Assortative Matching with Large Firms:
Span of Control over More versus Better Workers[*]

Jan Eeckhout[†]and Philipp Kircher[‡]

May 2012

Abstract

In large firms, management resolves a trade off between hiring more versus better workers. The span of control or size is therefore intimately intertwined with the sorting pattern. Span of control is at the center of many studies in macroeconomics, comparisons of factor productivity, trade, and labor. With heterogeneous workers, we analyze the worker assignment, firm size, and wages. The pattern of sorting between workers and firms is governed by an intuitive cross-margin-complementarity condition that captures the complementarities between qualities (of workers and firms), and quantities (of the work force and firm resources). A simple system of two differential equations determines the equilibrium allocation, firm size and wages. We can analyze the impact of technological change: skill-biased change affects wages and increases the skill premium; quantity-biased change affects firm size, especially of productive firms. We also introduce search frictions and investigate how unemployment varies across skills and how vacancies vary across firm size.

Keywords. Span of Control. Sorting. Firm Size. Wage Distribution. Skill-biased Technological Change. Unemployment. Supermodularity.

[*] We would like to thank numerous colleagues for insightful discussions and comments. Stefanie Huber provided invaluable research assistance. Eeckhout gratefully acknowledges support from the ERC, grant 208068, and from ECO2009-12157, and Kircher by the ERC, grant 184119.

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

[‡] Department of Economics, London School of Economics, and CEPR, `P.Kircher@lse.ac.uk`.

1 Introduction

Span of control – the number of workers under the control of management within a firm – attributes an essential role to the firm in economics. In the canonical macroeconomic context, firms predominantly make quantity decisions. Endowed with different management, technologies, or capital, companies choose the span of control accordingly, and this has important implications for the size of firms (Lucas 1978, Hopenhayn and Rogerson 1993). This labor factor intensity decision is both realistic and a convenient modeling device. It has been invoked to explain differences between countries (Restuccia-Rogerson, 2008), and to analyze technology adoption in evolving firms (Mortensen-Lentz, 2005, Jovanovic, 1982).

Yet, firms typically face a more complex tradeoff. They simultaneously choose the *quality* of the workers as well as the quantity. The retail arm of a company like Radioshack 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.[1] In the standard frictionless matching model (Becker, 1973), each firm consists of exactly one job, just as in most of the matching models with search frictions.[2] This 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*.

The aim of this paper is to investigate sorting in an otherwise conventional macro environment where firms simultaneously choose the quality as well as the quantity of the work force. This provides a much richer role for the firm and its span of control or size. For example, we shed light on why the high skilled management in a Walmart store has an enormous span of control over relatively low skilled workers, while in mom-and-pop retail stores the span of control is small and skills of both managers and workers are average. Or, we can evaluate what the consequences are of information technology that improves the ability to manage many workers, such as monitoring and GPS tracking devices. These examples illustrate how we address general questions: Are more productive firms larger? Do they hire better workers; or both? How does this affect managerial compensation and firm profits? And how does it depend on the characteristics of the particular industry or country that we are considering?

We formalize the tradeoff between the intensive and the extensive margin in a very simple framework. Firms differ in the quality of their endowment, such as managerial skill in a modern economy or quality of their arable land in an agrarian economy. In the spirit of the previous work in this literature, firms compete in the same industry and produce a homogeneous output good, but nevertheless the best firms do not employ all the labor because of decreasing returns. With homogeneous workers, this replicates the Lucas model with an intensive margin only. Better firms have a comparative advantage and span their control over more workers. With heterogeneous workers, a firm simultaneously chooses the worker type and its span of control. If decreasing returns are so stark that each firm optimally chooses only one worker, then the model replicates the standard Beckerian matching model. Only the extensive margin matters and all firms have the same size. The general specification gives rise to a rich but tractable framework to study the interaction between firm size and the skill of the workforce.

The paper contributes to existing work in four ways. First, we ask which workers are hired by which firms. We find a surprisingly simple condition for assortative matching that captures both the quality and quantity considerations. This condition is new and compares the different degrees of complementarity[3] along four margins:

[1] The canonical matching model has also extensively been used in the international trade literature, see amongst others Grossman and Maggi (2000), Grossman (2004) and Costinot (2009).

[2] Amongst many others, see Burdett and Coles (1997) and Shimer and Smith (2000).

[3] We will use the term complementarity and supermodularity interchangedly. For our purposes, it can best be thought of as the

(1) *type complementarity* captures the interaction between firm and worker types. Clearly, if better firms receive a exceptionally high return only from better workers, then they will end up hiring those workers. This is the only effect present in standard quality-sorting models in the spirit of Becker (1973). Additionally, there is the (2) *complementarity in quantities* of workers and resources, just as in the standard model with quantity choices only. There is the (3) *span-of-control complementarity* between the firm or manager type and the number of workers that features in Lucas (1978); how much of a higher marginal product do better managers have from supervising more workers of a given skill? Finally, there is the (4) *managerial resource complementarity*, the complementarity between worker skills and managerial or firm resources: do better workers have a higher marginal product of receiving more supervision time? A simple tradeoff between these four forces determines the pattern of sorting. It characterizes the efficient equilibrium outcome and is a measure of the efficiency losses that would result from misallocation.

Second, we can precisely pin down the composition of the work force across different firm types, i.e., how firms resolve the tradeoff of span of control over more versus better workers. The equilibrium allocation of types and quantities is entirely governed by a simple system of two differential equations. In particular, this gives a prediction for the firm's span of control, and therefore, for the firm size distribution. For example under positive assortative matching, better management supervises larger groups (better firms are larger) provided the span of control complementarity (3) outweighs the managerial resource complementarity (4).

Even if the theory is very stylized, it allows us to revisit our example of the retail industry where high productivity companies such as Walmart have high skilled management and hire many mainly low skilled workers, compared to the smaller mom-and-pop stores. This indicates negative sorting together with a size distribution that exhibits a density of workers that is increasing in firm productivity. In the light of our theory, this is consistent with a type complementarity (1) that is small in this industry, whereas the span of control complementarity (3) is large: while the complementarity between managers and workers is small, at the margin management in better firms is much better at managing large groups than management in low productivity firms. This may be the case for example because cash registers and inventories are nearly trivial to operate, while the firm heavily invests in management and control tools that allow the supervision of many workers through centralized information on performance on all registers and inventories. Not only does this lead to negative sorting, but also to a sharply increasing firm size in productivity, thus creating very large firms at the top. They do not need to spend much time with each employee because supervision time does not increase productivity much, and they have the tools to supervise many. Instead, in management consulting, with strong complementarities in manager and subordinate skill ((1) large) but moderate span of control technologies ((3) moderate), there is positive sorting. Top firms are only larger than bottom firms if their span of control (3) outweighs the benefits from training and interacting with employees (4). Given that the two counteract, top consulting firms tend to be only moderately larger than other firms in the industry.

The Walmart example clearly illustrates the role of technological change. Their way of doing business now is very different from how the retail sector worked half a century ago. Over time, information technology and investment in knowledge dramatically changes the production process. We can therefore analyze how technological change affects the firm size distribution and the composition in the work force. Skill-biased technological change is usually viewed as a change that makes the complementarity between worker skill and firm technology larger (see for example Krusell, Ohanian, Rios-Rull, and Violante, 2000). But much of technological change is in terms of information technology that changes the complementarity between manager skill and the amount of workers he supervises. This is quantity-biased technological change. In our model, skill-biased and quantity-biased

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).

technological change have discernibly different implications. Skill-biased technological change has a big impact on the wage premium and therefore the wage distribution. Yet, it does not affect much the size distribution or the equilibrium allocation. Instead, quantity-biased technological change leads to a sharp increase in the size of large firms, and a decrease in the size of small firms. As a result, it affects the firm size distribution, but not necessarily the wage distribution.

The paper contributes to existing work in a third way. We integrate labor market frictions into the model by means of directed search. The general setup is sufficiently flexible to allow for us to introduce directed search frictions. Ours is the first model that can analyze unemployment in the presence of *both* sorting *and* firm size simultaneously. This is important because in the empirical analysis of matched employer-employee data, both features are essential. We show that irrespective of the sorting pattern, unemployment rates are lower for more skilled workers, which is consistent with the empirically observed unemployment patterns. Instead, the vacancy rates across firm productivity levels are ambiguous. It depends on how firm size varies, which in turn is governed by the strength of the span of control complementarity relative to the managerial resource complementarity.

