A Static and Microfounded Theory of Zipf's Law for Firms and of the Top Labor Income Distribution*

François Geerolf †
University of California, Los Angeles

This version: December 1, 2014
First version: April 2014 **[Last Version]**

Abstract

I propose a theory of Zipf's Law for firm sizes and Pareto distributions for top labor incomes based on arbitrarily small differences in skills between workers, and no functional form assumption. A version of Garicano (2000)'s knowledge-based production hierarchies model is shown to generate Pareto distributions for high span of control of intermediary managers with coefficients equal to $2^L/(2^L-1)$ between L levels of management, under very limited assumptions on the underlying distribution of agents' skills. This breakdown of the aggregate firm size distribution receives considerable empirical support in the French matched employer-employee administrative data. The model provides a method to structurally estimate the relative importance of different factors in the recent rise in labor income inequality: the tail index of the Pareto top labor income distribution depends on a notion of race between education and technology, while the scale parameter depends on communication costs and heterogeneity. It also has striking implications for the literature in trade and macroeconomics on firm heterogeneity: in the model, heterogeneity in firm sizes grows when underlying productivity heterogeneity decreases.

Keywords: Pareto distributions, Wage Inequality, Hierarchies
JEL classification: D31, D33, D2, J31, L2

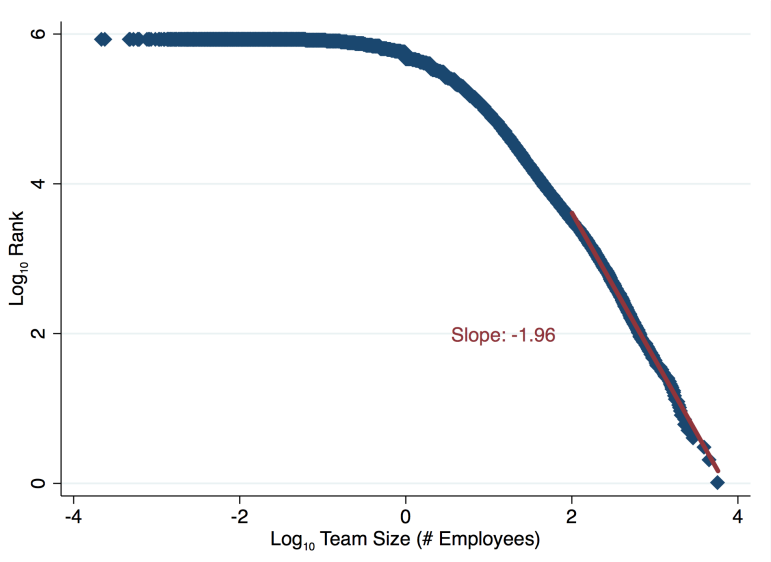
*I thank Daron Acemoğlu, Pol Antràs, David Atkin, Adrien Auclert, Simon Board, Ariel Burstein, Harold Cole, Arnaud Costinot, Jacques Crémer, Jan Eeckhout, Emmanuel Farhi, Xavier Gabaix, Cécile Gaubert, Christian Hellwig, Adriana Lleras-Muney, Martí Mestieri, Matthew Rognlie, François de Soyres, Ludwig Straub, Robert Ulbricht, Boris Vallée, Till von Wachter, Pierre-Olivier Weill, Iván Werning and numerous seminar participants at MIT, UC Berkeley, UCLA, UPenn and TSE for useful comments. I thank Toulouse School of Economics, where this project was started, for their hospitality; and I am particularly indebted to Maxime Liegey for presenting a graph on French team and plant sizes during a student workshop there. This work is supported by a public grant overseen by the French Research Agency (ANR) as part of the "Investissements d'Avenir" program (reference: ANR-10-EQPX-17 - Centre d'accès sécurisé aux données - CASD).

†E-mail: fgeerolf@econ.ucla.edu. This is preliminary, all comments are welcomed.

Introduction

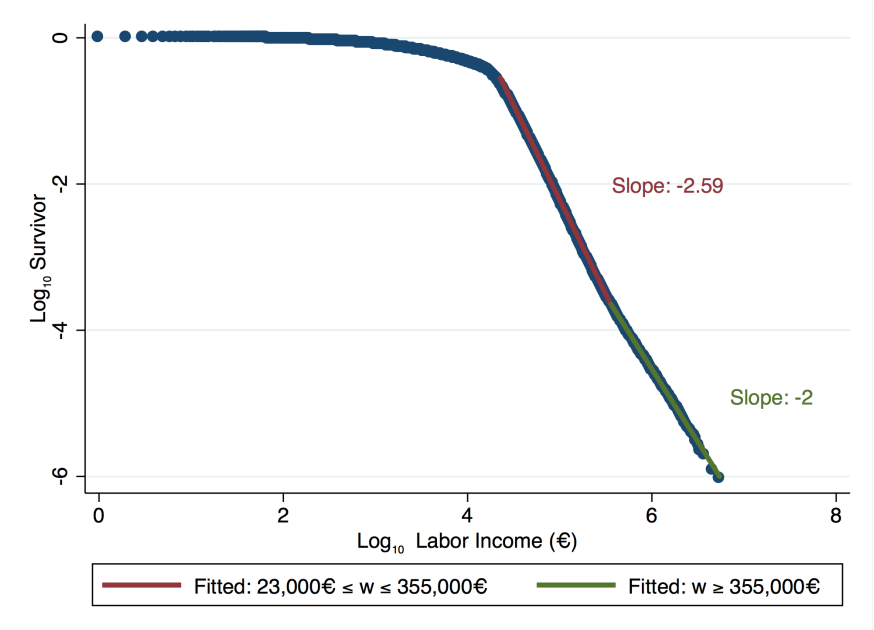
Economists usually agree that the distribution of firm sizes approximately follows Zipf's law in the upper tail. To some, it is even one of the few quantitative "laws" in economics, holding across time and countries (Gabaix (2014)). Existing models explaining this Pareto distribution all at some level rely on a stochastic proportional random growth process with statistical frictions, following Gibrat (1931)'s law, or on some other Pareto distribution underlying the economy's primitives, whether it be entrepreneurial skills or firms' productivities. In this paper, I propose a novel static mechanism, able to microfound the existence of Zipf's law, based on a production hierarchies model a la Garicano (2000). This model makes no functional form on the underlying distribution of primitives, in this model agents' skills. It rationalizes Zipf's law for firms by decomposing firms' hierarchies into elementary Pareto distributions for spans of control with a coefficient equal to 2 down one level of hierarchical organization. Hierarchies with two levels of management then follow Pareto distributions with a coefficient $4/3 \approx 1.33$, those with three levels have a coefficient of $8/7 \approx 1.17$. The Pareto coefficient converges to 1 (Zipf's law), as the number of levels increases. This decomposition receives considerable empirical support in the French administrative data, as Figure 1 exemplifies: in France, span of control down one level of management is indeed Pareto distributed with a coefficient close to two for large values of spans of control. The model can also account for the fact that Pareto holds only for the upper tail, and that very big firms are smaller than Zipf's law would predict. The theory rationalizes why establishment sizes are Pareto distributed, albeit with higher tail coefficients, in France and in the United States.

Figure 1: Span of Control down one level of Hierarchical Organization



Note: Source: French matched employer-employee Data, 2007. The construction of hierarchical levels follows Caliendo et al. (2014) and is based on the first digit of their occupation code. The Pareto fit is for span of controls higher than 100 employees ($\log_{10}(size) > 2$). See Section 2.7 for details and more supporting evidence.

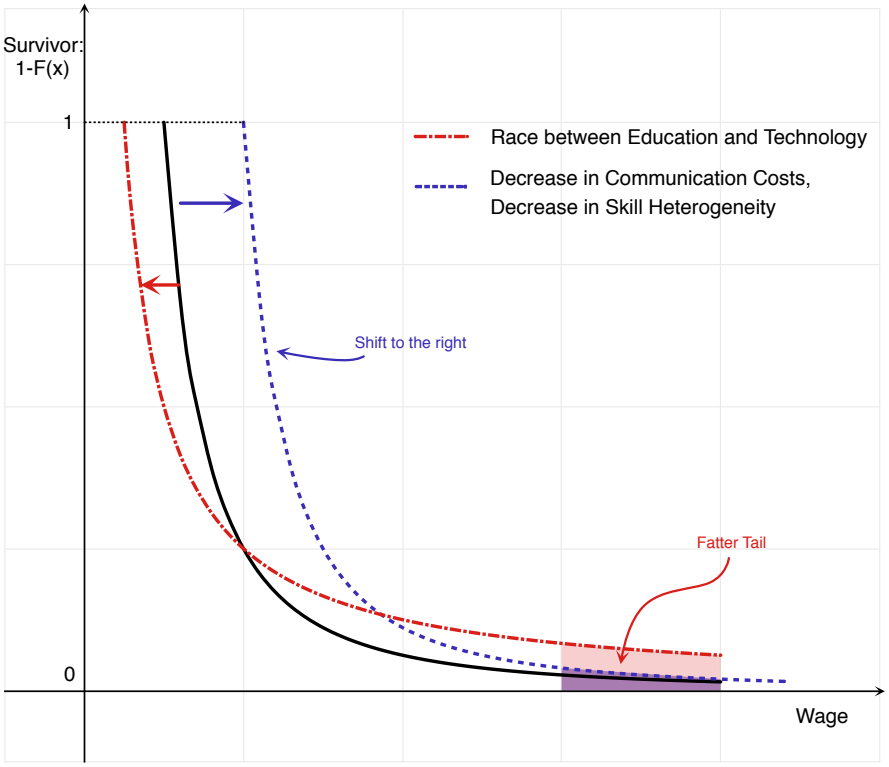
Microfounding Zipf's law in this way is not just interesting for its own sake but because many issues crucially rely on understanding why firm sizes are so heterogeneous in the economy. The most important of them is perhaps that high levels of income inequality are sometimes justified on the basis that top managers have high span of control, so that slight improvements in upper level decisions have a large influence as a whole by affecting the productivity of lesser ranking workers (Rosen (1982)). As a matter of fact, Vilfredo Pareto noticed in the 1890s using tax tabulations from Swiss cantons that the upper tail of the wage distribution is well approximated by a scale-independent "Pareto" distribution (Pareto (1896)) now named after him and shown on French data below. In these theories, the reason for why managers earn very high and very heterogeneous salaries, from ten to more than a thousand times as large as that of the average worker, in the end remains all about the underlying heterogeneity of the firms they run. In other words, one can explain why labor incomes are heterogeneous, but only given that a certain span of control distribution arises in the first place. In the random growth tradition, big firms are those which have been exceptionally lucky, and which have experienced a long and continued sequence of good idiosynchratic shocks. Very rich workers are then those who are skilled enough to be running these lucky firms. Another tradition following Champernowne (1953) has consisted in assuming random shocks to human capital at the individual level, and to view the workers with high earnings as the exceptionally lucky ones.[1] Finally, explanations relying on some primitive being distributed Pareto are a bit tautological, whether they rely on the productivity of firms or the talent of managers being distributed in this way. IQ or other measures of ability are certainly not distributed Pareto, but rather according to distributions with finite support.



Note: Source: see Figure 1. Bins with less than 5 employees are not shown for confidentiality.

[1] Quoting Mincer (1958): "From the economist's point of view, perhaps the most unsatisfactory feature of the stochastic models... is that they shed no light on the economics of the distribution process. It is difficult to see how the factor of individual choice can be disregarded in analyzing personal income distribution." (p. 283)

The theory I develop in this paper generates Pareto distributions for wages, without any heterogeneity in underlying primitives, or any notion of "luck". It shows that market forces can generate on their own a very large amount of inequality in outcomes, from only an epsilon amount of ex-ante heterogeneity. The mechanism actually develops on an idea which was put forward a long time ago by Tinbergen (1956) and then developed by Sattinger (1975): that a matching process with complementarities can lead to wage schedules which are convex functions of skills. In this paper, I take this insight to its furthest extent, in order to explain how such a mechanism can amplify ex-ante small (even infinitesimal) differences, and lead to Pareto distributions in wage schedules through Pareto distributions on span of control, in line with the evidence presented on the previous two figures. A key contribution of the paper, compared to Terviö (2008) or Gabaix and Landier (2008), is therefore to obtain Zipf's law endogenously from the matching process itself, instead of taking it as an assumption. The increase in firm sizes since 1980 can then be related to decreases in communication costs, as in Garicano and Rossi-Hansberg (2006), but also, perhaps counterintuitively, to decreases in skill heterogeneity which arises from a generalization of education to larger fractions of the population, and allowing the most skilled managers to leverage their talent more through higher span of control. (a new qualitative insight) The model even allows to distinguish these two factors from a third one, that of an increased race between education of skilled agents and technology, leading to more inequality through a reduction in the Pareto coefficient, as shown on the Figure below.



Note: The model allows to decompose the rise in labor income inequality into a "race between education and technology" component, changing the heavy tailedness of the distribution, and a decrease in communication costs or a generalized decrease in skill heterogeneity, shifting the top income distribution to the right.

Finally, the analysis also sheds a new light on the growing body of work which emphasizes heterogeneity in firm productivities as a potential source of misallocation across firms, or of reallocation of resources from least to most productive firms consecutive to openness to trade, as in Melitz (2003) type models. The model suggests on the contrary that underlying heterogeneity can actually be bounded, and even infinitesimal, and nonetheless generate Zipf's law for firms in the upper tail; which could suggest a much less important role for the effects just mentioned. In fact, the comparative statics one can draw from my analysis actually go in a rather counterintuitive direction: the less heterogeneity in firm primitives, the more heterogeneity in firm sizes. The reason is that when managers' and workers' skills become closer, the managers' span of control increases because workers need them less. This can suggest a new reason for why developing economies have fewer big firms than developed ones, or why the returns to education are lower there: the underlying skill heterogeneity is higher in developing countries.

The static model which forms the basis of this paper and generates a Zipf's law for firm sizes and Pareto distributions for top labor incomes builds on a well-known Garicano (2000) production hierarchies model, where time can be used for production as well as for communicating one's knowledge. In such a model, production workers encounter problems when producing. When they do not know the solution to this problem, they can potentially ask someone more knowledgeable than them for a solution. If the time of communicating knowledge, called "helping time", is lower than the time of producing by oneself, higher production can sometimes be achieved by having more knowledgeable agents specialize in communicating solutions to problems. Because workers ask these agents only when they do not know the solution themselves, these agents are called managers. The theoretical contribution of this paper is to show that in general, the distribution of span of control for managers over agents down L levels of hierarchical organization is a Pareto distribution of coefficient $2^L/(2^L-1)$ in the upper tail, under minimal assumptions for the distribution of problems as well as the distribution of workers' skills. Hence, Zipf's law obtains in the upper tail with a sufficiently large number of hierarchical levels. While Garicano (2000) and subsequent papers (Garicano and Rossi-Hansberg (2006)) have brought economists mostly qualitative insights about, for example, the impact of organization and decreases in costs of communication on wage polarization, I show that taking the distribution of skills as exogenous as well as allowing for multiple layers of management naturally gives rise to the Pareto properties described above. Compared to their work, my contribution is thus to show that precise and quantitative statements can be made about the upper tail of the span of control distribution as well as that of the wage distribution, *irrespective of the underlying distribution of skills.*

The intuition for why a Pareto distribution for span of control obtains in the upper tail is that the most knowledgeable managers work with the most knowledgeable workers in the competitive equilibrium of this model. This is because of a complementarity between workers' and managers' skills: at an optimal allocation, the most skilled managers must be shielded from answering easy questions, as they can make a more productive use of their time. To the limit, when heterogeneity goes to zero, or when the number of layers of hierarchical organization increases,

the most productive workers are able to solve almost all the problems by themselves, so that they almost never ask managers. Those very productive managers can therefore handle many workers, who could almost produce by themselves, and only ask for their help only when the question is very difficult. However, not all managers can work with those very productive, almost self-sufficient workers, for there is only a limited supply of very skilled workers. Less knowledgeable managers therefore have to "tap" progressively into the knowledge of less knowledgeable workers. The way in which they have to "tap" into this knowledge is given by a Pareto distribution of tail coefficient equal to two, for the same mathematical reason as in Geerolf (2013).[2] Just as in Geerolf (2013), the distribution of the largest firms' sizes corresponding to the most productive workers and the most productive managers does not depend on the functional forms underlying the primitives - distribution of arriving problems, and of agents' skills - because relatively few managers and few workers underly this distribution, so that to the limit, the distribution can be approximated by a uniform distribution. When there are not just one, but many levels of hierarchical organization, the formula for the endogenous Pareto coefficient is $2^L/(2^L-1)$ in general. The intuition for that is that managers at the top of the firm manage other intermediary managers, who themselves manage other workers. The tail of the span of control distribution is thus thicker with multiple levels of intermediary management than with just one of them. For example, in the case of two layers of management, I show that the number of workers that intermediary managers supervise is given by a Pareto distribution of coefficient four in the space of top managers: there are fewer managers than intermediary managers. Since their own span of control over intermediary management is given by a Pareto with coefficient two, total span of control therefore has a tail Pareto exponent given by $1/(1/2+1/4)=4/3$, since it is just a multiplication of the two above distributions. This reasoning generalizes straightforwardly to a case with L layers, with an exponent of $2^L/(2^L-1)$ in general.

One thing that crucially distinguishes this theory from a random growth mechanism for generating Zipf's law is therefore that it not only predicts a Zipf's law for firm sizes - the data lends support to both theories in this respect - but that it also predicts Pareto distributions with endogenous tail coefficients equal to $2^L/(2^L-1)$ for $L=1,2,...$ between intermediary levels of management, something that random growth theory does not. This is a very clear testable implication of the model. The fact that these Pareto distributions are indeed observed in the French data for firms, with very similar tail coefficients, as well as in the Census Data for US establishments, seems to lend support to the theory presented in this paper.

Another empirical success of this theory is that the way in which the distribution of spans of control with non zero heterogeneity differ from the exact Pareto distribution is similar to what is actually found in the data. On a log-log scale, the lower part is generally concave, which can help explain the discussion in the literature about whether Pareto or lognormal is a better fit for the bulk of the distribution. There are also fewer very large firms than would be expected under Pareto, again a well-known deviation from Pareto in the data. In the model, perfect Pareto

[2]There is a clear analogy between lenders and borrowers' leverage ratios in Geerolf (2013) and workers and managers' span of control in this paper.

means zero heterogeneity, while the amount of heterogeneity is backed out through the deviations from this benchmark. For example, to study the potential importance of misallocation, it is really the deviations from Pareto we should be looking at: less heterogeneity in firm sizes means more heterogeneity in primitives. The literature has so far been doing the opposite.

The model also generates Pareto distributions for top labor incomes through a mechanism reminiscent of Terviö (2008) and Gabaix and Landier (2008), whose model with exogenous span of control is a reduced form of the model presented in this paper. This partial equilibrium approach is sufficient for some purposes, for example to relate the increase in CEO wages to the increase in firm sizes which has occurred since 1980: Gabaix and Landier (2008)'s calibration exercice taking firm sizes' evolution as given would be done in the same way. However, endogenizing Zipf's law has several advantages, even to understand the evolution of top labor incomes.

First, one can explain the increase in firm sizes through the lens of the model. It can come from the development and diffusion of information technologies, a qualitative insight already present in Garicano and Rossi-Hansberg (2006). But it can also come from a decrease in overall skill heterogeneity, again because this allows the best managers to hire relatively more talented workers and hence to leverage their differential skill more. The partial equilibrium analysis above would lead to the mistaken opposite conclusion: with more similar skills between managers, marginal products and so wages should also be more similar. This effect is outweighed in the full model by the increase in span of control the best of them enjoy. This new qualitative insight could potentially provide a new rationalization for different evolutions of top labor income inequality across countries. For example, according to the Program for International Student Assessment, the United States and France are at opposite extremes for the evolution of the variance of skills at age fifteen (measured by test scores) between 2000 and 2007.

