}}{k^2lxy}\left(0.89\,(kx)^{0.25} + 0.11\left(\frac{ly}{r}\right)^{0.25}\right)^{3.0} \quad > \quad 0, \tag{98}$$

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 (14) 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 obtain 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.

Appendix C Extensions

C.1 Extension I. 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.[33] 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 (11). 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 (46) 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

[33] 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 aggregate 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.

for positive assortative matching

$$\left[\rho\tilde{F}_{xy}+(1-\rho)(\tilde{F})\frac{\partial^2\ln\tilde{F}}{\partial x\partial y}\right]\left[\rho\tilde{F}_{lr}-(1-\rho)l\tilde{F}\frac{\partial^2\ln\tilde{F}}{\partial l^2}\right] \tag{99}$$

$$\geq \left[\rho\tilde{F}_{yl}+(1-\rho)\tilde{F}\frac{\partial^2\ln\tilde{F}}{\partial y\partial l}\right]\left[\rho\tilde{F}_{xr}+(1-\rho)\left(l\tilde{F}_{xl}-l\tilde{F}\frac{\partial^2\ln\tilde{F}}{\partial x\partial r}\right)\right]. \tag{100}$$

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 (11). Otherwise, log-supermodularity also appears in the condition.

C.2 Extension II. 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.

C.3 Extension III. 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 equal 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 (14) to construct the type distribution after entry at all lower types.

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.

References

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.

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.*

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 B. Jovanovic (2011): "Occupational choice and development," *Journal of Economic Theory*, 147(2), 657–683.

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.

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

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), 915–938.

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.

Jovanovic, B. (1982): "Selection and the Evolution of Industry," *Econometrica*, pp. 649–670.

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.

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.

Lydall, H. F. (1959): "The distribution of employment incomes," *Econometrica: Journal of the Econometric Society*, pp. 110–115.

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.

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.

Simon, H. A. (1957): "The compensation of executives," *Sociometry*, 20(1), 32–35.

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.