As a final contribution, we point out that our theory provides a unifying framework for previous models, most of which are special cases of ours. Clearly, Becker (1973) and Lucas (1978) are special cases. A general setup was also proposed in Rosen (1982), but solved only for a functional form that is a special case of our model, that of efficiency units of labor. Our setup also includes as special or limiting cases the functional forms of several existing models in this line of research such as Sattinger (1975), Garicano (2000), Antràs, Garicano and Rossi-Hansberg (2006), and Van Nieuwerburgh and Weill (2010). We can also adjust the setup to match the features of the Roy model (Heckman and Honore, 1990). Here we consider the competitive equilibrium outcome of the general model. For applied purposes the generality of our model has the advantage that it can be estimated without taking a strong stance on the span-of-control specification, and then it can be checked ex-post to which extent the resulting estimates capture any of the special cases that it encompasses and that we discuss in detail in Section 3.3.[4]

The specific production structure of Garicano (2000) has also been successfully applied in the context of international trade and offshoring (e.g., Antràs, Garricano and Rossi-Hansberg, 2006). Since our approach captures it as a limiting case, it might prove useful as a broader model for studying how both the size and the skill composition of heterogeneous firms change when international integration changes the market structure.[5] In our setting the size of the firm is limited by decreasing returns to scale in production due to scare managerial resources, which offers a complimentary channel to the usual Dixit-Stiglitz reasoning based on limited demand for each variety (eg, Costinot, 2009). Our model can be studied without functional form assumptions, yet it can also be integrated into a Dixit-Stiglitz type framework, as we show in our extensions. Also, unemployment has attracted recent interest in trade settings (eg, Helpman, Itskohki and Redding, 2011), and our model with unemployment can readily be applied there as well.[6]

Finally, our approach might prove useful to understand cross-country differences in total factor productivity, which Restuccia-Rogerson (2008) and Hsieh-Klenow (2010) explain in the context of a Lucas span of control model of firm heterogeneity. There are huge differences across firm productivity and firm size is an important determinant explaining those. Yet, large differences remain that cannot be explained by the model and that are often attributed to frictions in financial markets or to differences in taxation. By introducing worker heterogeneity

[4] Most applied papers currently specify a specific span-of-control version. For example, amongst others Garicano and Hubbard (2008) and Roys and Seshadri (2012) consider the Garicano (2000) and Garicano and Rossi-Hansberg (2006) model, Fox (2010) and Sattinger (2008) consider versions of the models in Sattinger (1975, 1979), and Terviö (2008) and Gabaix and Landier (2007) consider the one-to-one matching model.

[5] Grossman, Helpman and Kircher (2012) embed a specific variant of this structure within a Heckscher-Ohlin model to understand the effects of trade openness on the size distribution, the matching outcome, the sorting across industries, and the wage distribution.

[6] In Helpman, Itskohki and Redding (2011) firms are heterogeneous and also workers draw a heterogeneous match quality once they meet a firm, but in their setting workers are ex-ante identical and earn identical expected-payoffs since heterogeneity only arises as a shock, while in many trade settings we would like to start from a situation where workers of different types exist in the population.

and sorting in this otherwise standard model, differences in skill distributions across countries may contribute to the explanation of cross-country differences, on top of the differences in the size distribution of firms. For example, the skill distribution in China and India is very different from that in the US. Under the equilibrium allocation, the resulting size distribution of firms in those countries will differ substantially even in the absence of frictions.

Related Literature

Our model relates to several strands of existing literature. Here we single out the four that are most relevant. Once we have laid out the model and derived the results, we discuss more formally in Section 3.3 how our framework captures and extends a number of the existing models in the literature that we mention here. In particular, there we make explicit the way in which those can be viewed as special cases of our setup.

The most common one-to-one matching models originating from Kantorovich (1942), Koopmans and Beckmann (1957), Shapley and Shubik (1971), and Becker (1973), restrict attention to settings where agents have to be matched into pairs, with the obvious limitation that these papers do not provide insights into the size of the firm and its capital intensity. They constitute a special case of our model with extreme decreasing returns. Notice that the matching models by Terviö (2008), and Gabaix and Landier (2007) 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. A number of both early and recent contributions have focussed on environments where managers can supervise more than one worker.[7] In Sattinger (1975), each employed worker type produces one unit of output, but requires supervision-time that depends on the manager type in a decreasing relation. A related structure arises in Garicano (2000), Garicano and Rossi-Hansberg (2006), and Antràs, Garricano and Rossi-Hansberg (2006). These models have the feature that both the quantity and quality of workers play a role, but in a rather stark manner where additional supervision time above the minimum has no additional benefits. Their conditions again arise as limiting cases of our model.

Rosen (1982) proposes a general setup with worker heterogeneity, where quantity and quality interact multiplicatively, which is a special case of ours. Rosen never solves his general setup, but assumes a functional form that guarantees perfect substitutability, i.e., workers of a given type generate exactly the same output as twice as many workers of half that type. This assumption is now commonly referred to as efficiency units of labor, and it is well-known that it generates no sorting implications. In contrast to his work and to some of the other contributions, we do not endogenize the number of hierarchies, though.

While the assortative matching literature has made rather specific assumptions for multi-worker firms that we attempt to generalize, the combinatorial matching and general equilibrium literature has instead stayed general but has focussed mainly on existence theorems rather than on characterizing the sorting or the wage patterns. The classic example in the combinatorial matching literature is Kelso and Crawford (1982). They propose a many-to-one matching framework in a finite economy and allow for arbitrary production externalizes across workers within the firm. While it is well-known that the stable equilibrium or the core may not exist, they derive a sufficient condition for existence, that of gross substitutes: adding another worker decreases the marginal value of each existing worker. This condition is satisfied in our setting since output is assumed to be concave in the number of 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, Milgrom and Hatfield (2005) analyze package bidding

[7] A different strand of literature has relaxed other parts of the one-to-one matching model, for example by considering non-transferable utility of information frictions. This retains the main resource constraint of one-to-one matching, though, while many-to-one matching requires a different constraint.

as a model of many to one matching.

Our model differs from settings such as the Roy (1951) model and its recent variants in e.g., Heckman and Honore (1990), where each firm (or sector) can absorb unbounded numbers of agents. In our setup, the marginal product decreases as the firm gets larger. Some models combine the Roy model with a demand by consumers that entails a constant elasticity of substitution (CES), which implies that the price falls when more workers produce output in a particular sector (see recently Costinot (2009)). The difference is that in such settings no agent internalizes the fact that the price falls when more output is produced. In section 4.3, we allow for a CES demand structure which results in a model of imperfect competition similar to Dixit and Stiglitz (1977), only that now two-sided heterogeneity and an extensive margin are allowed.

Finally, the extension to search frictions is linked to recent developments in the literature. In one-to-one matching, sorting has been integrated in models of search, see for example Shimer and Smith (2000), Shi (2001), Shimer (2005), Atakan (2006), and Eeckhout and Kircher (2010). Some of the techniques of the latter are useful also for our current setup, even though firm size was never a consideration there. Firm size has been modeled in different ways in the canonical search framework. The key challenge has been the wage setting mechanism. Smith (1999) resolves this with multi-agent sequential Nash Bargaining, whereas Hawkins (2011) and Kaas and Kircher (2011) determine market prices by means of wage posting and directed search. We use the latter. The real novelty of our approach here relative to the existing search literature is to combine *both* sorting under two-sided heterogeneity *and* firm size. This allows us to provide general conditions on the variation of the unemployment rate by skills and the vacancy rate by firm size. This is important for the estimation of search models using matched employer-employee data that feature both sorting and firm size variation.

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.

Agents. The economy consists of heterogeneous 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. Also firms are heterogeneous in terms of some proprietory input into production that is exclusive to the firm, such as scarce managerial talent or particular proprietary capital goods. In a modern business setting, this is the time endowment of an entrepreneur that he spends interacting with and supervising his employees. In an agricultural economy this is the amount of land that an agricultural firm possesses. Firms are indexed by their productivity type $y \in \mathcal{Y} = \mathbb{R}_+$, where $H_f(y)$ denotes the measure of firms with type below y. Unless otherwise stated, we focus on distributions H_f and H_w with non-zero continuous densities h_f and h_w on the compact subsets $[\underline{x}, \overline{x}] \subset \mathcal{X}$ and $[\underline{y}, \overline{y}] \subset \mathcal{Y}$, respectively, but especially for our main characterization result we also provide a proof for arbitrary distribution functions.

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

$$F(x, y, l_x, r_x)$$

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 total resources at the firm level r are fixed. Without loss we can therefore normalize $r = 1$. The firms can allocate resources over different skilled worker types x as long as its choice r_x satisfies the feasibility constraint $\int_{\mathcal{X}} r_x dx = 1$.

We will focus on a particular class of functions F, where the output of each worker depends only on his own type x, the type of the firm y, and the factor intensity r_x/l_x that each of the workers obtains. Total output F then has constant returns to scale in the quantity variables: output doubles when both the quantity of resources and of workers are doubled. We retain the assumption of constant returns to scale in the quantity variables throughout. Therefore, if we denote $\theta_x = l_x/r_x$, we can write output per r_x units of resources as[8]

$$f(x, y, \theta) := F(x, y, l_x/r_x, 1).$$

This also represents the production function of a firm that only hires one type of worker, in which case θ represents the firm's size. We also assume that F is twice continuously differentiable, and that it is strictly concave in each of the quantity variables at all interior points of the type space, and that it is strictly positive only if both quantity variables are strictly positive.[9] Even though we often refer to higher types as "better" types, we do not need to make any assumptions with respect to the quality variables except twice-differentiability to obtain our results.

Finally, for a firm that hires several worker types we assume that its total output is the sum of the outputs across all its worker types: firm y's total output is $\int F(x, y, l_x, r_x)dx$. This is an important assumption because it rules out complementarities between different worker types. The motive for this assumption is tractability. Abstracting from this one source of complementarity is obviously restrictive. It does allow us to solve the model (see Kelso and Crawford, 1982), and make progress in analyzing all the other cross-complementarities between quantities and qualities.