Second, the relationship between firm sizes and wages is not as straightforward as matching models with exogenous firm sizes would lead to believe. In particular, the largest wages do not necessarily accrue to CEOs managing the largest firms. These largest firms are found in sectors where the most talented managers are in relatively large supply, since they can then hire workers who are themselves very skilled, which allows them to increase their span of control. On the contrary, the largest wages correspond to sectors where the most talented managers are in relatively scarce supply and where firms are relatively small. Such a distinction cannot be made in a partial equilibrium model with exogenous firm sizes, since firm sizes do not then depend on the distribution of skills. This non-systematic relationship between firm sizes and wages provides a potential explanation for why the unconditional regression of CEO wages on their number of employees has an explanatory power only in the range of 20%, and why new tech firms (for example) have both relatively low firm sizes and high wages for top executives.

Third and finally, having a microfounded model of both firm sizes' and labor incomes' distributions can be used to answer potentially many different applied questions. For example, the model can be used to study optimal taxation of top labor incomes with endogenous labor supply, in a Mirrlees (1971) framework. Distorted labor supply decisions would both affect wages

directly as well as through a distorted firm formation process. In contrast, CEO wages are pure economic rents in Gabaix and Landier (2008), which could be taxed away with no efficiency cost. Another example would be to analyze the incentives to innovate in an endogenous growth model with firm formation, as the model allows to evaluate the returns to "discovering" certain types of skills. I come back to these potential extensions in the concluding section.

To summarize, the most important message in this paper is that heterogeneity in primitives, here agents' skills, can very well have finite support, even a support with an infinitesimally small measure, and yet heterogeneity in outcomes appear unbounded. Even if all agents can only produce an amount infinitesimally close to one unit of consumption by themselves, the market economy can lead the most productive of them to earn many millions of these units in equilibrium, while leaving the rest to earn only one unit. Organization and specialization allow the most productive to apply their slightly higher skills very broadly. To the best of my knowledge, this contrasts with all existing theories of inequalities which attribute outcomes with infinite support to very heterogeneous causes, ones with infinite support as well. As already stressed, managers' skills in Lucas (1978) have unbounded support - they are even distributed Pareto. Productivities are Pareto distributed in some models of heterogeneous firms with decreasing returns to scale and fixed costs. In "superstars" theories, all heterogeneity obtained in wages ultimately comes from the heterogeneity in firm sizes that is assumed. In contrast, the model I develop allows to map an empirically very unequal labor income distribution to an underlying potentially very equal distribution in skills, without incorporating any additional element of heterogeneity. Whether the "true model" of the economy has an unbounded dispersion of primitives (productivity, demand, skills), or a bounded one, as this paper would suggest, is likely to matter not just quantitatively, but also qualitatively. I come back briefly to this point at the end of the literature review below when referring to the relevant literature on models with heterogenous firms.

The rest of the paper proceeds as follows. Section 1 develops a model of production hierarchies, with exogenous skills and an endogenous number of layers of management. Section 2 focuses on the properties of the equilibrium span of control distribution, without reference to equilibrium prices. It shows that the span of control distribution between L layers of hierarchical organization is a Pareto distribution in the upper tail, and that its coefficient is equal to $2^L/(2^L-1)$. Section 3 calculates and studies the labor income distribution sustaining these allocations, as a function of three key underlying primitives of the model: heterogeneity, communication costs, and race between education and technology. Section 4 discusses the connections between the paper and the random growth literature and the literature on the economics of superstars. Section 5 concludes.

Related Literature

This paper provides an alternative to the random growth literature for generating Zipf (1949)'s Law. The intuition in random growth models for why heterogeneity in firm sizes, city sizes or incomes is very large is that firms, cities or individuals in the tail of the Pareto distribution had a particularly long and unlikely continued sequence of good idiosynchratic shocks. Scale independence, a key property of Pareto distributions, crucially results from the assumed scale independence in the underlying stochastic growth process, which translates into that of the stationary distribution. Models along these lines therefore crucially assume that Gibrat (1931)'s law holds, or that the distribution of firm growth rates is independent of their size, which is not exactly true, as variance of growth rates decreases a bit with size (see Mansfield (1962) for early evidence along these lines). The random growth literature is surveyed in Gabaix (2009) and Luttmer (2010), and comprise for example Simon (1955), Simon and Bonini (1958), Kesten (1973), Sutton (1997), Gabaix (1999), Axtell (2001), Luttmer (2007), Rossi-Hansberg and Wright (2007). Champernowne (1953), or more recently Jones and Kim (2014) are applications of the random growth mechanism to the income distribution. Importantly, these random growth models are not usually considered as being microfounded, as the source of idiosynchratic random shocks is not well understood, and these shocks are assumed to be uninsured. Moreover, these models need "statistical frictions" (like a reflecting barrier), for the variance not to grow without bound. Section 4.1 discusses the connections between random growth and this paper in more detail.

This paper most chiefly belongs to the span of control literature, developed since Lucas (1978). Lucas (1978) assumes a Pareto distribution in skills of managers to explain Pareto distributions in span of control (and, mechanically, in wages), an assumption my analysis can do away with. In contrast, the model presented in this paper is a microfoundation for the Pareto nature of the top labor income distribution in the upper tail, where heterogeneous salaries endogenously come from an assignment process with complementarities. The intuition develops on this idea that managers earn large salaries not because they really are more productive taken in isolation, but because the marketplace allows them to leverage their small differences of talents through managing more people. The model is therefore part of the large competitive assignment literature, comprising Tinbergen (1956), Rosen (1974), Sattinger (1975), Rosen (1981), Kremer (1993), Teulings (1995), Terviö (2008) and Gabaix and Landier (2008). In particular, a crucial assumption underlying Garicano (2000) as well as similar models of sorting is that quality and quality of workers are imperfect substitutes, and that there is no such thing as a concept of "efficiency units of labor" (Eeckhout and Kircher (2012)). More topically, because it provides a microfoundation for the Pareto nature of the labor income distribution, the model can potentially speak to the causes of the recent rise in labor income inequality, documented in particular by Piketty and Saez (2003) for the United States, and discussed extensively in Piketty (2014). It can try to address the different explanations which have been put forward for this rise: skill-biased technological change (Acemoglu and Autor (2011)), race between education and technology (Katz and Murphy (1992)), or decrease in communication costs (Garicano and Rossi-Hansberg (2006)).

The model also speaks to a large literature on organizational structure, a survey of which is given in Radner (1992) for the older literature. The paper most chiefly builds on Garicano (2000)'s production hierarchies model to investigate the distribution of firm sizes in the economy. More precisely, it draws heavily from the applied production hierarchies models developed in Garicano and Rossi-Hansberg (2004), Garicano and Rossi-Hansberg (2006), Antràs et al. (2006) and Caliendo and Rossi-Hansberg (2012). The closest to the paper is perhaps Garicano and Rossi-Hansberg (2006), for both its focus on wage inequality and the methods used. However, no mention to Pareto distributions is made in this paper. The authors further confine their analysis to the case of only two layers, restricting attention to qualitative insights about wage polarization for example. In contrast, I show that progress can be made without reference to the underlying distribution of skills by focusing only on the tail of the distribution.

This paper uses the same intuition, and a similar methodology, as Geerolf (2013) to generate a Pareto distribution of tail coefficient equal to two. The use of French data for production hierarchies, and in particular the occupation code as an indication of workers' position in a firm's hierarchical structure follows Caliendo et al. (2014) and Liegey (2014).

Finally, Pareto functional forms for primitives are used in a far more distant pieces of the economics literature, since research on firm heterogeneity takes place at the intersection of macroeconomics, labor economics, trade and industrial organization. For example a very large literature in trade with heterogenous firms following Melitz (2003) uses evidence of Pareto firm sizes to justify Pareto distributions in productivity. There is also a large literature in macroeconomics on misallocation, speaking to the welfare consequences of microeconomic and macroeconomic distortions, using abundantly the firm size distribution to back out productivity differences, among which Hopenhayn and Rogerson (1993), Restuccia and Rogerson (2008), Hsieh and Klenow (2009). In the light of this paper, it seems on the contrary that the shape of firm size's distribution says very little on the underlying distribution of productivity across them, even on the relative importance of their dispersion.

1 A Model Economy

1.1 Setup

I consider a static economy, with a continuum of agents indexed by i. There are two commodities, time and a consumption good. Agents are endowed with one unit of time, which they supply inelastically for producing the good or for helping others solve problems, as in Garicano (2000). Agents encounter problems in production, whose set is indexed on $[0, 1]$. The arrival of these problems is uniformly distributed on this interval, and the problems' indexes are higher when these problems are harder to solve - that is, less agents know how to solve them. This is without loss of generality, as from any initial draw of problems and their probability distributions, one can just relabel the problems such that this is verified. I follow the literature and denote the cumulative distribution function for the arrival of problems by $F(x) = x$ on $[0, 1]$.

Agents' heterogeneity. Agents i are assumed to be heterogeneous in terms of how many problems they can solve by themselves: they generally do not know everything. It is assumed that knowledge is cumulative, as in Garicano and Rossi-Hansberg (2004) or Antràs et al. (2006), so that an agent with a higher skill knows everything that an agent with lower skill knows - see Garicano (2000) for a discussion.

The skill of agent i is indexed by the hardest problem he can solve: agent i with skill z^i can solve all problems contained in the set $[0, z^i]$. The distribution of agents' skills over problems is given by a cumulative distribution function $G(.)$, with density $g(.)$ over $[0, 1]$.

It is assumed that the complexity of producing the good in the economy is such that only the most skilled of them would know how to produce the good by themselves. In other words that the density function has support $[1 - \Delta, 1]$, where Δ indexes, without summarizing in general, the level of skill heterogeneity in the economy.[3] Heterogeneity therefore has finite support: left to their own devices, agents can produce between $1 - \Delta$ and 1 unit of good by themselves.

Production. The organization of production follows Garicano (2000). Apart from producing with their time as a factor of production, agents can also communicate solutions to other agents. When production workers or intermediary managers know the solution to these problems, they can produce one unit of output per unit of time. But if they do not, they can ask their managers. Regardless of whether the manager knows the solution or not, it takes $h < 1$ units of time for a manager to communicate a solution to the worker.[4] h is called the helping time. Potentially, managers can also ask other agents to solve the problems if they do not know the solutions to them, so that hierarchies can form.

A crucial assumption is that workers cannot diagnose or label problems when they cannot solve them. They sometimes waste their direct managers' time by sending them problems that they do not know how to solve, and have to pass on higher up in the hierarchy. In other words, agents do not always directly ask the agents who know. I call manager of type l a manager who answers questions which have been transmitted l times: managers of type 1 directly supervise production workers, managers of type 2 answer questions of managers of type 1, and so on.

1.2 Equilibrium

In equilibrium, agent i with skill z^i chooses to allocate his time being a production worker, a manager of type $l \in \{1, 2, ...\}$, or self-employed in order to maximize his utility, or equivalently his income since there is just one consumption good and time is supplied inelastically.

[3] One could generalize this to having a non-trivial measure of agents able to solve all problems, but this would only add sub-cases, without adding much intuition to the model. On the other hand, the results rely on some agents being able to solve all problems arising in production, at least to the limit. It is doubtful that agents would in equilibrium decide to produce a good which would pose unsolvable problems in production even to the most skilled of them. Another option would be to endogenize the choice of the good being produced.

[4] $h < 1$ is assumed for agents to find it (sometimes) optimal to form teams: if the time of communicating knowledge is greater than the time of producing, then it is always better to engage in self-production.

Denote by L the maximum distance between two levels of hierarchical organization in this economy - I show later that when managers' time is indivisible, L is a finite number. I call agents who use their time producing "production workers". There are L types of different managers, who communicate solutions to problems.

Formally[5], an agent is denoted as having a negative position in production worker time $t^i_W(.)$ when he hires workers, and a positive one $t^i_W(z^i)$ when producing as a production worker. Symmetrically, an agent can hire managers of lower levels who pass on problems they cannot solve, so that the manager of type l time $t^i_{M_l}(z^i)$ can be positive or negative for any $l \in \{1, ..., L-1\}$. In contrast, an upper level manager by definition cannot be hired by anyone, agent i's time in this position is denoted by $t^i_{M_L}$. An agent can also spend time t^i_S being self-employed.

Denoting by $w_0(z)$ the wage from being a production worker when of skill z (or the price of production time at that skill level), and similarly by $w_l(z)$ the wage from being a manager of type $l \in \{1, ..., L-1\}$, and given that self-employment gives z as income, the agent chooses to occupy his time so as to maximize his income (I) subject to his total time constraint of one unit (TC) and his communication time constraints (CTC_0) and (CTC_l) given by the unit time of communicating h multiplied by the number of time he has to answer his workers' questions:

$$\max_{\left(t^i_S, t^i_W(.), \{t^i_{M_l}(.)\}, t^i_{M_L}\right)} \quad z^i t^i_S + \int_z w_0(z) t^i_W(z) dz + \sum_{l=1}^{L-1} \int_z w_l(z) t^i_{M_l}(z) dz + z^i t^i_{M_L} \qquad (I)$$

$$\text{s.t.} \quad (TC) \qquad t^i_S + \int_z \max\{t^i_W(z), 0\} dz + \sum_{l=1}^{L-1} \int_z \max\{t^i_{M_l}(z), 0\} dz + t^i_{M_L} \leq 1$$

$$\text{s.t.} \quad (CTC_0) \qquad \int_{z \neq z^i} h(1-z) t^i_W(z) dz \leq t^i_{M_1}$$

$$\text{s.t.} \quad (CTC_l) \qquad \forall l \in \{1, ..., L-1\}, \qquad \int_{z \neq z^i} h(1-z) t^i_{M_l}(z) dz \leq t^i_{M_{l+1}}$$

$$\text{s.t.} \qquad \forall z \neq z^i, \qquad t^i_W(z) \leq 0$$

$$\text{s.t.} \qquad \forall l \in \{1, ..., L-1\}, \forall z \neq z^i, \qquad t^i_{M_l}(z) \leq 0$$

Note that if $t^i_W(z) > 0$, then agent i is a production worker, if $t^i_{M_l} > 0$, then agent i is a manager of type l, and if $t^i_S > 0$, then agent i is self-employed. The before last constraint states that an agent can hire any type of production worker, but can only supply a positive amount of time $t^i_W(z^i)$ with his skill. This is why all other $t^i_W(z)$ for $z \neq z^i$ must be non positive. Similarly, an agent can hire any type of intermediary manager in principle, but can only supply a positive amount of intermediary time $t^i_{M_l}(z^i)$ for $l \in \{1, ..., L-1\}$ with his skill. A Competitive Equilibrium of this production economy $\mathcal{E}^L$ is then defined as follows.

[5]Readers who have only limited interest in technical details may wish to skip the formal definition of an equilibrium, and head directly to Section 2. Solving for the equilibrium can be done in a much more intuitive and straightforward way than by solving the model below. I however intend to show by this formalism that there is nothing more to finding Pareto distributions than solving for a particular General Equilibrium problem.

Definition 1 (Competitive Equilibrium of $\mathcal{E}^L$)**.** *A* Competitive Equilibrium for Economy $\mathcal{E}^L$ *is a wage function for production workers* $w_0(.)$*, wage functions* $w_l(.)$ *for all managers of type* $l \in \{1, L-1\}$ *and allocations of time* $\left(t_S^i, t_W^i(.), \{t_{M_l}^i(.)\}_{l=1}^{L-1}, t_{M_L}^i\right)$ *for all agents* i *such that agents maximize their income (I) under the time constraint (TC), the communication time constraints (CTC*$_l$*) for* $l \in \{0, ..., L-1\}$*, taking the wage functions* $w_l(.)$ *for* $l \in \{0, ..., L-1\}$ *as given, and the market for work time, and lower-level managers' time clears for all skill levels, that is:*

$$\forall z, \quad \int_i t_W^i(z)di = 0. \tag{MC_0}$$

$$\forall l \in \{1, ..., L-1\}, \quad \forall z, \quad \int_i t_{M_l}^i(z)di = 0. \tag{MC_l}$$

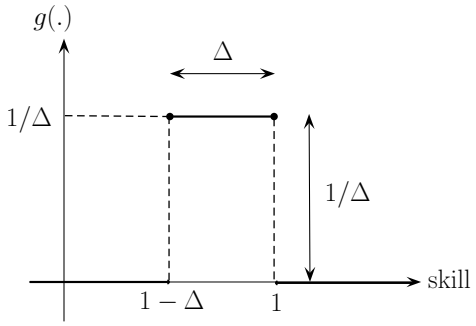
The program is linear in $\left(t_W^i(.), \{t_{M_l}^i(.)\}_{l=1}^{L-1}, t_S^i\right)$, so it is optimal in equilibrium for agents to do one of three things: use their whole time producing in a team, or managing and communicating solutions at a given level of management, or engage in self-employment. Following Caliendo and Rossi-Hansberg (2012), I assume that a top manager can only work in one firm, and not manage multiple firms.[6]

Assumption 1 (Indivisibility)**.** *A firm is run by a manager working full time at the top of his organization.*

Finally, a lot in this paper revolves around solving the model in closed form for the uniform distribution of skills. In the paper, I refer to the uniform distribution of skills with heterogeneity Δ which corresponds to the distribution in Definition 2.

Definition 2 (Uniform Distribution)**.** *The* uniform distribution of skills with heterogeneity parameter Δ *is defined as a function of* Δ *by:*

$$g(x) = \begin{cases} 0 & \text{if } x < 1-\Delta \\ 1/\Delta & \text{if } x \in [1-\Delta, 1] \\ 0 & \text{if } x > 1. \end{cases}$$



[6] I could as well assume that firms can be run by two or three agents, or that agents can run at most an integer number of firms, as long as it is a fixed number. In practice, conflict of interest or management issue would certainly arise in that case, so that the constraint can be thought of as a reduced form of an upper level model with an explicit decision-making process.

1.3 Discussion

In order to get quickly at the main results of the model, I assume a fixed, exogenous distribution of skills. Endogenous skill acquisition is potentially a fruitful extension, as in Garicano and Rossi-Hansberg (2006).

Unlike in Antràs et al. (2006), I do not here "force" agents to join teams. They work in cooperation with others only when they have an interest in doing so, that is when they gain more working in teams than working as self-employed. I show in Section 2.1 that it is the case when heterogeneity in skills, or the cost of communicating solutions to problems, is sufficiently low. Since occupational choice is a crucial element for the formation of Pareto distributions for span of control, I don't want to restrict the choice set exogenously in this dimension.

Another difference with Antràs et al. (2006) is that I do not restrict the maximum number of layers of hierarchical organization to be one, but the number of layers L is determined endogenously. This higher number of layers renders the mathematics as well as the notations a bit more complicated, but it is crucial to get the Pareto results.

2 Span of Control Distribution

As can be inferred from Definition 1, a Competitive Equilibrium of the model defined above is potentially a high dimensional object, notably because the number of layers L is potentially high.

Unlike the model in Geerolf (2013) however, this model has a block-recursive property. One can solve for the allocations independently from the prices sustaining these allocations. The decentralization of the allocations through wages, whose distributions also turn out to be Pareto under some conditions, is the subject of Section 3.

This Section is organized as follows. In Section 2.1, I show that when heterogeneity in skills is sufficiently low, there is no self-employment in equilibrium. In Section 2.2, I demonstrate positive sorting of workers and managers by skills and derive the main differential equation of the model, which drives the Pareto results. In Section 2.3, I assume that the distribution of skills is uniform, to illustrate the Pareto properties of the span of control distribution because it allows for closed form solutions. In Section 2.4, I provide an intuition of the Pareto distribution of coefficient 2, and the Pareto distributions of coefficient $2^L/(2^L-1)$, which gives an intuition for scale independence and allows to better understand why the result does not depend on the underlying distribution of skills.[7] Section 2.5 generalizes mathematically the argument to a bounded away from zero density function for skills, and Section 2.6 investigates the case where the density function is not bounded away from zero. Last but not least, Section 2.7 presents supporting evidence for this theory. In particular, it shows that, in the French matched employer-employee data, the distribution of span of control down one level of hierarchical organization is indeed distributed

[7] A skilled reader worried about his time constraint may therefore want to go directly to Section 2.4, which gives an intuition for Zipf's law, through graphical rather than mathematical arguments. Unfortunately, it does not convey the comparative statics about heterogeneity in particular.

Pareto with a coefficient close to 2 in the upper tail. The distribution of establishments' sizes and of number of establishment per firms also behaves in a way predicted by the theory.

2.1 Ruling out Self-Employment

The following lemma states the first result: for helping time h and heterogeneity Δ sufficiently low, there is no self-employment in the equilibrium of this model.

> **Lemma 1** (No self-employment Condition)**.** *For a sufficiently low value of skill heterogeneity Δ and of communication time h, there is no self-employment in equilibrium.*

Proof. See Appendix B.1. □

In the rest of the paper, I assume that Δ or h are indeed sufficiently low (Assumption 2). This avoids a cumbersome discussion of different cases and sub-cases and to go more quickly at the main results of the model. An alternative would have been to assume as in Antràs et al. (2006) that self-employment is not an option from the outset. The comparative statics exercices with respect to heterogeneity are always undertaken under Assumption 2. However, the reader can verify easily that the conclusions I draw from these comparative statics would be strengthened if self-employment was added to the picture (notably, that firm heterogeneity as well as inequality are decreasing functions of skill heterogeneity).

> **Assumption 2** (No self-employment)**.** *Δ and/or h are low enough, so that there is no self-employment in equilibrium.*

To get an idea of how restrictive this assumption might be, one can look at the shape of Assumption 2 in the case of a uniform distribution of skills with disagreement Δ (see Definition 2). The set $\mathcal{A}_2$ of (Δ, h) values such that the assumption is verified can then be expressed in closed form:

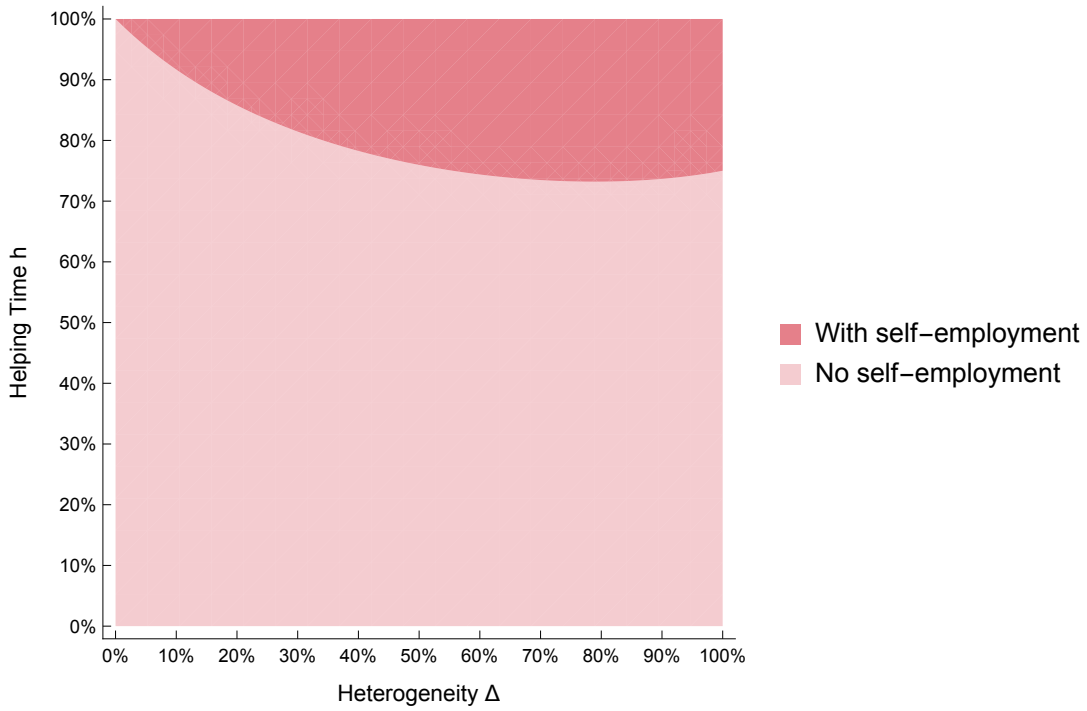
$$\mathcal{A}_2 = \left\{ (\Delta, h); \quad (\Delta, h) \in [0,1]^2; \quad h > \frac{2\sqrt{1+2\Delta-2\Delta^2}-1-\Delta}{1+2\Delta-3\Delta^2} \right\}$$

This particular expression is derived in Appendix B.3. The self-employment and no self-employment regions are drawn on Figure 2 as a function of the heterogeneity parameter Δ. Assumption 2 is valid across most of the parameter space. In particular, for any $h < 1$, there exists $\Delta > 0$ such that for any heterogeneity Δ lower than this threshold, the assumption is verified.

2.2 Positive Sorting and Allocations

Proposition 1 describes, given a maximum number of layers of management, the allocations in this model: the occupational choice of agents, and who works for whom (defining the boundaries

Figure 2: Validity of Assumption 2 in the Uniform Case

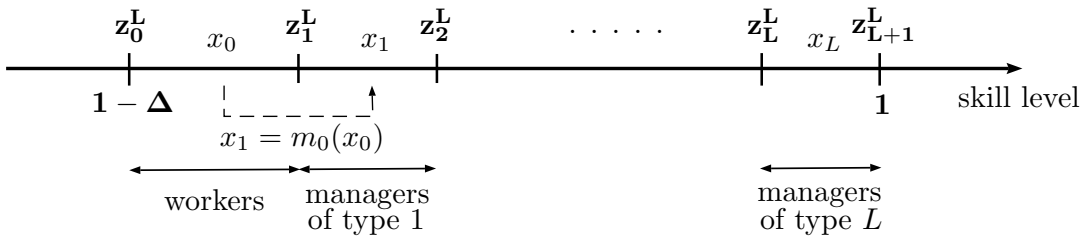


Note: The parameter space for which there is no self-employment is the upper-right panel of this Figure. Note that for any value of helping time h, even one very close to production time ($h = 100\%$), there is no self-employment in equilibrium if heterogeneity Δ is sufficiently low.

of the firm). Proposition 2 gives the equilibrium number of layers, for a finite number of workers - and when the skill distribution with probability distribution function given by $G(.)$ is the continuous limit of the underlying discrete skill distribution. Together, Propositions 1 and 2 uniquely define the allocations of a Competitive Equilibrium defined in Definition 1.

Before stating Proposition 1, it is convenient to denote the limiting bounds for the support of skills for any number of layers L as $[z_0^L, z_{L+1}^L] = [1 - \Delta, 1]$, as on Figure 3.

Figure 3: Equilibrium Allocations: Notations



Note: Only "workers" use their time producing, and draw problems in production. Managers of type l for $l \in \{1, ..., L\}$ do not draw problems, but only spend their time communicating solutions to problems. Workers do not know how hard the problem is, so they cannot go directly to top managers (of type L) if the problem is very hard. Instead they must waste every intermediary manager's time before the question reaches the top.

Definition 3 (z_0^L, z_{L+1}^L)**.** *Let z_0^L and z_{L+1}^L denote the bounds of the support of skills:*

$$z_0^L = 1 - \Delta \qquad z_{L+1}^L = 1.$$

Proposition 1 (Allocations, given L)**.** *Assume a maximum number of layers L of hierarchical organization. There exists L endogenous cutoffs $\{z_l^L\}_{l=1}^L$ uniquely given by a set of L equations:*

$$\forall l \in \{0, ..., L-1\}, \quad G(z_{l+2}^L) - G(z_{l+1}^L) = h \int_{z_l^L}^{z_{l+1}^L} (1-u)g(u)du,$$

such that agents with skills in $[z_0^L, z_1^L]$ are production workers, and pass on problems to managers of type 1 with skills in $[z_1^L, z_2^L]$ and so on, until managers of type L in $[z_L^L, z_{L+1}^L]$, the top managers of the firm.

Workers and different levels of managers pass on problems to each other according to increasing matching functions $\{m_l(.)\}_{l=0}^{L-1}$, respectively defined on $[z_l^L, z_{l+1}^L]$ with values in $[z_{l+1}^L, z_{l+2}^L]$ through:

$$\forall l \in \{0, ..., L-1\}, \quad \forall x_l \in [z_l^L, z_{l+1}^L], \quad m_l'(x_l)g(m_l(x_l)) = h(1-x_l)g(x_l) \quad \textit{and} \quad m_l(z_l^L) = z_{l+1}^L.$$

Proof. There is complementarity between the skills of workers and team managers, since the joint surplus of managers of type x_{l+1} supervising workers of type x_l is supermodular x_{l+1} and x_l and given by:

$$\frac{\partial^2 (N_{1+1\to 0}^l(x_l) w_{l+1}(x_{l+1}))}{\partial x_l \partial x_{l+1}} = w_{l+1}'(x_{l+1}) \frac{\partial N_{1+1\to 0}^l}{\partial x_l}(x_l) > 0$$

An optimality argument allows to state that in the competitive equilibrium, there is positive sorting of CEOs, lower ranking managers, and workers. The function mapping a worker with skill x_l with a team manager of skill x_{l+1} is denoted by $m_l(.)$, strictly increasing by positive sorting, defined on $[z_l^L, z_{l+1}^L]$, and such that $m_l(x_l) = x_{l+1}$. The differential equation results from the time constraint of a manager in layer $l+1$, who supervises agents with skill x_l. The number of problems these agents cannot solve is $1 - x_l$, and the communication time for a unit problem is h, so the matching function from workers to managers $m_l(x_l)$ must be such that the total time of managers with skills in interval $[m_l(x_l), m_l(x_l) + dm_l(x_l)]$ given by $g\left(m_l(x_l)\right) dm_l(x_l)$ is equal to the time it takes to answer questions of workers in interval $[x_l, x_l + dx_l]$, whose number is $g(x_l)dx_l$, and whose unit time requirement is $h(1-x_l)$. Formally, one just needs to manipulate the time constraint of managers and write that the market for skills clears:

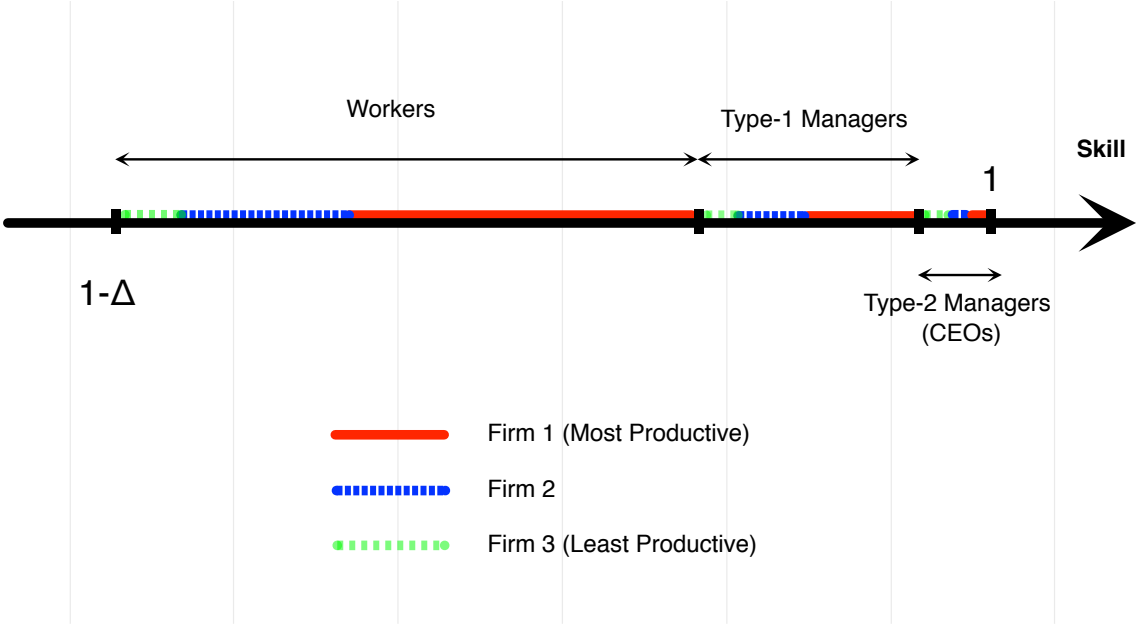
$$\int_i t_W^i(x_l)di = 0 \quad \Rightarrow \quad g(m_l(x_l))dm_l(x_l) = h\left(1-x_l\right) g(x_l)dx_l.$$

$$\Rightarrow \quad m_l'(x_l)g(m_l(x_l)) = h\left(1-x_l\right) g(x_l). \qquad (M_l)$$

□

The positive sorting the most skilled CEOs, with the most skilled intermediary managers and the most skilled workers in the same firms is illustrated on Figure 4.

Figure 4: Illustration of Positive Sorting - Example with $L = 2$, Three Firms, and a Uniform Distribution of Skills



Note: This chart shows how agents with different skills join different hierarchies in equilibrium. It is assumed that $L = 2$, that there are 3 firms, and that the distribution of skills is uniform. Each CEO spends his whole time at the top of his organization. The firm in red consists of the most productive CEO, the most productive type 1 managers, and the most productive workers. Because they are more skilled, type-1 managers of the red firm ask questions less often to their CEOs, which allows the CEOs of the red firm to achieve a higher span of control, and therefore hire a larger range of the skill distribution. The same thing happens for workers with type-1 managers.

Note that the matching functions depend on the assumed maximum number of layers L, just as the cutoffs, and so should normally be denoted by $\{m_l^L(.)\}_{l=0}^{L-1}$, but this is omitted for conciseness (see Figure 3 above for the detail of the notations).

A second result from the paper is that the equilibrium number of layers is given endogenously in this model by equation (L) in Proposition 2 below. This result is intuitive: when the number of layers increases, the number of agents in the upper layer becomes smaller and smaller, and tends to zero, as stated in the following Lemma 2.

Lemma 2 (Size of Top Layer). $z_L^{L+1} - z_L^L \to 0$ *when* $L \to \infty$.

Proof. This results straightforwardly from Proposition 1. See Appendix B.2. □

When the number of agents in this upper-level layer becomes lower than one, this add-in of a new layer becomes irrelevant, under the indivisibility assumption 1. Lemma 2 thus leads to Proposition 2.

Proposition 2 (Equilibrium L)**.** *For any finite number of workers N_w, the maximum levels of hierarchical organization is given by:*

$$L = \max_{L \in \mathbb{N}} \left\{ L \quad s.t. \quad 1 - z_L^L \geq \frac{1}{N_w} \right\} \tag{L}$$

It is useful to look at some numbers to get an idea of the speed of convergence. For example, the case of a uniform distribution over the maximum support of skill hetereogeneity $[0, 1]$ (that is, hetereogeneity is $\Delta = 1$) is investigated in Table 1. From this table, one notes that if the number of workers is one million, that is $N_w = 1,000,000$ then the integer constraint becomes binding for $L = 4$, so that the maximum levels of hierarchical organization is given by $L = 3$. In this model, the limited amount of time that a top manager is what determines the boundaries of the firm; and even the industrial organization of the economy. This is potentially interesting because scholars have long been interested in the distribution of firm sizes to know for example whether, if anything, something ought to be done about the high number of big firms. In this theory, both the number of firms and the number of employees in each firm are endogenously determined by the indivisibility of managers' time; and so is the industrial organization structure of a sector, even if the underlying technology has constant returns.[8]

Table 1: Uniform Skill Distributions: Cutoffs for Different values of L, with $\Delta = 100\%$ and $h = 75\%$

	$\mathbf{z_1^L}$	$\mathbf{z_2^L}$	$\mathbf{z_3^L}$	$\mathbf{z_4^L}$	$\mathbf{z_5^L}$	$\mathbf{z_6^L}$
$\mathbf{L = 1}$	0.666667					
$\mathbf{L = 2}$	0.625993	0.948538				
$\mathbf{L = 3}$	0.625000	0.947266	0.998958			
$\mathbf{L = 4}$	0.625000	0.947266	0.998957	1.000000		
$\mathbf{L = 5}$	0.625000	0.947266	0.998957	1.000000	1.000000	
$\mathbf{L = 6}$	0.625000	0.947266	0.998957	1.000000	1.000000	1.000000

Note: In this table, the cutoffs are calculated with distribution of skills in the population given by $G(z) = \frac{z-(1-\Delta)}{\Delta} \mathbb{1}(1-\Delta, 1)(z) + \mathbb{1}(1, +\infty)(z)$. Precision up to 6 digits is displayed. Successively, the maximum number of levels is restricted to being $L = 1$, $L = 2$, $L = 3$, $L = 4$, $L = 5$, $L = 6$. Heterogeneity in skills is taken to be maximal $\Delta = 100\%$ and so is helping time $h = 75\%$.

Proposition 1 and Proposition 2 together fully characterize the equilibrium allocations in the model. The main results from the paper then pertain to the endogenous span of control

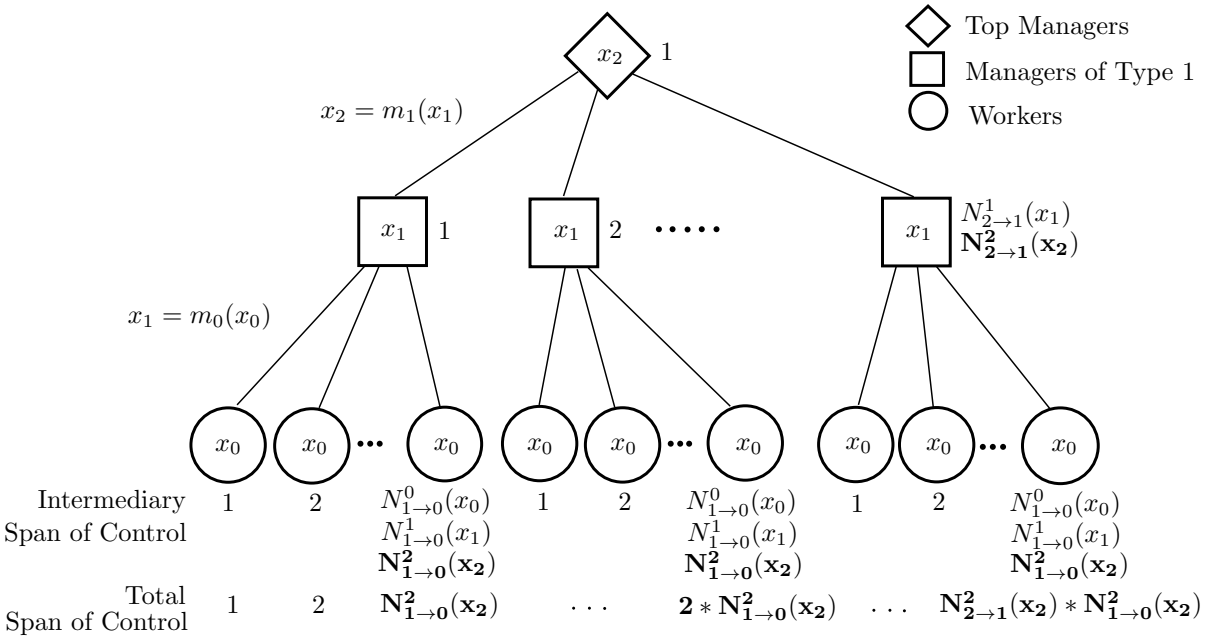
[8] Another option to endogenize the number of layers would be to assume a fixed cost of creating a new layer. Since the marginal gains to adding a new layer are decreasing with each new layer, the number of layers would be determined, and potentially different across firms. This is perhaps more consistent with the real-world counterpart.

distributions that result from these allocations, and from which Zipf's Law is obtained.