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

Firm optimality in a frictionless competitive market requires that a firm of type y maximizes its output minus wage costs as follows:

$$\max_{l_x, r_x} \int [F(x, y, l_x, r_x) - w(x)l_x]dx \tag{1}$$

where r_x can be any probability density function over x. Factoring out r_x from the square bracket reveals that the interior depends only on the factor intensity $\theta = l_x/r_x$, which can be freely chosen at any level in $\Theta = \mathbb{R}_+$ by adjusting the labor input appropriately. Because output across different types is additive, optimality requires that the firm places positive resources only on combinations of $x \in \mathcal{X}$ and $\theta \in \Theta$ that solve[10]

$$\max_{\tilde{x}, \tilde{\theta}} f\left(\tilde{x}, y, \tilde{\theta}\right) - \tilde{\theta} w(\tilde{x}). \tag{2}$$

If there is only one such combination that solves this maximization problem, then the firm will hire only one

[8] If $F(x, y, l, r)$ has constant returns to scale, we can write it as $F(x, y, l, r) = lF(x, y, 1, r/l)$, so that $g(x, y, r/l) = F(x, y, 1, r/l)$ represents the output per worker. This makes clear that our output function can be derived from a process where the output of each worker depends on his own skill, the skill of the manager, and the amount of resources allocated to him. Interaction among agents are present only between managers y and workers x, and between worker types x to the extent that more resources to one worker type might imply less resources to another. Alternatively, we can write output as $F(x, y, l, r) = rF(x, y, l/r, 1)$ and define $f(x, y, l/r) := F(x, y, l/r, 1)$ as the output per unit of resource. In our exposition we work with the latter, ie, from the firm's perspective, which is convenient in many derivations.

[9] Requiring strict concavity only at interior points allows for example the specification $F(x, y, l, 0) = 0$, which generates only weak concavity in l. The requirement that $F(x, y, l, 0) \leq 0$ is made for convenience as it rules out that workers are hired by firms that devote no resources to them.

[10] Problem (1) is equivalent to $\max_{r_{(\cdot)}} \int (r_x \max_{\theta_x} [F(x, y, \theta_x, 1) - w(x)\theta_x]) \, dx$, where $\theta_x = l_x/r_x$ can be adjusted through appropriate hiring of workers. Clearly, resources are only devoted to combinations of x and θ that maximize (2).

worker type, allocate all resources to this type, and hire an amount of labor $l = \theta$. Both firms and workers can abstain from the market and obtain a payoff normalized to zero, which means that profits and wages cannot fall below this level.[11]

Feasibility of the allocation implies that firms attempt to hire no more workers than there are in the population. Denote by $\mathcal{R}(x, y, \theta)$ the resource allocation in the economy, which describes the amount of resources that firms with a type below y devote to workers of a type below x that are employed with a factor intensity $l_x/r_x \leq \theta$. We use the convention that $\mathcal{R}(x, y, 0) = 0$. This implies that firms that do not want to hire any worker ($\theta = 0$) are counted as "unmatched" rather than employing zero workers of type x. This will only be important when we consider the definition of assortative matching. Let $\mathcal{R}(y|\mathcal{X}, \Theta)$ denote the marginal over y when the other two variables can take any value in their type space. It denotes the amount of resources used by firms with type below y. Similarly, let $\mathcal{R}(\theta, x|\mathcal{Y})$ be the resources spent by all firms on workers of type below x employed with intensity less than θ. Weighted by the intensity this yields the number of workers hired, so $\int_\Theta \theta d\mathcal{R}(\theta, x|\mathcal{Y})$ gives the number of workers hired with type up to x. Feasibility requires that these cannot exceed the number of agents in the population, ie, for all $y' < y$ and $x' < x$ it has to hold that

$$\mathcal{R}(y|\mathcal{X}, \Theta) - \mathcal{R}(y'|\mathcal{X}, \Theta) \quad \leq \quad H_f(y) - H_f(y') \tag{3}$$

$$\int_\Theta \theta d\mathcal{R}(\theta, x|\mathcal{Y}) - \int_\Theta \theta d\mathcal{R}(\theta, x'|\mathcal{Y}) \quad \leq \quad H_w(x) - H_w(x'). \tag{4}$$

We can now define an equilibrium as follows:

Definition 1 *An equilibrium is a tuple* $(w,\mathcal{R})$ *consisting of a non-negative hedonic wage schedule* $w(\cdot)$ *and a feasible resource allocation* $\mathcal{R}$ *such that*

1. *Optimality:* $(x, y, \theta) \in supp\mathcal{R}$ *only if it satisfies (2).*
2. *Market Clearing:* (4) *holds with equality if wages are strictly positive a.e. on* $(x', x]$.[12]

The market clearing condition simply states that if wages for some worker types are positive, their markets clear. Existence of an equilibrium when output is multiplicatively separable in quantity and quality variables and bounded has been proven, e.g., in Jerez (2012). Her proof technique can be adapted to the more general setting, and we provide an explicit construction of an equilibrium for our economy when our sorting conditions are satisfied. Our main focus in this work, though, 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. Let $\mathcal{R}(x, y|\Theta)$ be the marginal distribution of $\mathcal{R}$ over the firm and worker types at any level of intensity. It denotes the amount of resources devoted by firms with type below y to workers of skill below x, and we refer to it as the type allocation. Matching is positive assortative if (x, y) in the support of the type allocation implies that (x', y') is not in its support unless either both dimensions are weakly larger or both weakly smaller than (x, y). This means that strictly higher firm types never hire strictly lower worker types. Similarly, matching is negative assortative if (x, y) in the support of the type allocation implies that (x', y') is not in its support unless one dimension is weakly larger and one is weakly smaller than (x, y). In such a case strictly higher firm types never hire strictly higher worker types. For some derivations it will be particularly useful to focus on

[11] This is indeed simply a normalization. Consider true outside options $o(x)$ and $q(y)$ for firms and workers and output function $\tilde{F}(x, y, l_x, r_y)$. These can be incorporated into our framework by considering a normalized production function of form $F(x, y, l_x, r_y) = \tilde{F}(x, y, l_x, r_y) - l_x o(x) - r_y q(y)$ that captures the loss in outside option due to matching.

[12] This definition is suitable for type distributions without mass points. More generally applicable is the following generalization: $\int_{\theta\in\Theta, x\in\mathcal{A}} \theta d\mathcal{R}(\theta, x|\mathcal{Y}) = \int_{x\in\mathcal{A}} dH_w(x)$ for any measurable set $\mathcal{A}$ with $w(x) > 0$ H_w-a.e. on $\mathcal{A}$.

(strictly) differential assortativeness, where the support of $\mathcal{R}$ is concentrated only on points $(x, \mu(x), \theta(x))$ for some differentiable functions θ and μ, with the former strictly positive almost everywhere and the latter (strictly) monotone.

Alternative interpretations of our setup. In our exposition we assume that the number of firms is fixed, they each own a unit measure of a scarce resource and allocate it to the different workers that they hire. Only the workers are traded in the market. This is inspired by the idea of span of control of a line manager, whom has one unit of time for supervision, and who hires workers.

It might be worthwhile to note that there are alternative ways to set up our model that lead to identical results for sorting and factor prices. In our setup, we assumed that firms "buy" workers at wage $w(x)$. We could have chosen a different setup where workers buy resources for production at some endogenous price schedule $v(y)$. It turns out that our equilibrium profits according to (2) coincide with the equilibrium price $v(y)$ that arises in the alternative model where workers buy resources.

Finally, we could assume that there are both resource owners and workers, and both workers and resources are traded in the market at endogenous prices $v(y)$ and $w(x)$, respectively. Both workers and resources can be put together to produce output. Anybody can set up a production entity and make profits $\max_{x,y,l,r} F(x, y, l, r) - lw(x) - rv(y)$, which in equilibrium has to equal zero due to free entry, and demand has to equal supply of both workers and resources. Again, in equilibrium of this alternative model the wages are the same as in our equilibrium and the price of resources equals the firms' profits in our setup. In fact, this setup is identical to ours, only that we assumed that the unit measure of resources is tied to a particular manager who runs the firm and reaps as profits the price of his resource.

Even within our exposition the production function can be interpreted in broader terms. First, we interpreted r as the fraction of the firm's resources, implicitly using a unit measure of resources for each firm. This is natural in the example of managerial time, but in many other settings firms differ in their endowments. It turns out that this is easily captured in our setting, since the unit restriction in terms of resources is a normalization.[13] We can also accommodate a setting where firms can acquire additional resources.[14]

Additionally, one might want to follow many macroeconomic models and include some kind of generic capital good that can be bought in the world market for price i per unit and enters the production function as another factor. We return to this extension in Section 4. There we also cover the case where firms have to post vacancies in a frictional (competitive) search market, and the firm has to determine how many vacancies to post in order to attract the right level workers into production. This framework also allows us to incorporate unemployed workers in a large firm model with heterogeneity.

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 derive necessary and sufficient conditions for assortative matching and characterize the assortative equilibrium.

[13] If firms of type y have $T(y)$ resources and produce $\tilde{F}(x, y, l, t)$ by using t units of them, we can express this in terms of the fraction r of their resources: $F(x, y, l, r) = \tilde{F}(x, y, l, rT(y))$.

[14] If firms can create a unit of resources at cost $c(y)$, then in the ensuing equilibrium after resources are created the equilibrium profit per unit of resource of type y has to equal $c(y)$. It turns out that this makes it particularly easy to construct an equilibrium.