For this purpose, some notations need to be introduced. Denote by $N^{l_3}_{l_1\to l_2}(x_{l_3})$ the equilibrium span of control of layer l_1 over layer $l_2 < l_1$, expressed as a function of workers of layer l_3 skills, or in the "space" of workers of layer l_3 skills. For example, the time constraint of a team manager in layer $l+1$ with skill x_{l+1} helping a number $N^l_{l+1\to l}(x_l)$ of workers of skill x_l gives his equilibrium span of control $N^l_{l+1\to l}(x_l)$ through his time communicating constraint (CTC_l), which holds with equality at the optimum:

$$h\,(1-x_l)\,N^l_{l+1\to l}(x_l)=1 \quad \Rightarrow \quad N^{l+1}_{l+1\to l}(x_{l+1})=\frac{1}{h\left(1-m_l^{-1}(x_{l+1})\right)}.$$

Figure 5: Span of Control: Notations - Example with one Firm, $L=2$ Layers



Note: This chart is inspired from Rosen (1982)'s Figure 1. It represents a firm with $L=2$ layers, whose top manager (or manager of type-2) has skill x_2, supervises intermediary managers (managers of type-2) of skill $x_1 = m_1^{-1}(x_2)$, who themselves supervise workers of skill $x_0 = m_0^{-1}(x_1)$. The span of control of the top managers over intermediary managers is denoted by $N^1_{2\to1}(x_1)$ as a function of intermediary managers' skills, and $N^2_{2\to1}(x_2)$ as a function of their own skill (so that $N^2_{2\to1}\circ m_1 = N^1_{2\to1}$). Total Span of Control of a top manager over both intermediary managers and workers is given by $N^2_{2\to1}(x_2)*(N^2_{1\to0}(x_2)+1)$ which is behaves like $N^2_{2\to1}(x_2)*N^2_{1\to0}(x_2)$ when span of control $N^2_{1\to0}(x_2)$ is large.

Note in passing that $N^l_{l+1\to l} = N^{l+1}_{l+1\to l}\circ m_l$. All the other span of control functions $N^{l_3}_{l_1\to l_2}(x_{l_3})$ result from these fundamental span of control functions between intermediary levels of management, up sometimes to a small change of variable through matching functions $m_l(.)$ for some $l\in\{1,...,L-1\}$. The notations for different span of control distributions are illustrated in the case of a firm with $L=2$ layers on Figure 3, inspired by Figure 1 in Rosen (1982)'s seminal contribution.

Note that we do not need the wage function to solve for the span of control distribution in this economy, unlike in Geerolf (2013).[9] This is why the order of differential equations in this paper never is greater than one, which simplifies the problem considerably.

As can be seen in Table 1, the measure of top managers becomes very small when the number of layers increases. The measure of managers of type $L-1$ also becomes very small. Intuitively, this is the reason why the model's predictions on Pareto distributions do not depend on the underlying distribution of skills: the support of the distribution of skills which really matters for values of high span of controls is very small, and any smooth bounded away from zero, distribution, in such a region can as a first approximation be approximated by a uniform. Given that all the limiting results shown in the following are always given with respect to the uniform distributions, it is therefore useful first to solve for the case with a uniform distribution of skills, where all allocations and span of control distributions obtain in closed form, as I show in Section 2.3. In Section 2.5, I show that this case is actually relevant to understand the behavior of span of control distributions in the upper tail, any time the skill distribution is bounded away from zero. In Section 2.6, I show that Zipf's law also arises when the density of the skill distribution is not bounded away from zero, and the number of levels becomes large. But let us look first at the case of a uniform distribution of skill.

2.3 Uniform Distribution of Skills

The uniform distribution of skills allows for closed form solutions. (at the exception of the cutoffs) In this part, I therefore derive these closed forms, where the Pareto distributions become apparent. Section 2.4 derives these same results in a more heuristic way. Let us look first at the span of control distribution of managers down one level of hierarchical organization (Section 2.3.1), then go to two levels of hierarchical organization (Section 2.3.2) and finally generalize to L levels (Section 2.3.3).

2.3.1 One Level of Hierarchical Organization ("Teams")

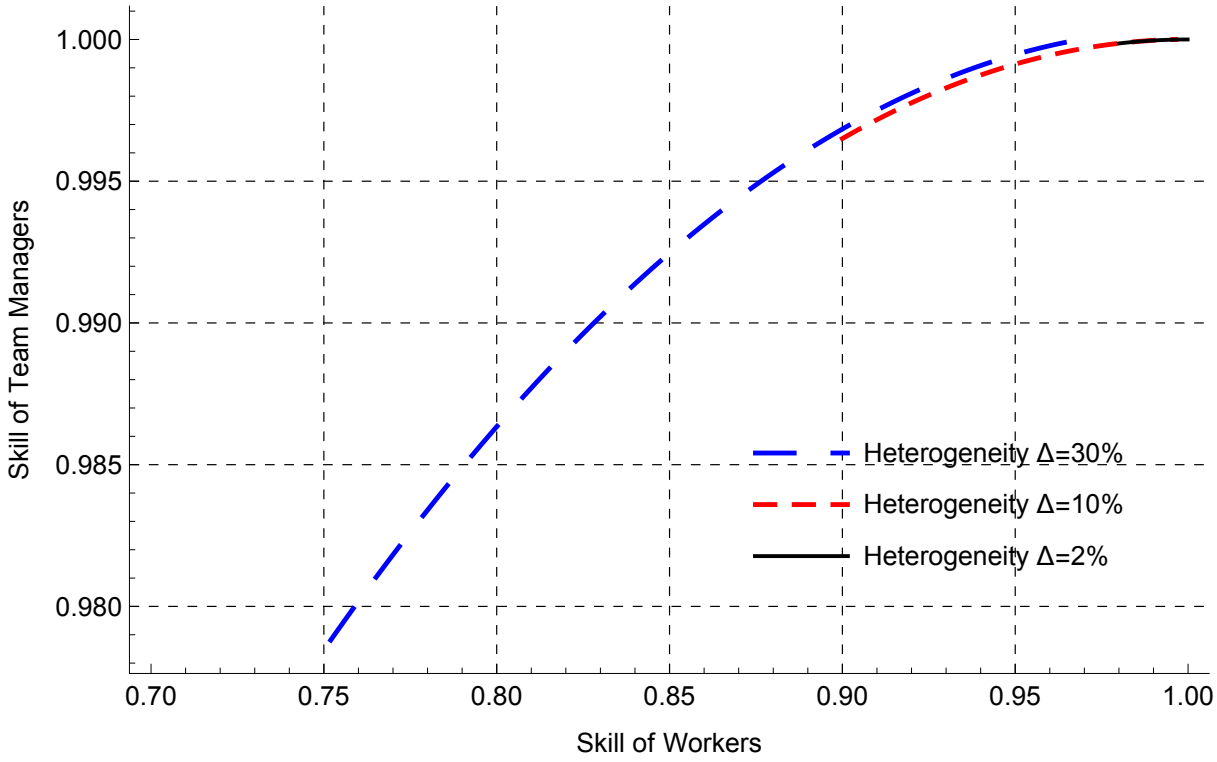
In the uniform case, the primitives of the model are summarized by the heterogeneity in skills Δ and the helping time h (see Definition 2). In particular, the inverse of the equilibrium matching functions is given in closed form through:

$$z_{l+2}^L - m_l(x_l) = h\int_{x_l}^{z_{l+1}^L}(1-u)du \quad \Rightarrow \quad 1 - m_l^{-1}(x_{l+1}) = \sqrt{(1-z_{l+1}^L)^2 - \frac{2}{h}(z_{l+2}^L - x_{l+1})}.$$

This allows to express the span of control between adjacent levels of hierarchical organization in the following Lemma 3.

[9] This is because in Geerolf (2013), the leverage factor depends on the price of bond contracts sold, which is a equilibrium outcome of the assignment process.

Figure 6: Matching Function from Workers to Team Managers, $h = 70\%$ - Uniform Distribution



Note: This figure gives the matching function from workers to team managers derived in Section 1, with the assumption that the distribution of skills in the population is given by $G(z) = \frac{z-(1-\Delta)}{\Delta} \mathbb{1}(1-\Delta, 1)(z) + \mathbb{1}(1, +\infty)(z)$. As heterogeneity Δ becomes small, this assignment function $m_l(x_l)$ becomes flat for $x_l \sim z^L_{l+1}$. This is because to the limit, the measure of managers $dy = dm_l(x) = m_l'(x_l)dx$ corresponding to a given measure dx of workers becomes small, as the most skilled managers work with workers who almost never ask for their help, which economizes on their valuable time.

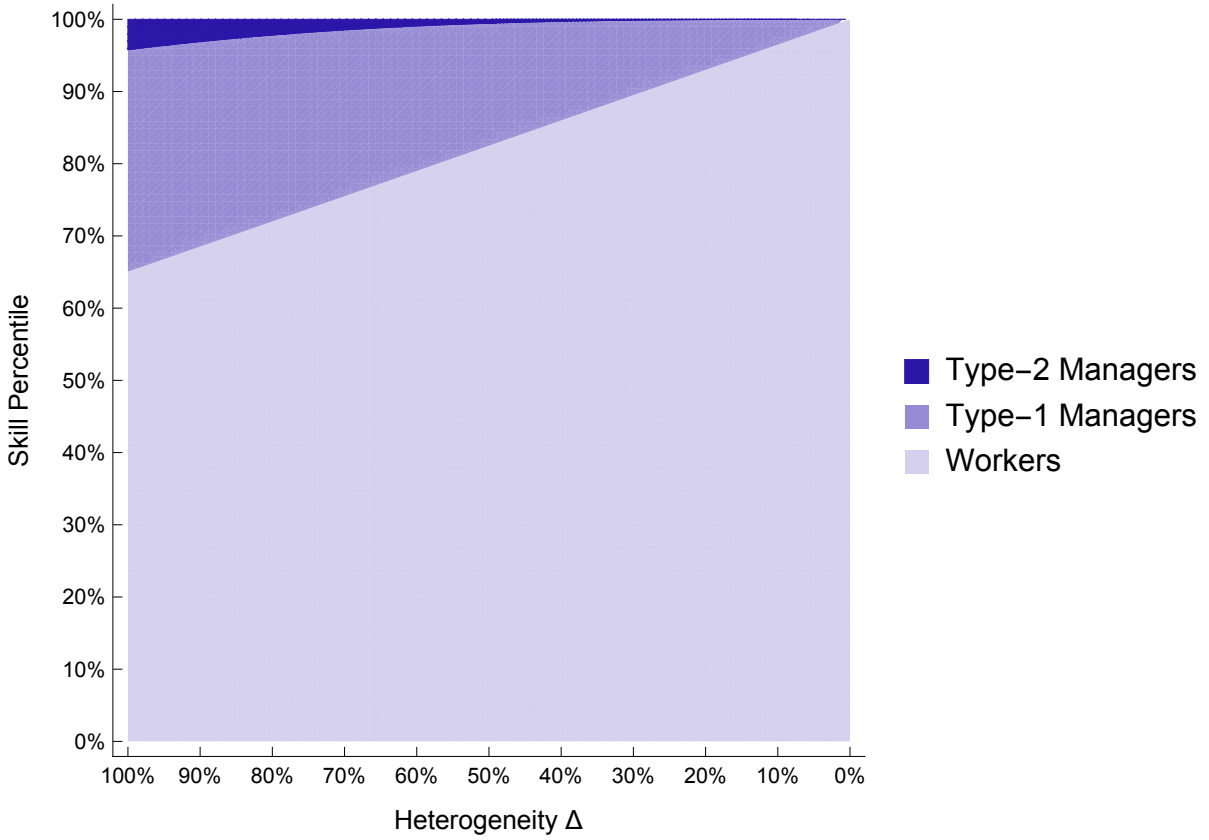
Lemma 3 (One Level, Uniform Case)**.** *Let the density of skills $g(.)$ be uniform with heterogeneity Δ. The span of control between adjacent levels of hierarchical organization is then given by a shifted Pareto distribution of coefficient equal to two:*

$$N^{l+1}_{l+1\to l}(x_{l+1}) = \frac{1}{h\sqrt{(1-z^L_{l+1})^2 + \frac{2}{h}(z^L_{l+2} - x_{l+1})}} \sim \text{Pareto}(2).$$

Pareto Approximation. When Δ or h become small, or when the number of layers increases, this becomes arbitrarily close to a Pareto distribution of tail coefficient equal to two, because in that case the cutoffs z^L_{l+1} and z^L_{l+2} approach one, and so the span of control distribution becomes arbitrarily close to a Pareto with a tail coefficient equal to 2, as for $x_{l+1} \in [z^L_{l+1}, z^L_{l+2}]$:

$$N^{l+1}_{l+1\to l}(x_{l+1}) \sim \frac{1}{\sqrt{2h}} \frac{1}{\sqrt{z^L_{l+2} - x_{l+1}}}.$$

Figure 7: Occupational Choices for a Uniform Distribution of Skills, and $h = 70\%$



Note: This figure gives the occupational choices of agents when $h = 70\%$, and the distribution of skills in the population is given by $G(z) = \frac{z-(1-\Delta)}{\Delta} \mathbb{1}(1-\Delta, 1)(z) + \mathbb{1}(1, +\infty)(z)$.

More precisely, it is a "shifted" Pareto in the sense that there is a maximum value for span of control, since the denominator does not quite attain zero even for extreme values of skills. This distribution is illustrated on Figure 8.

Because of the centrality of Pareto distributions in the paper, it is perhaps useful to remind the reader of the "Pareto principle".[10] As shown on Figure 8, the Pareto principle for span of control distributions states that the log of the probability that the number of workers supervised directly by a manager exceeds a given number of employees, is linear as a function of the log of this same number of employees. In the case of Figure 8, the relationship between the survivor function (or the countercumulative distribution) and the span of control is indeed linear on a log-log scale for the most part of the distribution. There are deviations for high values of span of

[10] Pigou (1920), describing the work of Pareto (1896), explains the Pareto principle very clearly, in the case of incomes: "Statistics of income in a number of countries, principally during the nineteenth century, are brought together. It is shown that, if x signify a given income and N the number of persons with incomes exceeding x, and if a curve be drawn, of which the ordinates are logarithms of x and the abscissae logarithms of N, this curve, for all the countries examined, is approximately a straight line, and is, furthermore, inclined to the vertical axis at an angle, which, in no country, differs by more than three or four degrees from 56°. This means (since tan(56°) = 1.5) that, if the number of incomes greater than x is equal to N, the number greater than mx is equal to (1/m 1.5)*N, whatever the value of m may be. Thus the scheme of income distribution is everywhere the same."

control, hence the term "shifted Pareto". On a log-log scale, the slope of the linear relationship if -2, which says that the shifted Pareto has a tail coefficient equal to 2. The deviations from Pareto become less severe when heterogeneity in skills goes to zero.

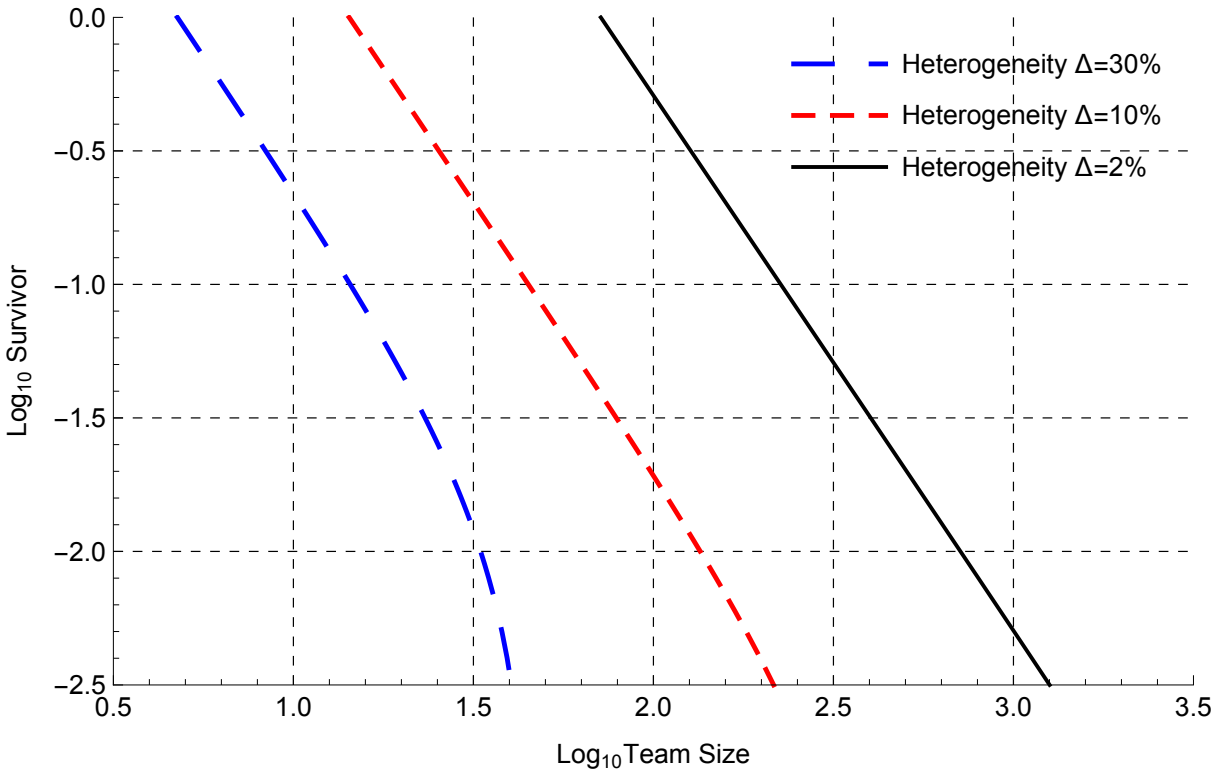
Note also that in that case, both z_{l+1}^L and z_{l+2}^L both go to one. The reader might at this stage worry that the exact Pareto results are very specific to having communication much more efficient than production (h low) or that the finding does not hold in economies with some non negligible amount of skill heterogeneity (Δ low). However, we shall see next that it is not the case. In fact, because when new layers of management are added, the size of the new layers endogenously become small, this in effect makes heterogeneity go to zero for the span of control of top managers over their direct subordinates, without any assumption on the total distribution of skills. Moreover, convergence is typically very fast. I come back to this point later.

Intuition. The intuition for why the span of control of top managers becomes very high, even more so when there is an arbitrarily small heterogeneity in skills in the economy, comes from the fact that all workers ask relatively few questions when heterogeneity goes to zero: they are then able to solve almost all problems by themselves. The fraction of managers required to answer these questions also goes to zero, in the limit, as can be seen on Figure 7. The marginal worker, who is just indifferent between being a worker and a manager, then is really able to solve almost all problems by himself, and the corresponding manager has a very high span of control. This reasoning actually also holds for relatively high values of heterogeneity: again as in Figure 7, adding new layers of management has the same effect as reducing heterogeneity in skills between the upper layer.

The fact that span of control is distributed like a shifted Pareto distribution in the upper tail comes from the mathematical reasoning above: the span of control is an inversely proportional function of the skill of the corresponding worker. This Pareto distribution has an endogenous tail coefficient equal to two, which can be seen on Figure 8, even for relatively high values of heterogeneity: on this Figure the slope of the linear part is -2 on a log-log scale when the survivor function is plotted against the value for span of control. Mathematically, this coefficient two comes from the shape of the matching function for the most skilled managers and workers, which becomes flat for those agents when heterogeneity goes to zero, as can be seen on Figure 6 above. The reason why this function is flat is that managers with high skills, close to one, are matched with workers who also have relatively high skills, and can in effect solve almost all problems by themselves. It is all the more true that heterogeneity is relatively low.

For non zero helping time or heterogeneity, the distribution of span of control we obtain could reasonably be called a "shifted Pareto": in particular, it has an upper bound (though it is very high for most values of the parameters) and does not quite have the Pareto property in the very upper tail. This fact can also be seen on Figure 6, the matching function is not exactly flat for non-zero heterogeneity. Note that this behavior is exactly what is observed in the data, where real economic variables are bounded, and the distribution of firms indeed has fewer firms in the

Figure 8: Negative Relationship between Heterogeneity in Agents' Productivities and Heterogeneity in Firm Sizes



Note: This figure gives the theoretical distribution of span of control down one level of hierarchical organization ("team size"), on a log-log-scale, with $h = 70\%$, for different values of skill heterogeneity Δ. The distribution of skills is taken as uniform. It is a "shifted" Pareto distribution with a coefficient equal to 2. As heterogeneity Δ goes to zero, the distribution approaches the full Pareto distribution in the upper tail. When the number of layers of hierarchical organization is endogenous, the number of agents in the last and before last layers endogenously go to zero, so that the full Pareto is obtained.

upper tail than the Pareto benchmark would suggest.[11]

Graphical, heuristic Intuition. A graphical, heuristic intuition is given in Section 2.4, through a simplification of the model. To understand the Pareto coefficient of coefficient 2, one needs to go through steps 1, 2, 3 of this graphical illustration (Figures 12, 13 and 14).

[11] There is an ongoing fierce debate in the literature about whether the distribution of firm sizes is Pareto or log-normal. The same debate rages about the distribution of city sizes (see Eeckhout (2004), Levy (2009)), as well as about that of trade flows (Head et al. (2014)). The model presented here generates a shifted Pareto only in the upper tail: in particular, the distribution of small firms depends on the distribution of skills; and the way in which the Pareto is shifted in the upper tail depends on the level of skill heterogeneity as well as on helping time.

2.3.2 Two Levels of Hierarchical Organization

Now that we have established that span of control down one level of hierarchical organization is a shifted Pareto of coefficient two in the uniform case, let us generalize this for the case of two layers of hierarchical organization. In particular, a case of interest is for the span of control of top managers, which pins down the equilibrium size distribution of firms, since each firm is run by a top manager:

$$N^L_{L\to L-1}(x_L) = \frac{1}{h\sqrt{(1-z_L^L)^2 + \frac{2}{h}(1-x_L)}} \sim \frac{1}{\sqrt{2h}}\frac{1}{\sqrt{1-x_L}}.$$

Of course, when $L > 1$, the agents below the top managers themselves manage other workers. Their span of control in turn is given by:

$$N^{L-1}_{L-1\to L-2}(x_{L-1}) = \frac{1}{h\sqrt{(1-z_{L-1}^{L-1})^2 + \frac{2}{h}(z_L^{L-1} - x_{L-1})}}$$

Expressed as a function of managers' skills, this span of control function is given by:

$$\begin{aligned} N^L_{L-1\to L-2}(x_L) &= N^{L-1}_{L-1\to L-2}\left(m_L^{-1}(x_L)\right) \\ N^L_{L-1\to L-2}(x_L) &= \frac{1}{h\sqrt{(1-z_{L-1}^L)^2 - \frac{2}{h}(1-z_L^L) + \frac{2}{h}\sqrt{(1-z_L^L)^2 + \frac{2}{h}(1-x_L)}}}. \end{aligned}$$

Similarly, this is becomes arbitrarily close to a Pareto distribution of tail coefficient equal to four, when Δ or h become small, or when the number of layers increases, as for $x_L \in [z_L^L, 1]$:

$$N^L_{L-1\to L-2}(x_L) \sim \frac{1}{\sqrt[4]{8h}}\frac{1}{\sqrt[4]{1-x_L}}.$$

Importantly, the total span of control of an upper level managers over workers down two levels of hierarchical organization is given by the product of the two spans of control, since agents of level $L-1$ themselves manage agents of level of $L-2$. The total span of control of a top level manager down two levels of hierarchical organization is therefore given by:

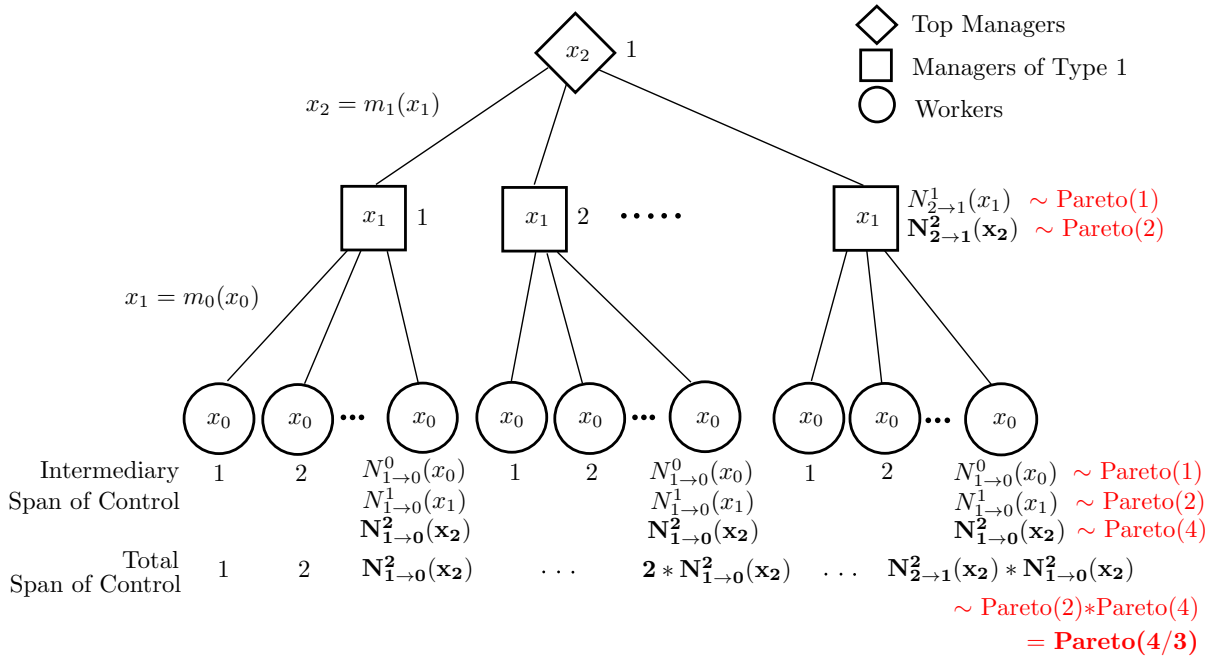
$$\begin{aligned} N^L_{L\to L-2}(x_L) &= N^L_{L\to L-1}(x_L) N^L_{L-1\to L-2}(x_L) \\ N^L_{L\to L-2}(x_L) &= \frac{1}{h^2\sqrt{(1-z_L^L)^2 + \frac{2}{h}(1-x_L)}} \frac{1}{\sqrt{(1-z_{L-1}^L)^2 - \frac{2}{h}(1-z_L^L) + \frac{2}{h}\sqrt{(1-z_L^L)^2 + \frac{2}{h}(1-x_L)}}}. \end{aligned}$$

Again, one can approximate this as Δ or h become small, or when the number of layers increases by:

$$N^L_{L\to L-2}(x_L) \sim \frac{1}{\sqrt{2h}}\frac{1}{\sqrt{1-x_L}}\frac{1}{\sqrt[4]{8h}}\frac{1}{\sqrt[4]{1-x_L}} \sim \frac{1}{\sqrt{2h}\sqrt[4]{8h}}\frac{1}{\sqrt[3/4]{1-x_L}}.$$

This is a Pareto distribution with coefficient equal to $4/3$. The steps of the reasoning for $L = 2$ levels of Hierarchical Organization are illustrated on Figure 9, where the firm has only two levels of hierarchical organization, in total. The total span of control of the top manager in that case is given by the multiplication of his span of control of managers of type 1 and of each one of these

Figure 9: Span of Control: Pareto Coefficients, $L = 2$ Layers



Note: This chart complements Figure 5 in giving an intuition for the formula $2^L/(2^L - 1)$ in the case $L = 2$. The Pareto distribution for total span of control has a coefficient equal to $4/3 = 2^2/(2^2 - 1)$, because of the multiplication of a Pareto with coefficient 2, that coming from the span of control of intermediary managers $N^2_{2\to1}$, and one with coefficient 4, that originating from the span of control of intermediary managers on workers, in the space of top managers' skills $N^2_{1\to0}$.

managers of type 1 on workers. The first distribution of span of control is given by a Pareto of coefficient two, as was explained before, and the second one if given by a Pareto of coefficient four.

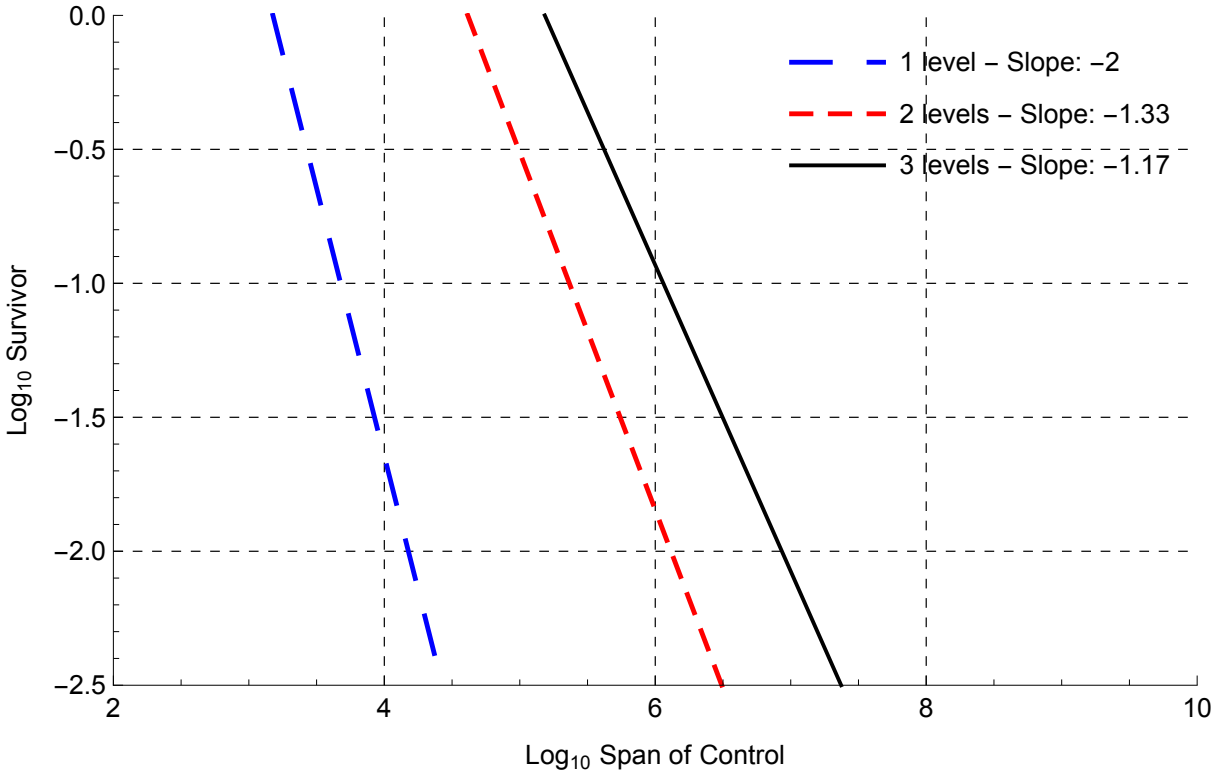
Graphical, heuristic Intuition. Again, a graphical, heuristic intuition is given in Section 2.4, through a simplification of the model. To understand the Pareto coefficient of coefficient 4/3, from that of coefficient 2, one needs to go through steps 4 and 5 of this graphical illustration (Figures 15 and 16).

2.3.3 Multiple Levels of Hierarchical Organization

More generally, what we have shown in the case $L = 2$ generalizes to any L levels of hierarchical organization. By recursion, the span of control distribution down one level of hierarchical organization, seen in the space of managers' of type L's skills, of $N^L_{n+1\to n}(x_L)$ is a shifted Pareto distribution of coefficient equal to 2^{L-n}, and allows to state the following lemma.

Lemma 4 (One Level, Different Space, Uniform Case)**.** *Let the density of skills $g(.)$ be uniform with heterogeneity Δ. The span of control between adjacent levels of hierarchical organization in the space of higher managers' skills, denoted by $N^L_{n+1\to n}(x_L)$ is a shifted Pareto distribution*

Figure 10: Simulation, Full Model: Decomposing Zipf's Law into its Components



Note: This Figure represents the theoretical distribution of span of control for top managers over employees down one to three levels or hierarchical organization, with the assumption that the distribution of skills in the population is given by $G(z) = \frac{z-(1-\Delta)}{\Delta}\mathbb{1}(1-\Delta,1)(z) + \mathbb{1}(1,+\infty)(z)$. The graph is drawn on a log-log scale, with $\Delta = 3/10$, $h = 9/10$ and the maximum $L = 3$. The slopes on a log-log scale are -2, $-4/3$, and $-8/7$ respectively.

of coefficient equal to 2^{L-n}*:*

$$N^L_{n+1\to n}(x_L) \sim \text{Pareto}(2^{L-n}).$$

For conciseness, I do not express these shifted Paretos explicitly. For example, the case with $L = n+2$ was solved previously:

$$N^L_{L-1\to L-2}(x_L) = \frac{1}{h\sqrt{(1-z^L_{L-1})^2 - \frac{2}{h}(1-z^L_L) + \frac{2}{h}\sqrt{(1-z^L_L)^2 + \frac{2}{h}(1-x_L)}}}.$$

The shifted Paretos are given explicitly by the formula:

$$N^L_{n+1\to n}(x_L) = \frac{1}{h\left(1 - m_n^{-1} \circ ... \circ m_{L-1}^{-1}(x_L)\right)} = \frac{1}{h\left[1 - \left(\bigotimes_{l=n}^{L-1} m_l^{-1}\right)(x_L)\right]},$$

where each one of the assignment functions are given as a function of the endogenous cutoffs

$\{z_l^L\}_{l=1}^L$ in Proposition 1, namely:

$$\forall l \in \{0, ..., L-1\}, \quad \forall x_{l+1} \in [z_{l+1}^L, z_{l+2}^L], \quad m_l^{-1}(x_{l+1}) = 1 - \sqrt{(1 - z_{l+1}^L)^2 - \frac{2}{h}(z_{l+2}^L - x_{l+1})}.$$

Proof. The proof for the tail coefficient of the Pareto proceeds by induction. The case for $L = 1$ was shown in Lemma 3. Assume that Lemma 4 is true for some $L-1$ with $L \geq 2$, let us show it is true for L. Noting that:

$$N_{n+1\to n}^L(x_L) = \frac{1}{h\left(1 - m_n^{-1} \circ ... \circ m_{L-1}^{-1}(x_L)\right)} = \frac{1}{h\left[1 - \left(\bigotimes_{l=n}^{L-1} m_l^{-1}\right)(x_L)\right]}$$
$$\sim \frac{1}{h\sqrt{1 - \bigotimes_{l=n+1}^{L-1}(m_l^{-1})(x_L)}} \quad \text{from Lemma 3.}$$

By induction hypothesis:

$$\frac{1}{1 - \bigotimes_{l=n+1}^{L-1}(m_l^{-1})(x_L)} \sim \text{Pareto}(2^{L-n-1}),$$

where $\text{Pareto}(2^{L-n-1})$ denotes a shifted Pareto distribution with a tail coefficient equal to 2^{L-n-1}. Then:

$$N_{n+1\to n}^L(x_L) \sim \text{Pareto}(2^{L-n}),$$

which proves the proposition for L and concludes the proof by induction. □

From Lemma 4, we are able to state the main result of the paper in the case of a uniform distribution in Lemma 5.

Lemma 5 (L Levels, Uniform Case)**.** *Let the density of skills $g(.)$ be uniform with heterogeneity Δ. Total span of control of managers of type n on levels of hierarchical organization down L levels, denoted by $N_{n\to n-L}^n(x_n)$ is a shifted Pareto distribution of coefficient equal to $2^L/(2^L-1)$:*

$$N_{n+L\to n}^{n+L}(x_{n+L}) \sim \text{Pareto}\left(\frac{2^L}{2^L - 1}\right).$$

Proof. This proposition, which is one of the main results of the paper (in the general case though), is just a corollary of the previous lemma since:

$$N_{L\to 0}^L(x_L) = N_{L\to L-1}^L(x_L) * N_{L-1\to L-2}^L(x_L) * ... * N_{1\to 0}^L(x_L) = \prod_{l=0}^{L-1} N_{L-l\to L-l-1}^L(x_L).$$

Applying the previous Lemma 4 to each one of the terms in the product $N_{L-l\to L-l-1}^L(.)$, which are therefore distributed according to shifted Paretos with a coefficient $2^{L-(L-l-1)} = 2^{l+1}$, allows to conclude by noting that the Pareto coefficient α_L is solution of:

$$\frac{1}{\alpha_L} = \sum_{l=0}^{L-1} \frac{1}{2^{l+1}} = \frac{1}{2}\frac{1 - \frac{1}{2^L}}{1 - \frac{1}{2}} = \frac{2^L - 1}{2^L} \quad \Rightarrow \quad \alpha_L = \frac{2^L}{2^L - 1}.$$

□

Let us denote by α_L the tail coefficient on the Pareto distribution for span of control given as $\alpha_L = 2^L/(2^L - 1)$. The following Table gives the values of α_L for $L = 1, 2, 3, 4, 5, ..., +\infty$.

L	1	2	3	4	5	...	$+\infty$
α_L (exact)	2	4/3	8/7	16/15	32/31	...	1
α_L (approx.)	2.00	1.33	1.17	1.07	1.03	...	1.00

These different coefficients are illustrated on Figure 10. In particular, the theoretical coefficients are close to those observed in the data (see Figure 1), and converge to 1 as the number of layers increases. Of course, the uniform distribution of skills is a very particular one. However I show in the next section that what was shown here is more generally true in the upper tail of the span of control distributions for any smooth distribution function, provided its density is continuous ad bounded away from zero for higher skills. This is an arguably rather limited set of assumptions.

Graphical, heuristic Intuition. Again, a graphical, heuristic intuition is given in the next Section, through a simplification of the model. To understand the Pareto coefficient of coefficient $2^L/(2^L - 1)$, from that of coefficient 2, one needs to go through steps 4, 5 and 6 of this graphical illustration (Figures 15, 16 and 17).

2.4 Intuition for the Pareto Distributions

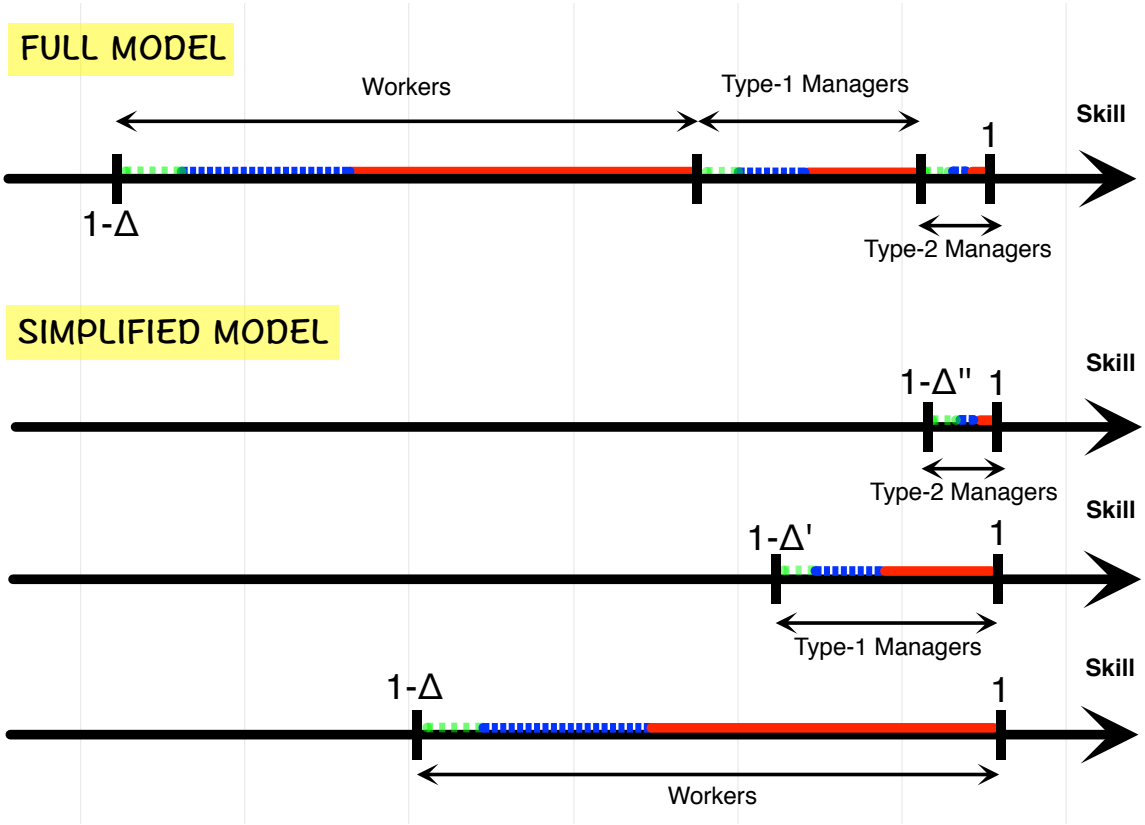
In this section, I give some heuristics for the Pareto results considering a simplified model, which is not exactly the one that is studied in this paper but for which the intuition is easier to convey. The idea is to capture the essence of the full model through simpler relations. The distribution of skills is assumed to be uniform, and some simplifying assumptions are made so that both the graphical intuitions and the mathematics are made more transparent.