3.1 Assortative Matching with Large Firms

In order to build intuition for our main proposition, it will be useful to focus first on differential assortativeness which allows us to make the derivation of our main condition transparent. Assume that the equilibrium is assortative, supported by some differentiable assignment function $\mu(x)$ and intensity $\theta(x) > 0$. By (2) this means that $(x, \theta(x))$ are maximizers of the following problem for a firm of type $y = \mu(x)$:

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

Assortative matching means that each firm only hires one type, and this problem can be understood as the problem of a firm that could choose any other worker type at any other quantity. As will become clear in the following, the wages are twice differentiable,[15] and the first order conditions for optimality are

$$\begin{aligned} f_\theta(x, \mu(x), \theta(x)) - w(x) &= 0 && (5)\\ f_x(x, \mu\left(x\right), \theta(x)) - \theta(x) w'(x) &= 0, && (6) \end{aligned}$$

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

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

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

$$f_{\theta\theta}[f_{xx} - \theta w''(x)] - (f_{x\theta} - w'(x))^2 \geq 0. \tag{7}$$

We can differentiate (5) and (6) with respect to the worker type to get

$$\begin{aligned} f_{x\theta} - w'(x) &= -\mu'(x) f_{y\theta} - \theta'(x) f_{\theta\theta} && (8)\\ f_{xx} - \theta(x) w''(x) &= -\mu'(x) f_{xy} - \theta'(x)\left[f_{x\theta} - w'(x)\right]. && (9) \end{aligned}$$

In the following three lines we successively substitute (8), (9) and then (6) into optimality condition (7):

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

For strictly positive assortative matching ($\mu'(x) > 0$) it has to hold that the term in square brackets in the last line is negative, for strictly negative assortative matching the term in square brackets in the last line needs to be positive. Focussing on positive assortative matching, and using the relationship in (6), we obtain the condition:

$$f_{\theta\theta} f_{xy} - f_{y\theta} f_{x\theta} + f_{y\theta} f_x/\theta \leq 0. \tag{10}$$

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.[16] A standard implication

[15] Given the assumed differentiability of μ and θ, the wage has to be differentiable as defined in (8) and (9) below.

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

of constant returns it then $F_x(x, y, \theta, 1) = \theta F_{xl} + F_{xr}$. We can now rewrite (10) in terms of $F(x, y, \theta, 1)$ and rearrange to obtain the following cross-margin-complementarity condition:

$$\begin{aligned} & F_{ll}F_{xy} - F_{yl}\left[F_{xl} - F_x/\theta\right] \leq 0 && (11)\\ \Leftrightarrow \quad & F_{ll}F_{xy} + F_{yl}F_{xr}/\theta \leq 0 && \\ \Leftrightarrow \quad & F_{xy}F_{lr} \geq F_{yl}F_{xr}. && (12) \end{aligned}$$

The condition depends on the cross-partials in each dimension, relative to the cross-partials across the two dimensions. Only if the within-complementarities in the extensive and intensive dimension on the left hand side exceed the between-complementarities from extensive to intensive margin on the right hand side, does positive assortative matching arise. The following sums up this finding: A necessary condition to have equilibria with positive assortative matching is that (12) holds along the equilibrium path. The reverse inequality is necessary for negative assortative matching.

The preceding argument heavily relies on local variations to establish the necessity of inequality (12). In the following we show that one does not need any additional requirements to achieve assortative matching. Also, we prove the Theorem for arbitrary type distributions, including those that might not have a continuous density.[17]

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

$$F_{xy}F_{lr} \geq F_{yl}F_{xr} \tag{13}$$

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. The proof relies on the first welfare theorem. Since we have quasi-linear utility, any equilibrium maximizes the sum of outputs in the economy. A feasible distribution $\mathcal{R}$ generates market output

$$S(\mathcal{R}) = \int F(x, y, \theta, 1)d\mathcal{R}.$$

In the appendix we prove sufficiency by considering the case where $(x_i, y_i, \theta_i) \in$supp$\mathcal{R}$ for $i \in \{1, 2\}$ and $x_1 > x_2$ but $y_1 < y_2$. We establish that output is strictly increased under a feasible variation yielding resource allocation $\mathcal{R}'$ that pairs some of the x_1 workers to some of the y_2 resources. Therefore, $\mathcal{R}$ cannot be optimal and cannot be an equilibrium, implying that sorting must be positive assortative, which establishes sufficiency. Since we construct an improvement path, we require the condition to hold at all possible values (i.e., in $\mathbb{R}^4_+$). A similar argument establishes that if (13) fails, then any matchings within the type space where it fails have to be negative assortative, as otherwise re-arranging would improve output. So if (13) fails we can find some type distribution with types in the region where it fails such that we have negative sorting, which means that positive assortative matching cannot hold for all type distributions. This establishes necessity. An analogue argument establishes the results for negative assortative matching. ■

[17] Also note that under homogeneity of degree one the condition $F_{xy}(x, y, l, r)F_{lr}(x, y, l, r) \geq F_{yl}(x, y, l, r)F_{xr}(x, y, l, r)$ is equivalent to $F_{xy}(x, y, l/r, 1)F_{lr}(x, y, l/r, 1) \geq F_{yl}(x, y, l/r, 1)F_{xr}(x, y, l/r, 1)$ when $r > 0$ and $l > 0$, which means that less combinations have to be checked.

This condition 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, one within quantities and two across quality-quantity dimensions. Observe that, as in the one-to-one matching model (Becker 1973), both positive and negative assorted allocations constitute an equilibrium if the condition holds with equality. Hence, the condition is only sufficient when it holds strictly.

On the left-hand side, a large value of the cross-partial on the quality dimensions (F_{xy}) captures strong *type complementarity* and means that higher firm types have ceteris paribus a higher marginal return for matching with higher worker types. This is reinforced by a higher cross-partial on the quantity dimension, the *complementarity in quantities*, even though under constant returns to scale this is always positive and might be viewed as a normalization. The terms 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. Finally, the *managerial resource complementarity* F_{xr} expresses how the marginal product of managerial time varies across better workers. If managerial time is particularly productive when spent with high skilled types, then it is positive and large. This would be the case for example if the learning by high types is faster. Instead, if time is more productive with low types, it is negative.[18] The overall condition 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.[19]

The condition for positive assortative matching then compares the product of within-complementarities on the left with the product of across-complementarities on the right. The left captures the traditional type complementarities and in the absence of a quantity dimension the right 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. If there is a strong quantity-quality complementarity, the traditional type complementarity F_{xy} must be strong enough. The discussion in Section 3.3 reveals that as the elasticity of substitution on the quantity dimension goes to zero – in the limit there is no substitution and agents can only be matched into pairs – the right hand side goes to zero.

Finally, one may wonder what happens when our homogeneity assumption does not hold and output is not proportional to the ratio θ of the labor force l to the amount of resources r. Conceptually, the problem is identical to the one we solve here (see the Appendix for the derivation). While the interpretation is much less transparent, the main sorting condition (11) is still necessary for differential positive assortative matching under increasing returns to scale, only the steps that require homogeneity do not apply.

3.2 Equilibrium Assignment, Firm Size Distribution and Wage Profile

In contrast to models with pairwise matching where assortativeness immediately implies who matches with whom (the best with the best, the second best with the second best, and so forth), this is not obvious in this framework as particular firms may hire more or less workers in equilibrium. In the appendix we show how to

[18] 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.

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

construct a differentiable positive assorted equilibrium when (13) holds on the full domain. Our main focus is the characterization. For the following we will consider output functions that are increasing in types, which ensures that all types above some cutoff are matched. If output can fall for higher types, holding all other variables constant, than there might be holes in the matching set, and the following characterization can only be applied on each connected component. The results hold even if output can fall, as long as it is ensured that on the equilibrium path all agents above some cut-off trade. The next proposition fully characterizes the equilibrium.

Proposition 1 *If matching is assortative and differentiable and output is 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:*

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

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

where $\mathcal{H}(x) = h_w(x)/h_f(\mu(x))$*.*

Proof. Consider the case of PAM – the case of NAM can be derived in a similar way. The equilibrium condition for market clearing condition implies $H_w(\overline{x}) - H_w(x) = \int_{\mu(x)}^{\overline{y}} \theta(\tilde{x})h_f(\tilde{x})dx$. Differentiating with respect to x delivers the second differential equation in (14).

The initial condition in the case of PAM is $\mu(\overline{x}) = \overline{y}$. From (6) we know that $w' = f_x/\theta$, which gives the third equation since $F_x = f_x$. From the first-order condition in equation (5) we know that $f_\theta(x, \mu(x), \theta(x)) = w(x)$. Then from equation (8), after substituting for w' and μ' we obtain:

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

Using the same substitutions that we used in connection with equation (10) we obtain the first equation in (14). The initial condition for this differential equation obtains from running down the allocation from the top to the bottom and where the boundary condition holds either when the lowest type is attained or when the number of workers goes to zero. An equilibrium allocation simultaneously solves the differential equation for μ' and θ' with the respective boundary conditions. ■

An immediate implication of interest of these equilibrium conditions is that the size distribution $\theta(x)$ may change if we hold the production function and the distribution of firms type constant. This occurs when the distribution of workers changes. In particular, for some distributions of worker skills better firms will be smaller, while for other distributions better firms might be larger. This is important in the misallocation debate: firm distributions vary even without mismatch, once the skill distribution is taken into account (see discussion below).