I consider a case where occupations are not endogenous, but exogenous, as shown on Figure 11. For example, when $L = 2$, there are only managers and workers, and the matching is two sided ex-ante, with managers' skills distributed uniformly in $[1 - \Delta', 1]$, and that workers' skills are distributed uniformly in $[1 - \Delta, 1]$. In short, this amounts to making the occupational choice cutoffs $\{z_l^L\}_{l=1}^L$ exogenous. As shown below, the mathematics are then much simpler.

However only the proofs above and below show that the exercise in simplification I engage in here is valid. In particular, note that everything I show here turns out to be true for *any bounded from* 0 *density function of skills* at the top, not just for the uniform distribution (as shown in Section 2.5). Intuitively, this is because a very small interval of skills underlies the very high values for span of control, and to a first approximation, any bounded away from zero density of skill can be approximated by a uniform distribution.

Step 1/6 (Figure 12). The number of problems that workers of skill x cannot solve is given by $1 - x$, it takes h units of time to communicate a solution to a worker, and each manager has

Figure 11: Example: Full Model, Simplified Model



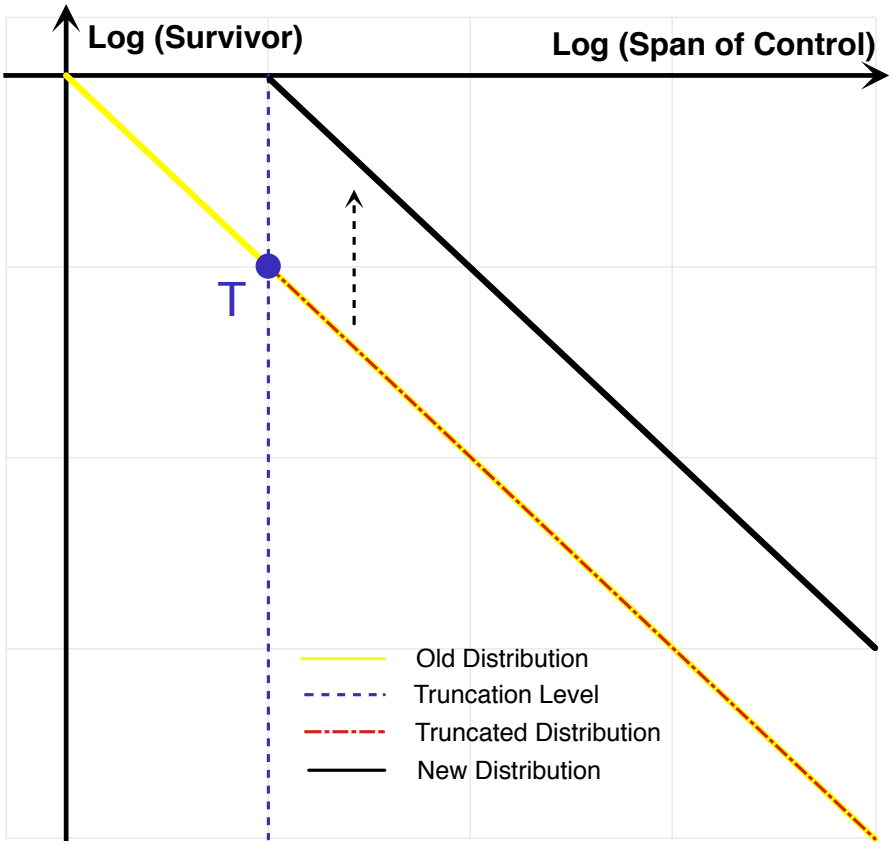
one unit of time. The size of a firm a worker works in is inversely proportional to how many problems he cannot solve, and to the communication cost h. In mathematical terms, this size $N(x)$ of a firm which employs workers of skill x is given by:

$$N(x) = \frac{1}{h(1-x)}.$$

Therefore, the firm size distribution "in workers' space" (that is, when the same firm is counted three times if it employs three workers), is the inverse of a uniform distribution over $[0, \Delta]$, up to the h constant. This distribution is clearly a Pareto distribution of coefficient one: a truncation of the distribution at some cutoff, say $\Delta_c < \Delta$ (or skill $1 - \Delta_c > 1 - \Delta$), displays the exact same span of control distribution in workers' space, where all span of controls are multiplied by Δ/Δ_c. One can then plot the new distribution of spans of control, in this subsample, which is given by the "New Distribution" on Figure 12 below. One could get the false impression that this is where Zipf's law comes from, but it is not at all the case: this does not at all correspond to a firm size distribution, as the next step shows.

Step 2/6 (Figure 13). The relevant measure is not the firm size distribution in workers' space, but the firm size distribution as measured in the data. We are therefore still a few steps away from understanding Zipf's law. Again, the firm size distribution in workers' space means

Figure 12: Scale Independence, Pareto (1) in the space of workers' skills



that the same firm is counted 3 times if it employs 3 workers. One therefore needs to investigate the mapping from workers to firms, or equivalently the mapping from workers to managers (one manager runs one firm).

Luckily, in the hypothetical economy described above, this mapping also is a scale independent function. More precisely, the function mapping the number of problems that workers do not know how to solve to the equivalent quantity for managers is an exact square function.

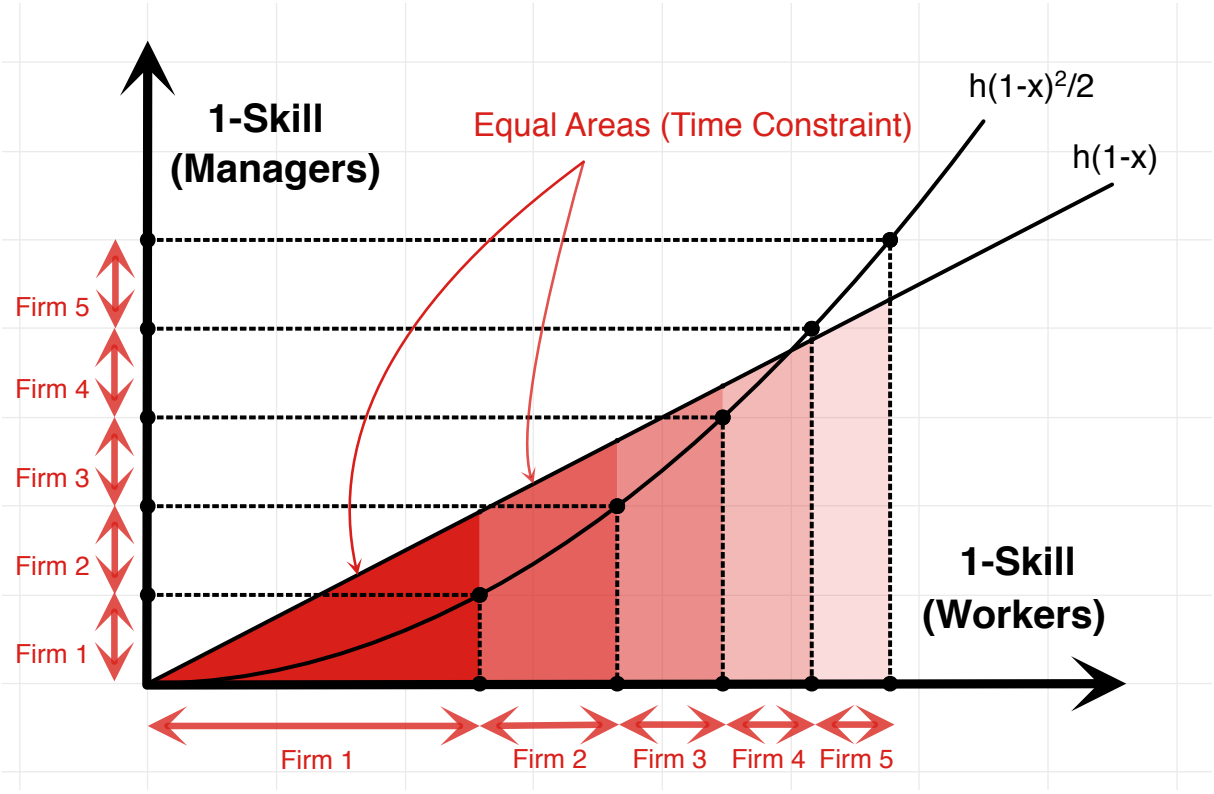
The intuition is given on Figure 13 and comes from the integration of a linear function. The number of problems that workers of skills x cannot solve is given by $1-x$. The first manager therefore employs more workers, than the second manager who has the same time constraint. In other words, the integration of a scale independent function is another scale independent function with an exponent one unit higher. Mathematically, denoting by $y(x)$ the Skill of managers working with workers of skill x:

$$d(1-y) = h(1-x)d(1-x) \quad \Rightarrow \quad 1-y(x) = h\frac{(1-x)^2}{2} \quad \Rightarrow \quad 1-x = \sqrt{\frac{2}{h}}\sqrt{1-y}$$

The matching function therefore is a square function (a power function with a coefficient 2).

Step 3/6 (Figure 14). From Step 1 and Step 2, one can infer that the firm size distribution

Figure 13: Scale-independent Matching Function, h=75%



is also scale independent because it is a transformation of the two above scale independent distributions, and that the corresponding Pareto coefficient is 2.

The reason why the Pareto tail is less thick for the firm size distribution as measured in the data than that as a function of workers' skills is illustrated in Figure 14: there are many more workers in large firms by definition, so large firms are disproportionately more represented when the firm size distribution is expressed in the "space of workers". In the bottom-right corner, one sees the firm size distribution expressed in the "space of workers", which is a Pareto distribution with a coefficient equal to 1; in the upper-right corner, one sees this firm size distribution with a much thinner tail when expressed in the relevant skill space, that is managers' space.

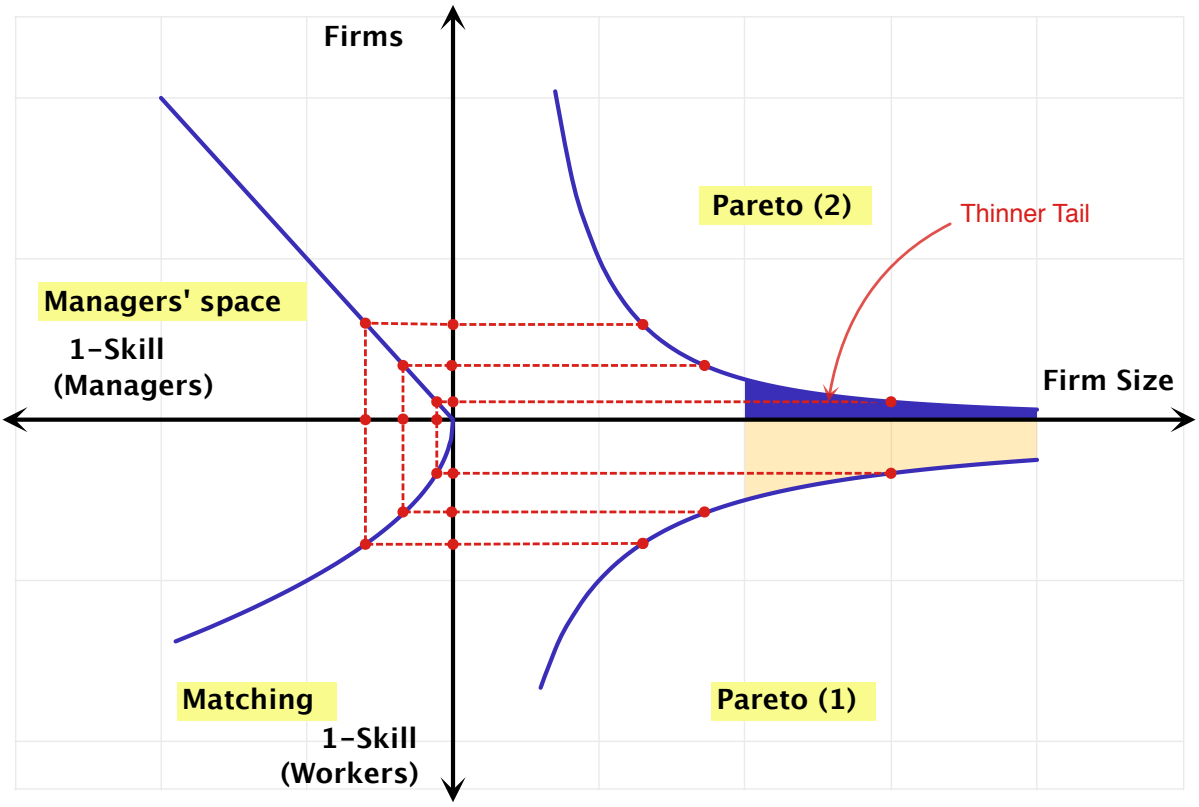
Another way to see this transparently from the mathematics, which are much simpler in this case than in Section 2, is to express the firm size as a function of managers' skills, which are distributed uniformly (with a slight abuse of notation, denoting again by N the number of firms):

$$N(y) = \frac{1}{1-x} = \sqrt{\frac{h}{2}}\frac{1}{\sqrt{1-y}}.$$

The firm size distribution therefore is in that case, exactly a Pareto distribution with a coefficient equal to 2.

Step 4/6 (Figure 15). It is also useful to write the distribution of span of control "in the space of top managers' skills": that is, to count the distribution of span of control as if the unit

Figure 14: From Pareto (1) to Pareto (2)



of observation was not the manager of that team, but the manager of the team above him. One can then see why that amounts to doing the same operation as was done in Step 3, and this leads to an even thinner tail than previously. Figure 15 illustrates the compounding. It shows that the tail of the span of control distribution is made thinner by both matching functions, from the lower-right corner, to the upper-right corner.

Again, another way to see it transparently is through the mathematics which are much simpler than in Section 2. Denoting by z the type of the top manager, the mapping $z(y)$ from intermediary managers to top managers is given by:

$$d(1-z) = h(1-y)d(1-y) \quad \Rightarrow \quad 1 - z(y) = h\frac{(1-y)^2}{2}.$$

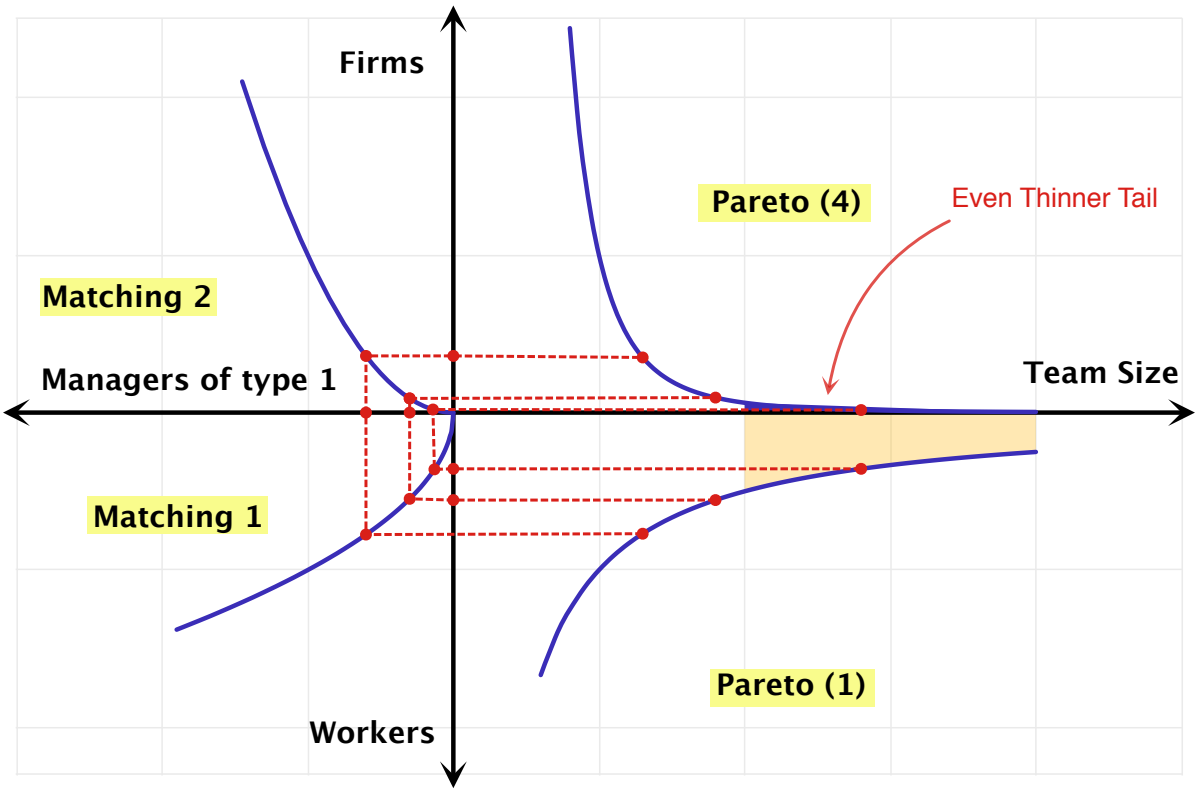
The span of control of intermediary managers over workers is therefore given in the space of managers' skills, by the following distribution:

$$N(z) = \frac{1}{1-x} = \sqrt{\frac{h}{2}}\sqrt[4]{\frac{h}{2}}\frac{1}{\sqrt[4]{1-z}}.$$

This is exactly a Pareto distribution with a coefficient equal to 4.

Step 5/6 (Figure 16). The next step is to multiply the span of control distributions obtained in Steps 3 and 4, to obtain the total span of control of managers down two levels of hierarchical

Figure 15: From Pareto (1) to Pareto (4)



organization. To do that simply, it is now useful to think of the previous Figure 14 and Figure 15 on a log-log scale. The Pareto distribution with a coefficient 2 now appears as a line with slope 1/2 on a Log (Survivor) and Log (Span of Control) scale. Similarly, a Pareto distribution with coefficient 4 now appears as a line with slope 1/4. The total span of control distribution then results from the multiplication of these two spans of control: the one of a top manager over intermediary managers, and that of intermediary managers over workers. Taking logs, these corresponds to adding spans of control, so to add the two curves along the x-axis. Straightforward algebra then says that the slope of this new curve α is solution of:

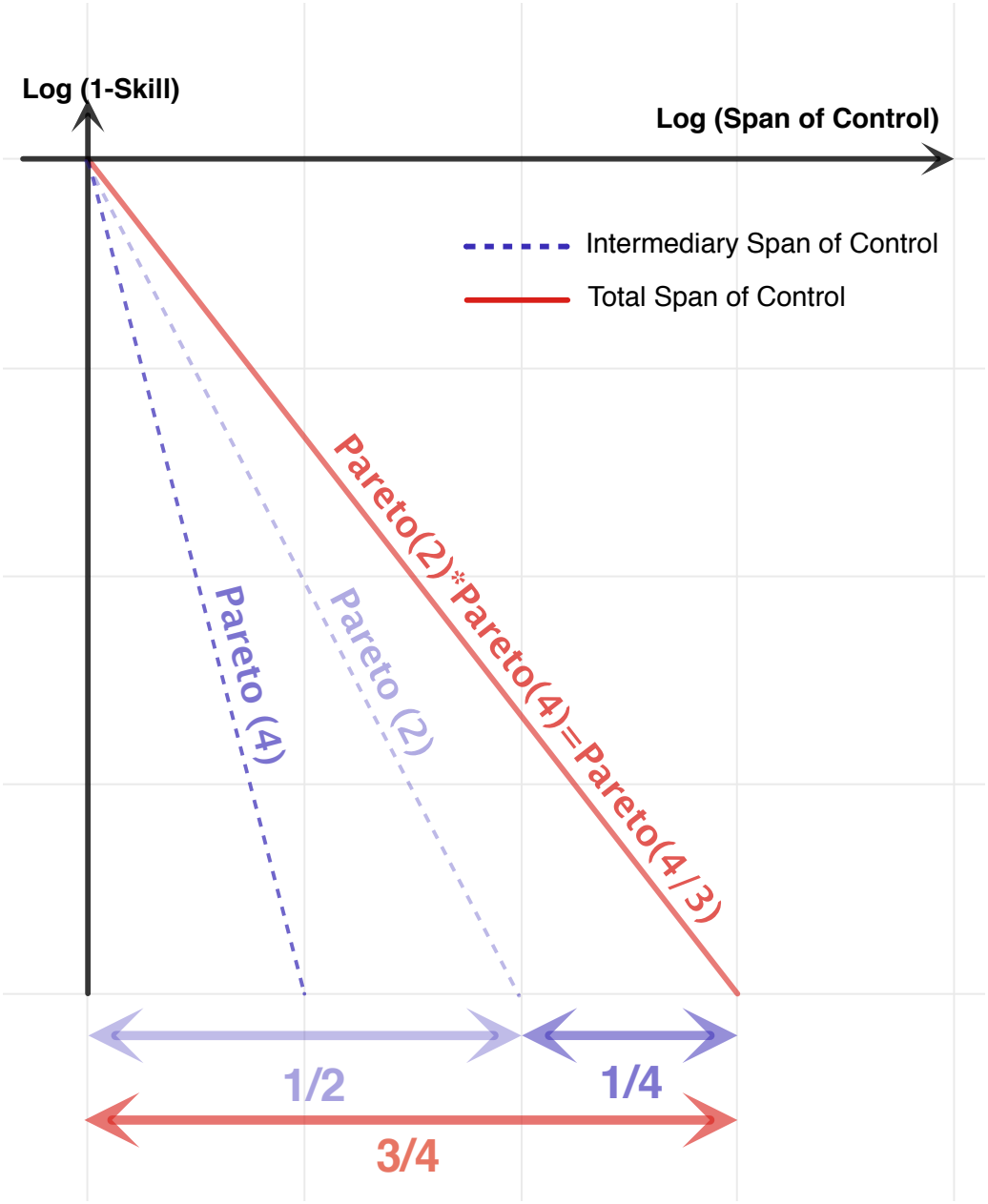
$$\frac{1}{\alpha} = \frac{1}{2} + \frac{1}{4} \quad \Rightarrow \quad \alpha = \frac{4}{3}.$$

Step 6/6 (Figure 17). The last step relies on recursion of the argument in step 5. Just as coefficient 4/3 is obtained by summing the two linear functions on a log-log scale along the x axis (to multiply span of control), coefficient 1 to the limit is obtained by summing an infinite number of linear functions on a log-log scale along the x axis. The fact that it converges straightforwardly follows from the well-known geometric sum:

$$\sum_{l=0}^{\infty} \frac{1}{2^{l+1}} = 1.$$

Note how this algebraic identity actually has a geometric counterpart on the graph, as it can be seen very simply why 1 is indeed equal to 1/2 plus 1/4, etc.

Figure 16: Multiplying a Pareto (2) and a Pareto (4) to get a Pareto (4/3)

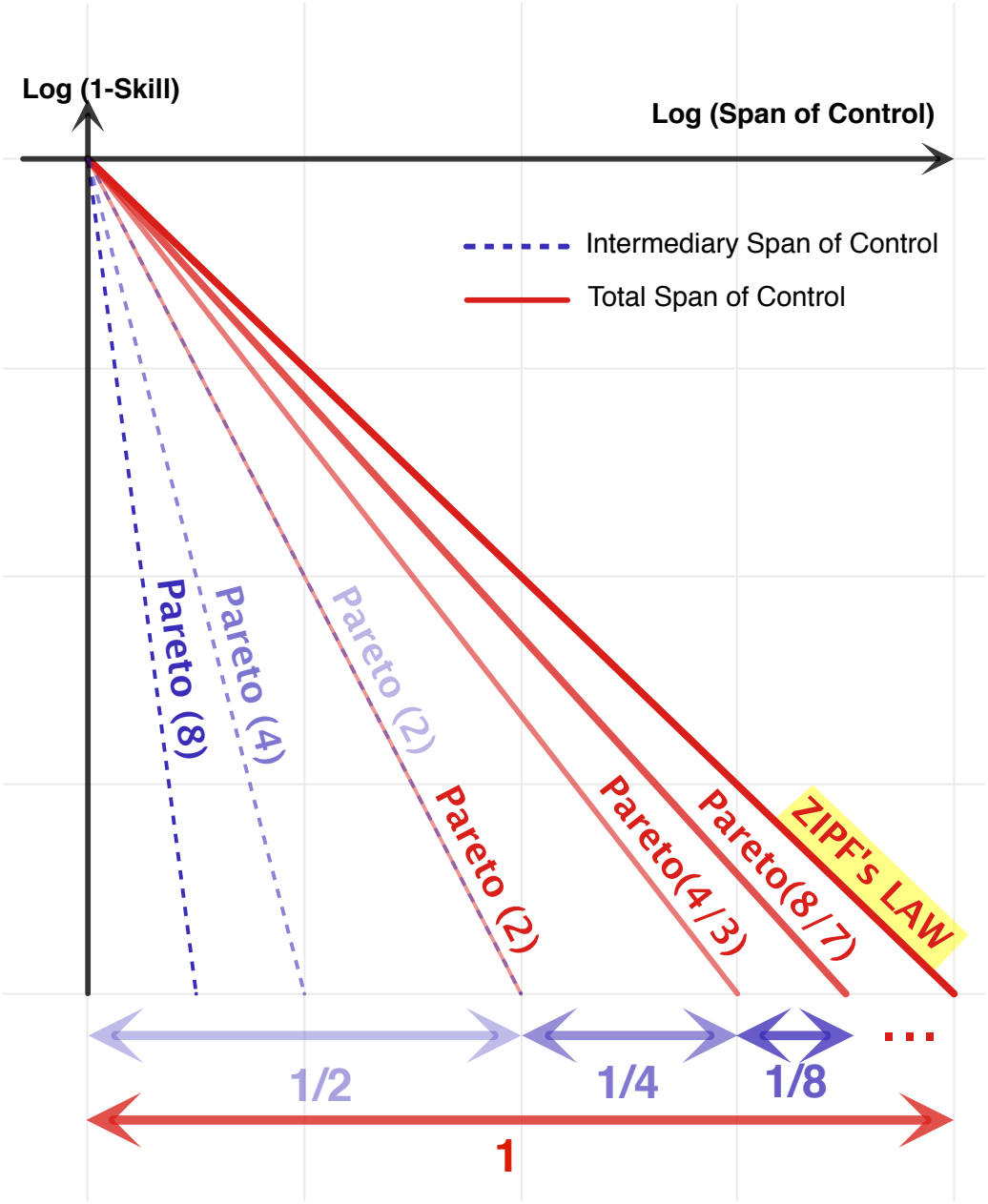


2.5 Skill Density Bounded away from Zero at the Top

I next show that no matter what the distribution of skills in the population is, provided that it is bounded away from zero for high skills, and continuous, the Pareto distributions for span of control obtain in the upper tail, exactly in the same way as in the uniform case. Let me first formalize these assumptions in Assumption 3.

Assumption 3 (Density Function of Skills)**.** *The density function of skills $g(.)$ is continuous*

Figure 17: Multiplying Paretos (2^n) to get Zipf's Law



on $[1-\Delta, 1]$*, and bounded away from zero near one:*

$$\exists m > 0, \quad \exists \eta > 0, \quad \forall x \in [1-\eta, 1], \quad g(x) \geq m > 0.$$

The fact that the conclusions on Pareto distributions, and in particular on their tail coefficients, do not depend on the underlying distribution of skills (apart from them satisfying Assumption 3) may seem as a perhaps surprising result at first sight. The intuition is that as the number of layers increases (or heterogeneity goes to zero), the distribution of skills which matters is drawn for a smaller segment of the total skill distribution. To the limit, any function can be

approximated by a uniform distribution under minimal regularity assumptions. In particular, the very large heterogeneity underlying the upper tail of the span of control distribution actually hinges on the behavior of a relatively small segment of managers' and workers' skill distribution. Proposition 3 only generalizes Lemma 3 in the case of a uniform distribution to any distribution of skills satisfying Assumption 3.

Proposition 3 (One Level, General Case)**.** *The span of control of a manager of type $l+1$ over employees down one level of hierarchical organization $N^{l+1}_{l+1\to l}(.)$ is arbitrarily close in the upper tail to the shifted Pareto distribution of coefficient two obtained in the uniform case:*

$$\forall \epsilon > 0, \quad \exists \eta > 0, \quad \forall x_{l+1} \in [z^L_{l+2} - \eta, z^L_{l+2}],$$
$$\frac{1}{h\sqrt{(1-z^L_{l+1})^2 + \frac{2}{Ah(1-\epsilon)}(z^L_{l+2} - x_{l+1})}} \leq N^{l+1}_{l+1\to l}(x_{l+1}) \leq \frac{1}{h\sqrt{(1-z^L_{l+1})^2 + \frac{2}{Ah(1+\epsilon)}(z^L_{l+2} - x_{l+1})}}.$$

This is denoted as:

$$N^{l+1}_{l+1\to l}(x_{l+1}) \simeq \text{Pareto}(2).$$

The shifted Pareto on the left-hand side is a lower bound in the upper tail:

$$\frac{1}{h\sqrt{(1-z^L_{l+1})^2 + \frac{2}{Ah(1-\epsilon)}(z^L_{l+2} - x_{l+1})}} \sim \frac{\sqrt{A(1-\epsilon)}}{\sqrt{2h}} \frac{1}{\sqrt{z^L_{l+2} - x_{l+1}}}.$$

Symmetrically, the shifted Pareto on the right-hand side is an upper bound:

$$\frac{1}{h\sqrt{(1-z^L_{l+1})^2 + \frac{2}{Ah(1+\epsilon)}(z^L_{l+2} - x_{l+1})}} \sim \frac{\sqrt{A(1+\epsilon)}}{\sqrt{2h}} \frac{1}{\sqrt{z^L_{l+2} - x_{l+1}}}.$$

As ϵ goes to zero, these shifted Paretos become infinitesimally close to each other, so that $N^{l+1}_{l+1\to l}(x_{l+1})$ is also arbitrarily close a shifted Pareto in the upper tail, which I denote in the following as:

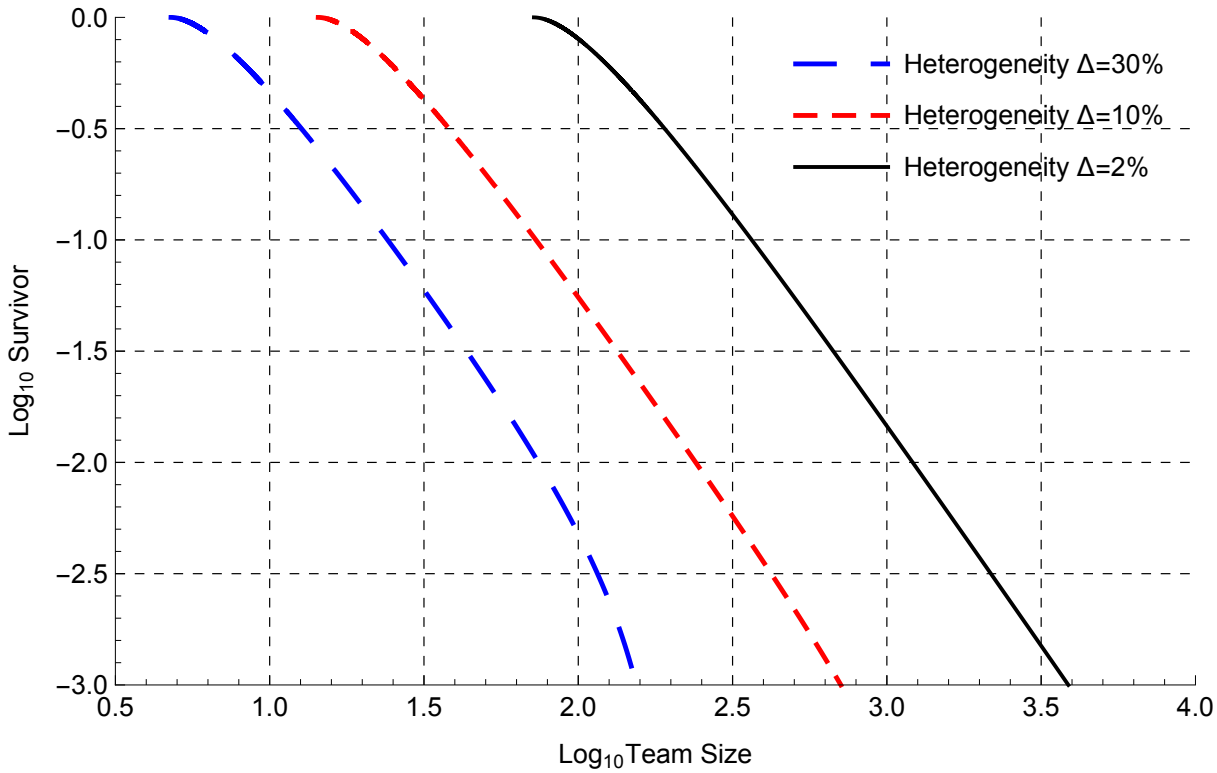
$$N^{l+1}_{l+1\to l}(x_{l+1}) \simeq \text{Pareto}(2).$$

As an illustration, Figure 18 plots the equivalent of Figure 8 in the case where the distribution of skills is given by an increasing function. One can see how the behavior in the lower tail of the distribution differs a bit from the uniform case, but that the upper tail behavior, and in particular the slope -2 for the survivor function on a log-log scale, remains unchanged. (see Appendix D, for more Figures in the increasing case)

Proof. From the market clearing equation for skills:

$$m'_l(x_l) = h(1-x_l)\frac{g(x_l)}{g(m_l(x_l))}$$

Figure 18: Independence of the Upper tail Behavior from the Underlying Skill Distribution



Note: This figure gives the theoretical distribution of team sizes, with the assumption that the distribution of skills in the population is given by $G(z) = \frac{(z-(1-\Delta))^2}{\Delta^2}\mathbb{1}[1-\Delta, 1](z) + \mathbb{1}[1, +\infty](z)$, instead of the uniform case. One can compare to Figure 8 for an equivalent in the uniform case: still the graph is on a log-log scale, down One Level of Hierarchical Organization ("Teams") with always $h = 70\%$. Note that the lower tail behavior differs a bit with this increasing distribution. (as in the data, see Figure 30) However, similarly as in the uniform case, as heterogeneity Δ goes to zero, the distribution approaches the full Pareto distribution with coefficient two in the upper tail.

Denote by A the limit of $\frac{g(x_l)}{g(m_l(x_l))}$ when $x_l \to z_{l+1}^L$. Fix $\epsilon > 0$ as small as one wants. Then there exists η such that for all $x_l \in [z_{l+1}^L - \eta, z_{l+1}^L]$, we have:

$$Ah(1-x_l)(1-\epsilon) \leq m_l'(x_l) \leq Ah(1-x_l)(1+\epsilon)$$

Integrating between $[x_l, z^L_{l+1}]$, this gives:

$$Ah(1-\epsilon)\left(z^L_{l+1} - \frac{(z^L_{l+1})^2}{2} - x_l + \frac{x_l^2}{2}\right) \leq z^L_{l+2} - m_l(x_l) \leq Ah(1+\epsilon)\left(z^L_{l+1} - \frac{(z^L_{l+1})^2}{2} - x_l + \frac{x_l^2}{2}\right)$$

$$\Rightarrow \quad 2\frac{z^L_{l+2} - m_l(x_l)}{Ah(1+\epsilon)} - 2z^L_{l+1} + (z^L_{l+1})^2 \leq x_l^2 - 2x_l \leq 2\frac{z^L_{l+2} - m_l(x_l)}{Ah(1-\epsilon)} - 2z^L_{l+1} + (z^L_{l+1})^2$$

$$\Rightarrow \quad 2\frac{z^L_{l+2} - m_l(x_l)}{Ah(1+\epsilon)} + \left(1 - z^L_{l+1}\right)^2 \leq (1-x_l)^2 \leq 2\frac{z^L_{l+2} - m_l(x_l)}{Ah(1-\epsilon)} + \left(1 - z^L_{l+1}\right)^2$$

$$\Rightarrow \quad \sqrt{(1-z^L_{l+1})^2 + \frac{2}{Ah(1+\epsilon)}(z^L_{l+2} - x_{l+1})} \leq 1 - m_l^{-1}(x_{l+1}) \leq \sqrt{(1-z^L_{l+1})^2 + \frac{2}{Ah(1-\epsilon)}(z^L_{l+2} - x_{l+1})}$$

$$\Rightarrow \quad \frac{1}{h\sqrt{(1-z^L_{l+1})^2 + \frac{2}{Ah(1-\epsilon)}(z^L_{l+2} - x_{l+1})}} \leq N^{l+1}_{l+1\to l}(x_{l+1}) \leq \frac{1}{h\sqrt{(1-z^L_{l+1})^2 + \frac{2}{Ah(1+\epsilon)}(z^L_{l+2} - x_{l+1})}}.$$

□

Similarly, Lemma 6 generalizes Lemma 4 and Proposition 4 generalizes Lemma 5, to a case where Assumption 3 holds.

Lemma 6 (One Level, Different Space, General Case)**.** *The span of control between adjacent levels of hierarchical organization in the space of higher managers' skills, denoted by $N^L_{n+1\to n}(x_L)$ is arbitrarily close to the shifted Pareto distribution of coefficient equal to 2^{L-n} obtained in the uniform case:*

$$N^L_{n+1\to n}(x_L) \simeq \text{Pareto}(2^{L-n}).$$

Proposition 4 (L Levels, General Case)**.** *Total span of control of managers of type n on levels of hierarchical organization down L levels, denoted by $N^n_{n\to n-L}(x_n)$ is arbitrarily close to the shifted Pareto distribution of coefficient equal to $2^L/(2^L-1)$ obtained in the uniform case:*

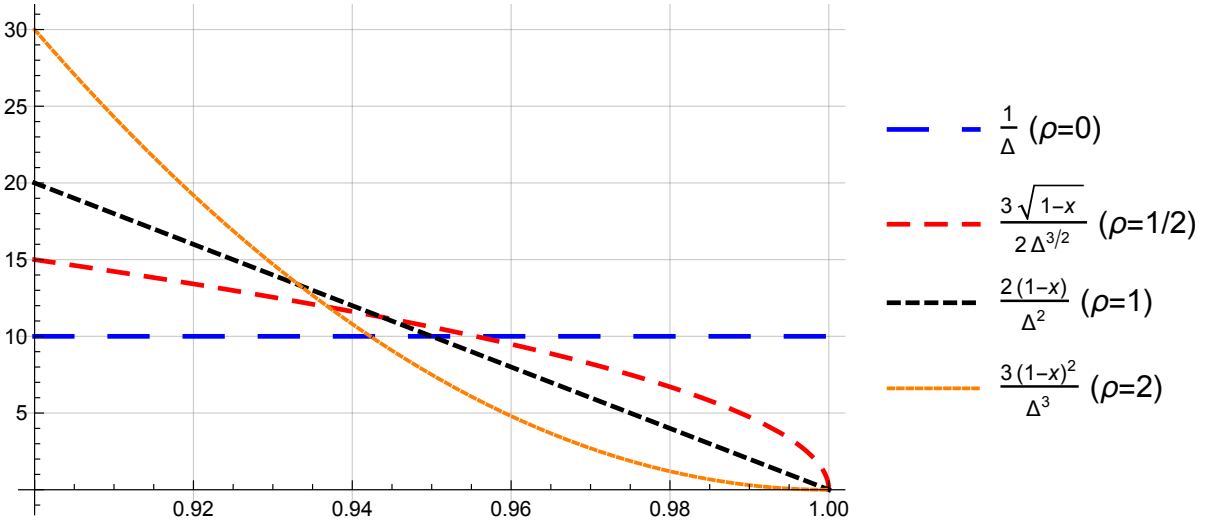
$$N^{n+L}_{n+L\to n}(x_{n+L}) \simeq \text{Pareto}\left(\frac{2^L}{2^L-1}\right).$$

Proof. These two Propositions result directly from Proposition 3 in the same way as Lemmas 4 and 5 result from Lemma 3. The proof for the tail coefficient of the approximating shifted Pareto proceeds by induction, using the case for $L = 1$ as a starting statement, assuming that the statement is true for some $L-1$ with $L \geq 2$ and showing it is true for L. Similarly, the proof of Proposition 4 follows from multiplying individual span of control distributions together, which gives the same formula for the total span of control distributions' tail coefficient $\alpha_L = 2^L/(2^L-1)$. Refer to the proofs of Lemmas 4 and 5 for details, replacing $\sim$ by $\simeq$ equivalence relations. □

2.6 Skill Density Non Bounded away from Zero at the Top

Let us drop now Assumption 3, and look at what happens when $g(1) = 0$ instead. In that case, it is useful to approximate the density function in the neighborhood of 1 as follows:

$$g(x) = A(1-x)^{\rho} + O\left((1-x)^{\rho+1}\right).$$



Note: The polynomial density functions for skills, for which closed form solutions are obtained, are drawn for $\rho = 0$ (uniform density function), $\rho = 1/2$, $\rho = 1$ and $\rho = 2$ respectively, with $\Delta = 10\%$. ρ governs the behavior of the density function for high skills. Only $\rho = 0$ corresponds to Assumption 3.