Proposition 1 then immediately implies the following result on the size of the different firms:

Proposition 2 *If matching is assortative and differentiable, and output is increasing in types, better firms hire more workers if and only if along the equilibrium path:*

1. $\mathcal{H}(x)F_{yl} > F_{xr}$ *under PAM,*

2. $\mathcal{H}(x)F_{yl} > -F_{xr}$ *under NAM.*

Proof. This follows readily from Proposition 1, once one realizes that under NAM $\theta'(x) < 0$. ■

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

An important feature of the model is that from the size distribution, the allocation and the wages, we can ascertain the technological determinants of differences between industries. In the retail industry for example, high productivity firms such as Walmart have invested heavily in the ability of the management team to supervise many workers. This is achieved through information technology that tells the store manager exactly what is in stock, how much each cash register is taking in, and how each individual employee is performing. In local mom-and-pop stores technological sophistication is usually present to a lesser extent. Thus, the span-of-control is increasing in firm type, i.e., F_{yl} is positive and large. And since Walmart tends to employ lower ability workers, there is NAM. From condition (13) we therefore infer that the type complementarity F_{xy} is not too high relative to the span-of-control complementarity F_{yl}. Moreover, since Walmart is much larger than the local retail stores, the firm size is steeply increasing in firm type y. This allows us to infer that the 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 is positive assortative. From this we infer that the type complementarityF_{xy} must be large. While it seems natural that the best managers benefit more from having many team members in order to leverage their skills ($F_{yl} > 0$), it is also very beneficial to spend time with the very talented team that they assembled to transfer their knowledge ($F_{xr} > 0$). The type complementarity must be large to outweigh the product $F_{yl}F_{xr}$. Firm size changes according to $F_{yl} - F_{xr}$. The fact that top consultancy firms do not operate much larger groups than lower level ones indicates that the difference between these two complementarities is small.

Interestingly, if matching is PAM and $F_{yl} = F_{xr}$ exactly holds, 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.

Skill-biased versus Quantity-biased Technological Change. In our framework, it is natural to investigate the impact of different types of technological change, in particular the difference between skill-biased technological change and quantity-biased technological change. Skill-biased technological change can be thought of as an increase in F_{xy}. This changes the wages, but may affect the factor allocations and the size distribution only marginally. In particular, if the model is symmetric and in the region of positive assortative matching, firm size is constant. Firms benefit from better workers, but also better workers benefit from better firms. The wage differences between workers according (14) depend on F_x along the equilibrium assignment, and is larger for higher types when F_{xy} is larger because the fact that they are matched to better partners implies a larger marginal product.

In contrast, quantity-biased technological change increases F_{yl}. This directly affects the size distribution. Better firms not only have better managers, but they also invest in resources to supervise. If quantity-biased technological change manifests itself in information and communication systems that allow management to supervise larger groups (F_{yl} large), then it immediately implies that better firms are bigger. Such quantity-biased technological change is therefore the direct determinant of a change in firm size (or team size within firms) whereas the notion of skill-biased technological change is the direct determinant of changes in the wage premium. Typically of course, the model is not symmetric and general equilibrium effects will kick in, possibly off-setting some of these effects.

Mismatch Debate. Our model can help shed light on the mismatch debate. In recent work, Restuccia and Rogerson, 2008, and Hsieh and Klenow, 2010, amongst others, have argued that large GDP differences across countries can be accounted for in part by considering firm heterogeneity. In a Lucas span of control model, the firm size distribution then identifies cross-country differences that can in part be attributed to differences in the allocation of resources. If firms are less productive, they optimally hire fewer workers. Such firm heterogeneity explains a large part of the cross-country differences in factor productivity. Nonetheless, after accounting for heterogeneity in firm productivity and firm size, substantial differences remain in factor productivity across countries. Those remaining differences are attributed to taxation regimes or financial and other market frictions.

Models in this literature usually lack a role for heterogeneity in skills. The skill distribution in India is first order stochastically dominated by the distribution in the US. Surely, even if all firms had the same technologies, the equilibrium allocation of skilled workers to firms will differ. For one, because high skilled workers are more scarce in India than in the US which would be reflected in wages. This will affect both the skill composition across firms and the firm size distribution. In particular, whether more productive firms are larger depends in part on the relative availability of high skilled workers, as Proposition 2 makes clear. What is currently attributed to financial frictions might in large part be due to differences in the countries' skill distribution.

To give an idea of the technological differences, consider the following nested CES version of our general technology:

$$F(x,y,l,r) = r\left[\mu y^{\alpha} + (1-\mu)\left(\lambda_1 x^{\beta} + \lambda_2\left(\frac{l}{r}\right)^{\beta}\right)^{\frac{\alpha}{\beta}}\right]^{\gamma}. \tag{16}$$

Then a special case where $\beta = 0$ and $\lambda_1 = \lambda_2 = 1$ is a technology in efficiency units of labor: skills are heterogeneous but the distribution is indeterminate since skills are perfect substitutes:

$$F(x,y,l,r) = r\left[\mu y^{\alpha} + (1-\mu)\left(x\frac{l}{r}\right)^{\alpha}\right]^{\gamma}.$$

And when $\alpha = \frac{1}{\gamma} = 0$ and $r = 1$ then we get

$$F(x,y,l,r) = y^{\mu}\left(xl\right)^{1-\mu}$$

which is the Lucas span of control model in efficiency units xl. The original Lucas model had $xl = l$ for all x, and therefore had no role for worker heterogeneity at all. A more general form such as (16) nests these other models, but allows a role for both skill and quantity bias. Now the distribution of skills in conjunction with the distribution of firm productivities determines the equilibrium size distribution. Countries with relatively few high skill workers for example may adopt smaller firm sizes as an optimal response to the scarce endowment of skilled workers.

3.3 Special Cases

This section documents how our model characterizes a number of existing setups that have been heavily used in the literature. It also highlights that it can capture new settings that have not been analyzed before. It also shows that it is not easy to start with certain separability assumptions (for example between the quality and quantity dimensions as in the second example below), because a lot of formulations in the literature have used different ways of interacting the variables that can be captured in our setup but not in more specialized versions.

1. Efficiency units of labor. A particularly common assumption in the literature is the case of efficiency units of labor, where the output remains unchanged as long as the multiplicative term xl_x remains unchanged. In such a case workers of one type are completely replaceable by workers of half the skills as long as there are twice as many of them. Sorting is then essentially arbitrary: Each firm cares only about the right total amount of efficiency units, but not whether they are obtained by few high-type workers or many low-type workers. Our setup captures efficiency units of labor under the production function $F(x, y, l, r) = \tilde{F}(y, xl, r)$. Taking cross-partials immediately reveals that we always obtain $F_{xy}F_{lr} = F_{yl}F_{xr}$ in this case.

2. Multiplicative separability. A particularly tractable case arises under multiplicative separability of the form $F(x, y, l, r) = A(x, y)B(l, r)$. In this case the condition (13) for positive assortative matching can be written as $[AA_{xy}/(A_xA_y)][BB_{lr}/(B_lB_r)] \geq 1$. If B has constant elasticity of substitution ε, we obtain an even simpler condition $AA_{xy}/(A_xA_y) \geq \varepsilon$.[20]

3. Becker's one-on-one matching model as a limit case. Consider some output process $F(x, y, l, r)$. In the spirit of most of the sorting literature, we can now consider the restricted variant where only "paired" inputs can operate: every worker needs exactly one unit of resource and any resource needs exactly one worker, otherwise it is not used in production. The output can then be represented by $F(x, y, \min\{l, r\}, \min\{r, l\}) = F(x, y, 1, 1)\min\{l, r\}$, where the equality follows from constant returns to scale. This nicely corresponds to the multiplicatively separable setup discussed in the previous point. While our framework is build around the idea that more resources or more labor inputs improve production, this Leontief setup is on the quantity dimension exactly the limit case of a CES function with zero elasticity ($\varepsilon \to 0$). From the previous point we therefore know that sorting arises in this limit if $F_{xy} \geq 0$, which is exactly the condition in Becker (1973).

4. Sattinger's and Garicano's span of control problem as a limit cases. One of the few contributions that provides clear conditions for sorting in a many-to-one matching model is presented in Sattinger (1975). His production function assumes that each worker produces the same, but a worker of type x needs $t(x, y)$ units of supervision time from manager of type y, where better types need to spend less time. The manager can only hire as many workers as he can supervise, so that $F(x, y, l, r) = \min\{r/t(x, y), l\}$, where the first terms in the minimization operator captures the number of workers that can be supervised and the second the number of workers hired. Our model allows for more flexibility in the substitution between inputs, but a CES extension that takes $r/t(x, y)$ and l as inputs again has the previous Leontief specification as the inelastic limit.[21] Inspecting (13) and taking the inelastic limit reveals that positive sorting arises only if $t(x, y)$ is log-supermodular. This exactly recovers the condition found by Sattinger.

Related is Garicano's (2000) formulation where each worker can solve problems with difficulty equal to his type, and passes the remaining problems to the supervisor who needs one unit of time to consider each problem passed to him. Here the supervision "time" $t\,(x)$ depends only on the worker type, but each manager can solve problems up to category y himself, leading to output $F = y\min\{r/t(x), l\}$. Approximating this with the appropriate CES highlights that better managers prefer to hire better workers to leverage their skills.

[20] If jobs match with workers one-to-one according to a matching function that separates markets as in Jerez (2012), Eeckhout and Kircher (2010) and Shi (2001), the resulting output per market has such a multiplicative structure and can be analyzed as a "firm" in our setting. If ε is in the unit interval, this condition is equivalent to root-supermodularity, i.e., it is equivalent to $\sqrt[n]{A(x, y)}$ being supermodular with $n = (1 - \varepsilon)^{-1}$ as shown by Eeckhout and Kircher (2010). If $\varepsilon > 1$ this requires conditions on $A(x, y)$ that are stronger than log-supermodularity.

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

5. Extension of Lucas' Span of Control, and Rosen's general production: Lucas (1978) assumed a production function that is multiplicatively separable in the firm type and the amount of labor, where all labor is identical. Consider the following extension to heterogeneous labor: $rF(x,y,\theta,1) = ryg(x,\theta)$, indicating that the return firm productivity is leveraged both by quality and quantity of workers. The new condition for assortative matching is $g_\theta g_{x\theta} \geq g_x g_{\theta\theta}$. If production is increasing in worker type and strictly concave, this means that sorting will be positive unless better workers types indeed dislike to work together because that limits the amount of resources they can obtain ($g_{x\theta}$ sufficiently negative).

This is related to Rosen's (1982) setup where $F = rh(y)g(x, y\theta)$ for the first level of supervision and the production workers, which has somewhat different separability assumptions.[22] He allows more flexibility by allowing for multiple layers of hierarchy and a choice on who performs on which layer. But he analyzes the model only for the case of linear homogeneity of g, which is equivalent to efficiency units of labor analyzed in point 1 above. Since his general model is a special case of our setup, our sorting conditions apply directly to this setting. Again, one can easily write the conditions in terms of g: $g_\theta g_{x\theta} - g_x g_{\theta\theta} \geq g_x g_\theta / \theta$, but does not get that much additional insight above and beyond those we have discussed already for the general model.

6. Spatial Sorting Within the Mono-centric City. The canonical model of the mono-centric city can explain how citizens locate across different locations, however there is no spatial sorting. All agents are identical and in equilibrium they are indifferent between living in the center or in the periphery by trading off commuting time for housing space and prices.[23] We therefore consider a model of spatial sorting within the city. Let there be a continuum of locations y, each with housing stock $r(y)$. Let $y \in [0,1]$, where y is the center and y is the inverse of a measure of the distance from the center. Agents with budget x have preferences over consumption c and housing h represented by a quasi-linear utility function $u(c,h) = c + v(h)$. With consumption as the numeraire good and $p_h(y)$ as the price per unit of housing in location y, the budget constraint is $c + p_h(y)h = xg(y)$, where x is the worker skill and $g(y)$ is an increasing function representing the time at work rather than in commute. The closer to the center, the less time is spent on commuting and the more time is earned. Then we can write the individual citizen x's optimization problem as $xg(y) + v(h) - p_h(y)h$. The total supply of housing in location y is r and as a result, $l \cdot h = r$. 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 $F_{xy} = g'(y)l$, $F_{lr} = -\frac{r}{l^2}v''\left(\frac{r}{l}\right)$, $F_{xr} = 0$ so that if $v(\cdot)$ is concave there is positive assortative matching of the high income earners into the center and the low income earners in the periphery. A similar functional form is used in Van Nieuwerburgh and Weill (2010) to consider differences between cities rather than within the city, where the term $xg(y)$ is replaced by a more agnostic worker-output $u(x,y)$ depending on worker skill x and city type y. Sorting is again fully determined by the cross-partial of x and y because $F_{xr} = 0$.

4 Extensions

Our baseline model set up is very general. So far, we have given it the interpretation of a managerial assignment problem that optimizes both the worker quality and the firm's span of control. The advantage of the generality of the setup is that we can readily interpret the basic model in different settings and extend it with minor modifications. Our principal extension is the introduction of unemployment. This is then followed by several other interpretations.

[22] Rosen (1982) equation (1) for the output per worker can be written as $h(y)\xi(yr/l,x)$ for some functions h and ξ,so that total output is constant returns to scale. Output per resource is therefore $yg(yl/r,x)$ after appropriate transformation (so that $g(y\theta,x) := \xi(y/\theta,x)/\theta$).

[23] Also Lucas and Rossi-Hansberg (2002) model the location of identical citizens but their model incorporates 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.

4.1 Frictions and Involuntary Unemployment

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

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

Consider a situation where the workers are unemployed and can only be hired by firms via a frictional hiring process. As part of this process, each firm decides how many vancancies v_x to post for each worker type x that it wants to hire. Posting v_x vacancies has a linear cost cv_x. It also decides to post wage ω_x for this worker type. Observing all vacancy postings, workers decide where to search for a job. Let q_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 $m(q_x)$, which is assumed to be strictly increasing and strictly concave. Since there are q_x workers queueing per vacancy, the workers' job-finding rate for these workers is $m(q_x)/q_x$. The job finding rate is assumed to be strictly decreasing in the number of workers q_x queueing per vacancy. Firms can attract workers to their vacancies as long as these workers get in expectation their equilibrium utility, meaning that q_x adjusts depending on ω_x to satisfy: $\omega_x m(q_x)/q_x = w(x)$. Note the difference between the wage ω_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 (1) the new problem

$$\max_{r_x,\omega_x,v_x} \int [F(x,y,l_x,r_x) - l_x\omega_x - v_x c]\, dx \qquad (17)$$
$$\text{s.t. } l_x = v_x m(q_x); \quad \text{and} \quad \omega_x m(q_x)/q_x = w(x)$$

and r_x integrates to unity. The first line simply takes into account that the firm has to pay the vacancy-creation cost, and that the number of hires depends on the amount of hiring per vacancy which is in turn related to the wage that it offers. There are two equivalent representations of this problem that substantially simplify the analysis. It can easily be verified that problem (17) is mathematically equivalent to both of the following two-step problems:

1. Let $G(x,y,s,r) = \max_v [F(x,y,vm(s/v),r) - vc]$, and solve $\max_{s_x,r_x} \int [G(x,y,s_x,r_x) - w(x)s_x]dx$ where r_x integrates to unity.

2. Let $C(l,x) = \min_{v,q}[cv + vqw(x)]$ s.t. $l = vm(q)$, and solve $\max_{l_x,r_x} \int [F(x,y,l_x,r_x) - C(l_x,x)]dx$ where r_x integrates to unity.

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

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

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

$$w(x)q(x) = \frac{\eta(q(x))}{1-\eta(q(x))}c \tag{18}$$

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

Proposition 3 *Assume higher worker types create more output ($F_x > 0$). In the competitive search equilibrium with large firms, higher skilled workers have lower unemployment rates.*

Proof. The term $\eta(q)/[q(1-\eta(q))] = m'(q)/[m(q) - qm'(q)]$. This term is strictly decreasing in q, since the numerator is strictly decreasing and the denominator is strictly increasing in q. Since output at any firm is increasing in skill ($F_x > 0$) it follows immediately that in any equilibrium $w(x)$ is increasing in x. Implicit differentiation of (18) implies that $q(x)$ is decreasing, which in turn implies that the chances of finding employment are increasing in x. ■

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

Interestingly, the finding in Proposition 3 implies that under positive assortative matching the firm-size can be increasing in firm type even though the number of workers that apply for jobs is decreasing. This can be seen mathematically as follows. The amount of labor that is actually hired, $l(x)$, relates to the actual number of searchers and their queue per vacancy as $l(x) = s(x)m(q(x))/q(x)$, implying:

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

The change in the number of searchers (s') is determined by (14) under appropriate change of variables (θ and f replaced by s and g). Even if the number of workers that search for employment at better firms is not increasing, the number of hires might still be increasing because the second term is strictly positive. This is due to the fact that high productivity firms put more resources into creating jobs for their high-skilled applicants. If a firm tries to attract workers for whom their time-constraints make it very costly to apply, it will invest resources to make sure that the applicants perceive a sufficiently high probability that they will find a suitable appointment in the hiring process. In this model this is captured through creating a sufficient number of different vacancies.

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

Proposition 4 *The vacancy rate is ambiguous in firm size.*

Proof. Consider PAM (likewise for NAM). The vacancy rate ($1/q$) is increasing in x, and under PAM then also in y. However, from Proposition 2, firm size ambiguous in y. In particular, it is increasing if $G_{yl} \geq G_{xr}$ and decreasing if $-G_{yl} \leq G_{xr}$. ■

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

4.2 Capital Investment

Consider a production process that not only takes as inputs the amount of labor and of proprietary firm resources, but also some amount k of a generic capital good, and creates output $\hat{F}(x, y, l, r, k)$. The generic capital can be bought on the world market at price i per unit. Optimal use of resources requires $F(x, y, l, r) = \max_k \left[\hat{F}(x, y, l, r, k) - ik\right]$, where F is constant returns in its last two arguments if $\hat{F}$ is constant returns in its last three arguments. Rewriting the cross-margin-complementarity condition (13) in terms of the new primitive yields the following condition for positive assortative matching: $\hat{F}_{xy}\hat{F}_{lr}\hat{F}_{kk} - \hat{F}_{xy}\hat{F}_{lk}\hat{F}_{rk} - \hat{F}_{xk}\hat{F}_{yk}\hat{F}_{lr} \geq \hat{F}_{xr}\hat{F}_{yl}\hat{F}_{kk} - \hat{F}_{xr}\hat{F}_{yk}\hat{F}_{lk} - \hat{F}_{xk}\hat{F}_{yl}\hat{F}_{rk}$. We expect that particular functional form assumptions for the way that generic capital affects the production process will simplify this condition and make it more amenable for interpretation in specific cases.

4.3 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 (1977), 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.[24] The difficulty in this setup is that, despite the fact that output is constant returns to scale in employment and

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

firm resources, the revenue of the firm has decreasing returns to scale. Therefore, we cannot directly apply (13). 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 (10) directly. Rearranging and using $\tilde{F}(x,y,l,r) = r\tilde{f}(x,y,l/r)$ we get the condition for positive assortative matching

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

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 (13). In gereral, the condition relies not only on supermodularities in the production function, but also on log-supermodularities. This should not be surprising. Even in the standard models supermodularity is the relevant condition when the marginal consumption value of output is normalized to one (Becker 1973), while sorting when output is CES-aggregated requires log-supermodularity. If $\tilde{F}$ is multiplicatively separable between quantity and quality dimension, and the quality dimension is CES, then as the quality dimension becomes increasingly inelastic it is easy to show that the condition reduces to log-supermodularity in x and y.

4.4 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 $F(x,y,l,r) = -rc(x,y) - \alpha r d(r/l)$, 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.

4.5 Endogenous type distributions, technology choice, and 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 the profits of firms according to (2) equal the entry cost $c(y)$ for all active firms. In fact, this formulation is easier to construct: We know that the highest types match, so that $\mu(\bar{x}) = \bar{y}$. The problem is usually how to determine at which ratio they match, i.e., to find $\theta(\bar{x})$. But here it is given simply by the requirement that the profits of the highest firm equals the entry costs. Substituting the first order condition (5) 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.

5 Concluding Remarks

We have proposed a matching model 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 trade off of complementarities between and across types and quantities. The equilibrium allocation is completely characterized by the solution to the system of two differential equations that pins down the allocation, the firm size distribution and the wage distribution.

Observed allocations of workers to firms and the firm size distribution can inform us about the underlying technological characteristics. In the retail sector, for example, large stores like a Walmart center with an unskilled workforce coexist with small corner stores. We argue that this can be explained by a strong span of control complementarity. More productive management is disproportionately more productive in managing larger groups of workers. We also illustrate how technological change leads to different conclusions whether it is skill-biased or quantity-biased. Skill-biased change directly affects the skill premium and therefore the wage distribution whereas quantity-biased change directly affects the firm size distribution.

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

6 Appendix

Remainder Proof of Proposition 1

Proof. Part I: sufficiency. Focus on positive assortative matching. The same logic applies to negative assortative matching. Strict cross-margin-supermodularity $F_{xy}F_{lr} > F_{xr}F_{yl}$ for all (x, y, l, r) is by (10) equivalent to $f_{\theta\theta}f_{xy} - f_{y\theta}f_{x\theta} + f_{y\theta}f_x/\theta < 0$ for all (x, y, θ). Assume a feasible resource allocation $\mathcal{R}$ such that (x_i, y_i, θ_i) $\in$supp$\mathcal{R}$ for $i \in \{1, 2\}$ and $x_1 > x_2$ but $y_1 < y_2$. Focus on strictly positive intensity $\theta_i > 0$. This is either already the case, or otherwise the facts that (x_i, y_i, θ_i) $\in$supp$\mathcal{R}$ and $\mathcal{R}(x, y, 0) = 0$ guarantee that there is a closeby combination with strictly positive intensity on which we can focus. Let r_i denote the measure of resources at combination (x_i, y_i, θ_i). If the type distribution has mass-points, this is exactly the measure of types assigned to this combination. We will consider this case here. (If the type distribution only has a density, then it is the mass of resources in an arbitrarily small area around (x_i, y_i, θ_i). By continuity all resources in this area have output very close to $f(x_i, y_i, \theta_i)$, and since the inequalities below are strict, the argument applies also to this case.)

We will establish that output is strictly increased under a feasible variation yielding resource allocation $\mathcal{R}'$ that pairs some of the x_2 workers to some of the y_2 resources. 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 r resources are deployed in this match (and are paired to θr workers). For the variational argument, we are interested in the marginal benefit of pairing an additional measure r' of resources of type y' with workers of type x. The optimal output is generated by withdrawing some optimal measure $\theta' r'$ of the workers that were supposed to be working with resource y and reassigning them to work with resource y'. The joint output at (x, y) and (x, y') is given by

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

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

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

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

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

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

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

2. Not PAM has strictly positive marginal benefits from matching the high types:

We started under the assumption that matching is not assortative since $x_1 > x_2$ but $y_1 < y_2$. In particular, consider y_1 matched to x_2 with θ_1 and y_2 matched to x_1 with θ_2, where $x_2 > x_1$ and $y_2 > y_1$. For (x_1, y_2) and (y_1, x_2) to be matched, optimality requires that the marginal benefit of types $y^v = y_1$ are higher when paired with x_2, while types $y^v = y_2$ yield higher benefit when paired with x_1 :

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

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

where $\beta(\cdot; \cdot, \cdot, \cdot)$ was defined in (19). We will show that if (22) holds, then (23) cannot hold, which yields the desired contradiction. We will show this by proving that the benefit $\beta(y'; x_1, y_1, \theta_1)$ on the right hand side of (22) and (23) always remains above the benefit $\beta(y'; x_2, y_2, \theta_2)$ on the left hand side. By (22) this has to be true at $y' = y_1$, and we will show that it remains true when we move to higher y'. The marginal increase of β with respect to its first argument y' is given by

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

where θ' is again determined as in (21). Assume there is some $y' \geq y_1$ such that marginal benefits are equalized, i.e., $\beta(y'; x_2, y_2, \theta_2) = \beta(y'; x_1, y_1, \theta_1)$. We have established the result when we can show that $\beta_1(y'; x_2, y_2, \theta_2) < \beta_1(y'; x_1, y_1, \theta_1)$.

By (24) this equivalent to showing that $f(x_2, y', \theta_2') < f(x_1, y', \theta_1')$, where $\theta_1' = \theta'(y'; x_1, y_2, \lambda_2)$ and $\theta_2' = \theta'(y'; x_2, y_1, \lambda_1)$ as in (21). To show this, define $\xi(x)$ for all x in resemblance of (20) by the following equality

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

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

Part II: necessity. Assume that (13) 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 $\theta' = l'/r'$ (see also Footnote 17). 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 (13) cannot fail if we want to obtain positive assortative matching. Since we want to ensure positive assortative matching for all type distributions, we can choose the support of x and y within this neighborhood. But since θ is endogenous, this requires slightly more work. Assume that $X = [x', x' + \varepsilon]$ and $Y = [y', y' + \varepsilon]$, and uniform type distributions with mass $H_w^\varepsilon(x' + \varepsilon) = \theta'$ and $H_f^\varepsilon(y' + \varepsilon) = 1$. For small enough ε', firms make nearly identical profits. Since they can only match with nearly identical types, identical profits require them to have nearly identical factor ratios $\theta(x)$. These have to be close to the average ratio in the population. Therefore, for ε small enough all matches lie in $\mathcal{N}$, which rules out that matching can be positive assortative for all type distributions if (13) fails. ■

Construction of a differential positive assortative equilibrium.

Assume (13) holds at all $(x, y, l, r) \in \mathbb{R}_+^4$. Assume also that $F(x, y, \theta, 1)$ is strictly increasing in x and y and has a finite strictly positive maximum in θ (for example because of positive outside options as in outlined in Footnote 11). The former implies that higher types are matched in equilibrium whenever lower types are. We know there is sorting, so $\mu(\bar{x}) = \bar{y}$. Take a guess for $\theta(\bar{x})$. Then the first two equations in (14) evaluated at $(x, \mu(x), \theta(x))$ give a differential equation system that uniquely gives $\mu(x)$ and $\theta(x)$ at all values of x below $\bar{x}$. Equation (5) gives the associated wages $w(x)$, and output minus the wages gives the firms' profits $\pi(\mu(x))$. Stop the differential equation at $x^\star$ and $y^\star = \mu(x^\star)$ when for the first time one of the following conditions occurs: (a) $x^\star = \underline{x}$, (b) $y^\star = \underline{y}$, (c) $w(x^\star) = 0$ or (d) $\pi(y^\star) = 0$. One has found an equilibrium if one of the following holds: (a)&(b) which means that the lowest types are matched and might both get positive payoff; (a)&(d) which means that not all firms are matched, with the lowest firms remaining unmatched and therefore earning zero; or (b)&(c) in which case some workers are unmatched and the lowest worker type gets zero; or (c)&(d) hold in which case there are both unmatched workers and firms at the lower end. Clearly, the first order conditions are satisfied because the differential equation was constructed to satisfy (5) and (6), and the second order condition is locall satisfied because (13) holds. One can show that (13) also implies that no firm wants to deviate globally. Finally, for any guess of $\theta(\bar{x})$ such that $w(\bar{x}) > 0$ and $\pi(\overline{y}) > 0$ one of the end-point conditions (a)–(d) arises. At very low guesses either (b) or (d) holds (few workers per firm means that one either exhausts the firms or depletes their profits), while at very high levels of $\theta(\bar{x})$ either (a) or (c) holds. Since the system changes continuously in the initial guess, the set of guesses that gives rise to a particular condition is compact. Given that at low guesses (b) or (d) holds while at high guesses (a) or (c) hold, there has to be some intermediate guess where two conditions hold at the same time: (a)&(b), (a)&(d), (b)&(c) or (c)&(d), constituting an equilibrium.

The Non-Homogeneous Production Technology

Let output of the firm be $F(x, y, r, s)$, and the firm of type y chooses the worker type and the labor intensity l. As before, let the capital intensity r be given. Then the problem of a firm that chooses exactly one type x is

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

The first order conditions for optimality are

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

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

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

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

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

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

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

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

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

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

References

[1] Antràs, Pol, Luis Garicano, and Esteban Rossi-Hansberg, “Offshoring in a Knowledge Economy,” *Quarterly Journal of Economics*, 2006, **1**, 31-77.

[2] Acemoglu, Daron, and Robert Shimer, “Efficient Unemployment Insurance,” *Journal of Political Economy* **107**, 1999a, 893-928.

[3] Atakan, Alp, “Assortative Matching with Explicit Search Costs”, *Econometrica* **74**, 2006, 667–680.

[4] Becker, Gary S., “A Theory of Marriage: Part I”, *Journal of Political Economy* **81(4)**, 1973, 813-46.

[5] Burdett, Kenneth, and Melvyn Coles, “Marriage and Class,” *Quarterly Journal of Economics* **112**, 1997, 141-168.

[6] Costinot, Arnaud, “An Elementary Theory of Comparative Advantage,” *Econometrica* **77(4)**, 2009, 1165-1192.

[7] Dixit, A. K. and J. E. Stiglitz, “Monopolistic competition and optimum product diversity,” *American Economic Review* **67**(3)*,* 1977, 297-308.

[8] Eeckhout, Jan and Philipp Kircher, “Sorting and Decentralized Price Competition”, *Econometrica* **78(2)**, 2010, 539-574.

[9] Eeckhout, Jan and Roberto Pinheiro, “Diverse Organizations and the Competition for Talent”, University of Pennsylvania mimeo, 2008.

[10] Fox, Jeremy, “Estimating the Employer Switching Costs and Wage Responses of Forward-Looking Engineers,” *Journal of Labor Economics* **28**(2), 2010, 357-412.

[11] Gabaix, Xavier, and Augustin Landier “Why has CEO Pay Increased so Much?”, *Quarterly Journal of Economics* **123(1)**, 2008, 49-100.

[12] Garibaldi, Pietro, and Espen Moen, “Job-to-Job Movements in a Simple Search Model,” *American Economic Review* Papers and Proceedings **100(2)**, 2010, 343-47.

[13] Garicano, Luis, “Hierarchies and the Organization of Knowledge in Production,” *Journal of Political Economy* **108 (5)**, 2000, 874-904.

[14] Garicano, Luis and Thomas Hubbard, ”The Returns to Knowledge Hierarchies”, 2008, mimeo.

[15] Garicano, Luis and Esteban Rossi-Hansberg. ”Organization and Inequality in a Knowledge Economy.” Quarterly Journal of Economics **121**(4), 2006, 1383-1436.

[16] Grossman, Gene, “The Distribution of Talent and the Pattern and Consequences of International Trade,” *Journal of Political Economy* **112**, 2004, 209-239.

[17] Grossman, Gene, and Giovanni Maggi “Diversity and Trade,” *American Economic Review* **90**, 2000, 1255- 1275.

[18] Grossman, Gene, Ehanan Helpman, and Philpp Kircher “Dance with the One Who Brought You - Matching and Sorting in a Global Economy,” 2012, mimeo.

[19] Guerrieri, Veronica, Robert Shimer and Randall Wright, “Adverse Selection in Competitive Search Equilibrium,” *Econometrica* **78(6)**, 2010, 1823-1862.

[20] Gul, Faruk, and Ennio Stacchetti, “Walrasian Equilibrium with Gross Substitutes,” *Journal of Economic Theory* **87**, 1999, 95-124.

[21] Hawkins, William, “Competitive Search, Efficiency, and Multi-worker Firms,” University of Rochester mimeo, 2011.

[22] Heckman, James J., and Bo E. Honore, "The Empirical Content of the Roy Model," *Econometrica* **58(5)**, 1990, 1121-1149.

[23] Helpman, Elhanan, Oleg Itskhoki, and Stephen Redding, "Trade and Labor Market Outcomes," CEPR Discussion Papers 8191, C.E.P.R. Discussion Papers, 2011.

[24] Hsieh, Chang-Tai and Peter J. Klenow, "Development Accounting," *American Economic Journal: Macroeconomics* **2(1)**, 2010, 207-23.

[25] Hopenhayn, Hugo, and Richard Rogerson, "Job Turnover and Policy Evaluation: A General Equilibrium Analysis," *Journal of Political Economy* **101(5)**, 1993, 915-938.

[26] Jerez, Belen, "Competitive Equilibrium with Search Frictions: a General Equilibrium Approach", mimeo 2012.

[27] Jovanovic, Boyan, "Selection and the Evolution of Industry," *Econometrica* **50(3)**, 1982, 649-670

[28] Kaas, Leo and Philipp Kircher, "Efficient Firm Dynamics in a Frictional Labor Market", mimeo, 2011.

[29] Kantorovich, L.V., "On the Translocation of Masses", (Translated into English and reprinted in Management Science **5(1)**, 1958, 1-4), *Comptes Rendus (Doklady) de l'Academie des Sciences de L'URSS* **37**, 1942, 199-201.

[30] Kelso, Alexander S., and Vincent P. Crawford, "Job Matching, Coalition Formation, and Gross Substitutes," *Econometrica* **50(6)**, 1982, 1483-1504.

[31] Koopmans, Tjalling C., and Martin Beckmann, "Assignment Problems and the Location of Economic Activities," *Econometrica* **25(1)**, 1957, 53-76.

[32] Krusell, Per, Lee E. Ohanian, Jose-Victor Rios-Rull and Giovanni L. Violante, "Capital-Skill Complementarity and Inequality: A Macroeconomic Analysis,"*Econometrica* **68(5)**, 2000, 1029-1053.

[33] Lentz, R. and Mortensen, D. T., "An Empirical Model of Growth Through Product Innovation." *Econometrica,* **76**, 2005 1317-1373.

[34] Lucas, Robert E., "On the Size Distribution of Business Firms," *The Bell Journal of Economics* **9(2)**, 1978, 508-523.

[35] Menzio, Guido, and Espen Moen, "Worker Replacement," *Journal of Monetary Economics*, **57**, 2010, 623-636.

[36] Milgrom, Paul R., and John W. Hatfield, "Matching with Contracts," *American Economic Review* **95(4)**, 2005, 913-935.

[37] Peters, Michael, "Ex Ante Pricing in Matching Games: Non Steady States", *Econometrica* **59(5)**, 1991, 1425-1454.

[38] Restuccia, Diego, and Richard, Rogerson, "Policy distortions and aggregate productivity with heterogeneous establishments," *Review of Economic Dynamics* **11**, 2008, 707-720.

[39] Rosen, S., "Authority, Control, and the Distribution of Earnings," *Bell Journal* **13(2)***,* 1982, 311-322.

[40] Roy, A. D. "Some Thoughts on the Distribution of Earnings," *Oxford Economic Papers* **3(2)**, 1951, 135-146.

[41] Roys, Nicolas, and Ananth Seshadri, "Economic Development and the Organization of Production," *mimeo.*

[42] sattinger, Michael, "Differential rents and the distribution of earnings," *Oxford Economic Papers* **31**(1), 1979, 60–71.

[43] sattinger, Michael, "Comparative Advantage in Individuals," *The Review of Economics and Statistics* **60**(2), 1978, 259-267

[44] Sattinger, Michael, "Comparative Advantage and the Distributions of Earnings and Abilities", *Econometrica* **43(3)**, 1975, 455–468.

[45] Shapley, Lloyd S. and Martin Shubik, "The Assignment Game I: The Core", *International Journal of Game Theory* **1(1)**, 1971, 111-130.

[46] Shi, Shouyong, "Frictional Assignment. 1. Efficiency," *Journal of Economic Theory*, **98**, 2001, 232-260.

[47] Shimer, Robert, "The Assignment of Workers to Jobs in an Economy with Coordination Frictions," *Journal of Political Economy* **113(5)**, 2005, 996-1025.

[48] Shimer, Robert, and Lones Smith, "Assortative Matching and Search," *Econometrica* **68**, 2000, 343–369.

[49] Smith, Eric, "Search, Concave Production and Optimal Firm Size," *Review of Economic Dynamics*, **2**, 1999, 456-471.

[50] Terviö, Marko, "The Difference that CEOs Make: An Assignment Model Approach," *American Economic Review* **98(3)**, 2008, 642-668.

[51] Van Nieuwerburgh, Stijn, and Pierre-Olivier Weill, "Why Has House Price Dispersion Gone Up?" *Review of Economic Studies*, **77**, 2010, 1567-1606.