As before, the only relevant thing for the shape of the span of control distribution in the upper tail is coefficient ρ. Symmetrically to the benchmark of a uniform distribution the case of boundedness from zero, it suffices to look at the following simple polynomial functions:

$$g(x) = \frac{\rho+1}{\Delta^{\rho+1}}(1-x)^{\rho}.$$

They integrate up to one since the corresponding cumulative distribution function is given by:

$$G(x) = 1 - \frac{(1-x)^{\rho+1}}{\Delta^{\rho+1}}.$$

In that case, the matching function is given through the differential equation:

$$\begin{aligned}
&\frac{\rho+1}{\Delta^{\rho+1}}\left(1-m(x)\right)^{\rho} m'(x) = h(1-x)\frac{\rho+1}{\Delta^{\rho+1}}(1-x)^{\rho} \\
\Rightarrow \quad &\left[\frac{(1-m(u))^{\rho+1}}{\rho+1}\right]_{x_l}^{z_{l+1}^L} = h\left[-\frac{(1-u)^{\rho+2}}{\rho+2}\right]_{x_l}^{z_{l+1}^L} \\
\Rightarrow \quad &(1-x_l)^{\rho+2} = \frac{1}{h}\frac{\rho+2}{\rho+1}\left[(1-m_l(x_l))^{\rho+1} - (1-z_{l+2}^L)^{\rho+1}\right] + (1-z_{l+1}^L)^{\rho+2} \\
\Rightarrow \quad &1 - m_l^{-1}(x_{l+1}) = \left[(1-z_{l+1}^L)^{\rho+2} - \frac{1}{h}\frac{2+\rho}{1+\rho}\left[(1-x_{l+1})^{\rho+1} - (1-z_{l+2}^L)^{\rho+1}\right]\right]^{\frac{1}{\rho+2}}.
\end{aligned}$$

The span of control distribution down 1 level of hierarchical organization is therefore given by:

$$N_{l+1\to l}^{l+1}(x_{l+1}) = \frac{1}{h\left[(1-z_{l+1}^L)^{\rho+2} - \frac{1}{h}\frac{2+\rho}{1+\rho}\left[(1-x_{l+1})^{\rho+1} - (1-z_{l+1}^L)^{\rho+1}\right]\right]^{\frac{1}{2+\rho}}} \sim Pareto\left(\frac{2+\rho}{1+\rho}\right).$$

Note that for $\rho = 0$, one obtains the formula derived earlier for a uniform distribution of skills. Symmetrically as before, one can generalize this to the case of L levels of hierarchical organizations, in which case the Pareto coefficient on $N_{L\to 0}^L(x_L)$, such that $N_{L\to 0}^L(x_L) \simeq (1-x_L)^{1/\alpha_L}$ is given by:

$$\frac{1}{\alpha_L} = \sum_{l=0}^{L-1}\left(\frac{1+\rho}{2+\rho}\right)^{l+1} = \frac{1+\rho}{2+\rho}\frac{1-\left(\frac{1+\rho}{2+\rho}\right)^L}{1-\frac{1+\rho}{2+\rho}} \quad \Rightarrow \quad \alpha_L = \frac{1}{1+\rho}\frac{(2+\rho)^L}{(2+\rho)^L-(1+\rho)^L}.$$

Note that the Pareto coefficient for the span of control distribution as measured in the data, given by $N_{L\to 0}^{top}(x) \simeq (1-x)^{1/\beta_L}$ in turn is given by:

$$\beta_L = \frac{(2+\rho)^L}{(2+\rho)^L-(1+\rho)^L}.$$

This is because:

$$N_{L\to 0}^{top}(x) = N_{L\to 0}^L\left(G^{-1}(x)\right) \simeq \left(1-G^{-1}(x)\right)^{1/\alpha_L} \simeq (1-x)^{\frac{1}{\alpha_L(\rho+1)}}.$$

The last equivalence comes from:

$$1-G(x) \sim \frac{(1-x)^{\rho+1}}{\Delta^{\rho+1}} \quad \Rightarrow \quad 1-G^{-1}(x) \sim \Delta(1-x)^{\frac{1}{\rho+1}}.$$

This allows to state Proposition 5 in the general case where Assumption 3 does not hold.

Proposition 5 (L Levels, General Case, $g(1) = 0$)**.** *Assume that in the neighborhood of* 1, $g(x) \sim_{x\to 1} (1-x)^\rho$ *for some* ρ*. Then total span of control of managers of type* n *on levels of hierarchical organization down* L *levels, in the space of top managers' skills, denoted by* $N_{n\to n-L}^n(x_n)$ *is arbitrarily close to the shifted Pareto distribution of coefficient equal to that obtained in the polynomial of degree* ρ *case:*

$$N_{n+L\to n}^{n+L}(x_{n+L}) \simeq \text{Pareto}\left(\frac{1}{1+\rho}\frac{(2+\rho)^L}{(2+\rho)^L-(1+\rho)^L}\right).$$

The span of control distribution of top managers as measured in the data then follows:

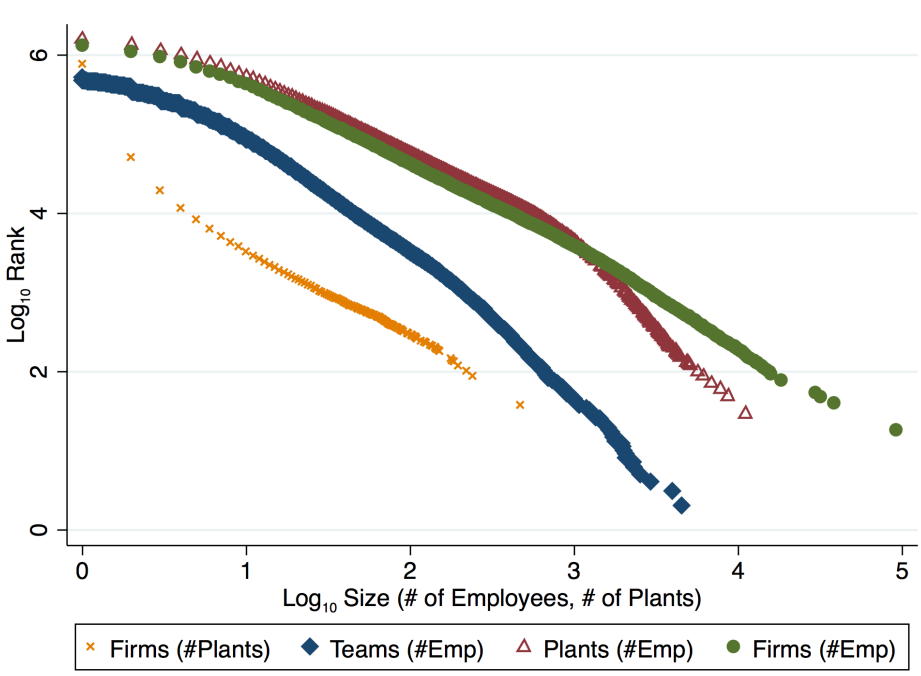
$$N_{n+L\to n}^{top}(x) \simeq \text{Pareto}\left(\frac{(2+\rho)^L}{(2+\rho)^L-(1+\rho)^L}\right).$$

2.7 Evidence

The model presented in this paper gives very precise and testable predictions about how Zipf's law arises in the data. In particular, one striking implication of the model is that not only does firm size distribution follow a Pareto distribution of coefficient one, but that intermediary span of control distributions also follow Pareto distributions, with higher tail coefficients.

In particular, it is shown that in theory, "team sizes", that is the span of control of CEOs or intermediary managers over one level of hierarchical organization, follow Pareto distributions with a coefficient equal to 2 in the upper tail. As discussed in the Introduction and shown in Figure 1, the French administrative matched employer-employee data gives a strong support to this proposition.

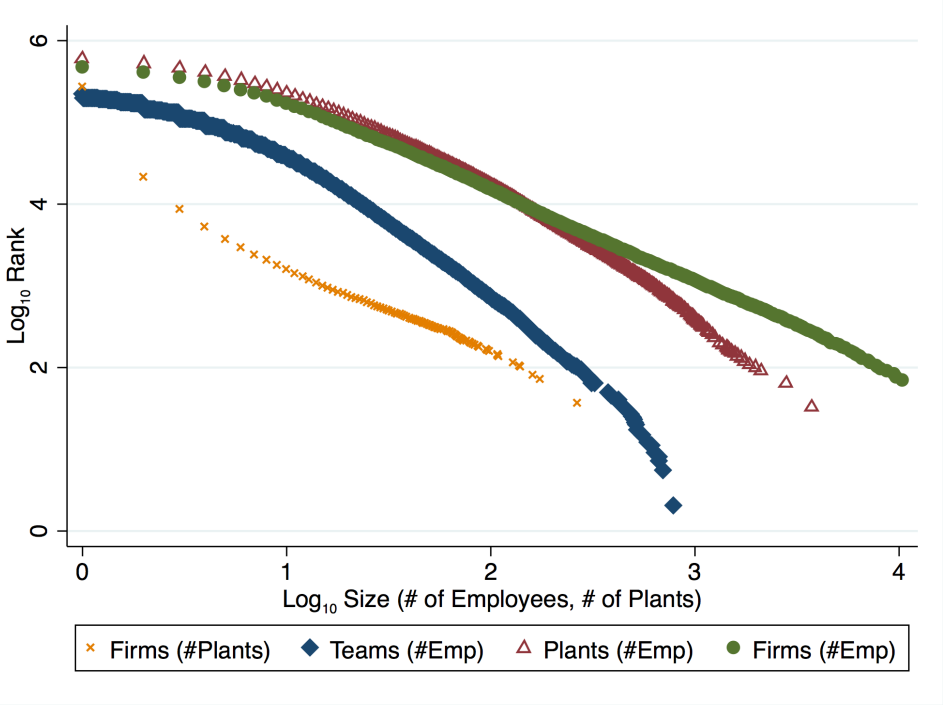
Figure 19: All sectors (Administrative Data)



Data. I next describe the data behind Figure 1 more in detail. I use a sample of French Déclarations Annuelles de Données Sociales (DADS). I report only the data from a sample in year 2007, however other time periods show very similar pictures. The sample originally contains 55, 979, 881 observations of employee-employer matched pairs.

I follow Caliendo et al. (2014) and use the first digit of the PCS variable (PCS stands for "Profession, Catégorie Socioprofessionnelle", or social class based on occupation) as a proxy for the hierarchical position of workers in a firm. That allows me to divide workers into subgroups of high ranking managers, middle managers, workers, manual workers. More precisely, the first digit of the PCS variable is one of six alternatives. The first category contains the farmers. The second group corresponds self-employed and owners (for example, plumbers, firm directors, Chief

Figure 20: Wholesale Trade, Accommodation, Food (Administrative Data)



Note: Source: French matched employer-employee data, year 2007.

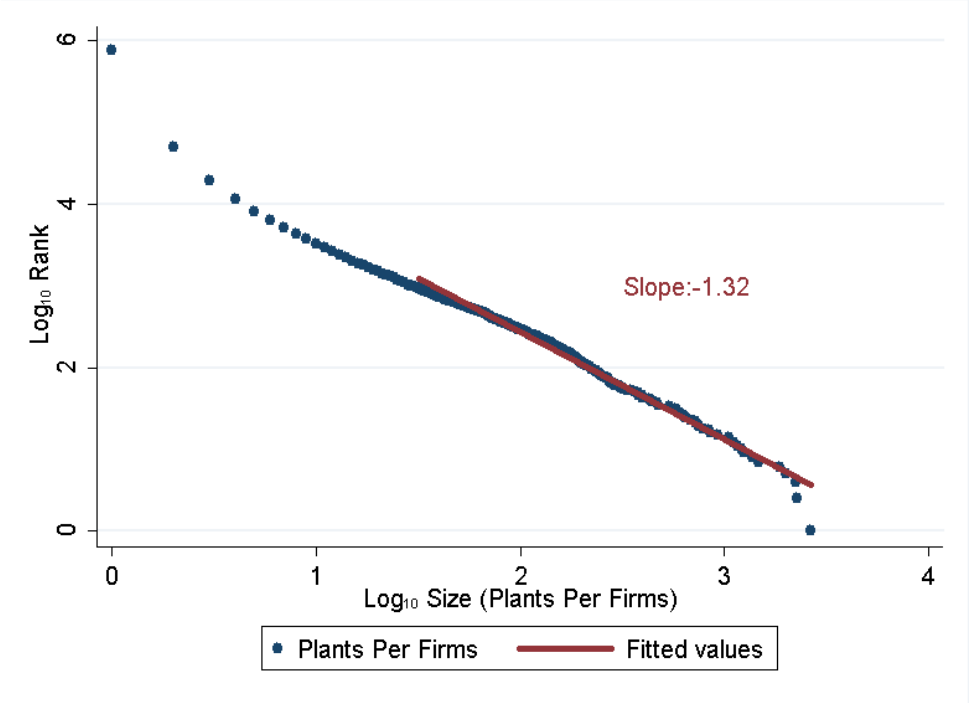
Executive Officers). The third groups senior staff or top management positions (for example, chief financial officers, heads of human resources, or purchasing managers). The fourth corresponds to employees at the supervisor level (for example, quality control technicians, sales supervisors). The fifth puts together clerical or white collar employees (for example, secretaries, human resource or accounting, and sales employees). Finally, the sixth and last group correspond to blue collars workers, for example assemblers, machine operators or maintenance workers.

Unfortunately, the PCS variable is not available for all workers, and farmers and manual workers are in separate categories (first digit equals to 1 and 2), so I drop them. This leaves me with employer-employee matches corresponding to the PCS variable having a first digit equal to 3, 4, 5, or 6.

Results. Proposition 3 states a very general result: the distribution of span of control down one level of hierarchical organization is close to a Pareto distribution with coefficient 2, whatever the underlying distribution of skills. (provided it is bounded away from zero for high skills) Figure 1 plots the corresponding distribution of span of control that is obtained empirically, with the dataset described above. The total number of employees in a given layer, as defined by the first digit of the PCS variable, is divided by the total number of employees occupying the layer above them (again, this is seen through the PCS variable). This measure is calculated at the establishment level. These ratios are referred to on the x-axis as "Team Sizes". The distribution of them is shown on a log-log plot, with the rank on the y axis (plotting the survivor function

would give a similar picture). As can be seen on Figure 1, the data does seem to point to a Pareto distribution for large spans of control in the upper tail, and a coefficient close to that predicted by the theory, 1.96.

Figure 21: Plants Per Firms Distribution (Administrative Data)



Note: Source: French matched employer-employee data, year 2007.

Figure 19 and Figure 20 add further support to the theory developed in this paper. They are constructed with a similar methodology as Figure 1. Figure 19 shows that the distribution of firm sizes is indeed more fat tailed than that of teams, establishments, or the number of plants per firms. Figure 20 illustrates the fact that this pattern is not specific to a particular sector. Indeed, the sector "wholesale trade, accommodation, food" shows very similar patterns as the overall economy, and so do other two digit NAICS industries in unreported graphs. Finally, Figure 21 shows that the distribution of the number of establishments per firms is strikingly close to what should be expected from the theory. As shown in Section 2.5, the distribution is expected to be closer to a Pareto distribution as one approaches the highest levels of hierarchical organization, as the uniform approximation for the distribution of skills becomes a very good approximation for the top (given that the corresponding cutoffs z_l^L are very close to 1). This is exactly what can be seen on Figure 21, compared for example to Figure 1, where Pareto is a good approximation for most of the distribution, not just for the upper tail. Moreover, note also how the Pareto is shifted for very high levels of span of control (that is, concave, and bounded), just as the theory predicts for non zero heterogeneity in skills. The coefficient estimated for this distribution (1.33) further seems to suggest that there are two levels of hierarchical organization between the CEO of the firm and each one of the establishment managers.

3 Top Labor Income Distribution

The equilibrium spans of control for managers was derived without any reference to equilibrium prices. This was possible because the model has a block-recursive property. Zipf's law can therefore be understood independently from the supporting labor income distribution.

However, the prices which sustain these equilibrium span of control distributions are equally, if not more, interesting from an economic point of view. When the race between technology and education parameter ρ is positive, Pareto distributions for top incomes arise. The model therefore provides a new rationalization of the Pareto nature of the top labor income distribution, one which does not rely on any Pareto functional form assumptions for primitives, and which does not have any relation to random growth theory.

I first discuss the theory (Section 3.1), then provide some comparative statics exercices (Section 3.2), and finally provide a preliminary look at the evidence (Section 3.3). The relationship to models where the distribution of firm sizes is taken as exogenous, as in Terviö (2008) or Gabaix and Landier (2008), is discussed in Section 4.2.

3.1 Theory

The competitive equilibrium can be decentralized with a set of transfers inside the firm, which can be viewed as internal labor markets. And just as the span of control distribution in the upper tail does not depend on the distribution of skills in the population, the upper tail of the labor income distribution is essentially invariant to the shape of the skill distribution.

Denote by $w_0(.)$ the wage that production workers get, and by convention $w_L(.)$ the wage that the top level of the hierarchy "implicitly" gets through production, given therefore by:

$$\forall x_L \in [z_L^L, z_{L+1}^L], \quad w_L(x_L) = x_L.$$

The return of a manager of level l is then given by the number of problems that the hierarchy solves on the aggregate minus what is given back, for each unit of problem, to lower level workers:

$$R_{l+1}(x_{l+1}) \equiv \left(w_{l+1}(x_{l+1}) - w_l(x_l)\right) N_{l+1\to 0}^l(x_l).$$

Note that the total wage of the workers in layer l are actually given by how many of those problems are solved by the hierarchy on top of which they are:

$$W_l(x_l) \equiv N_{l\to 0}^l(x_l) w_l(x_l).$$

By convention, let us similarly define the return function for a production worker as:

$$\forall x_0 \in [z_0^L, z_1^L], \quad R_0(x_0) = w_0(x_0).$$

All the income a production worker gets is indeed its wage: he is not able to leverage his skills, and neither does he need to pay any employee.

Lemma 7 (Internal Labor Markets)**.** *The implicit wages received by each layer of the hierarchy are given through inverse recursion, starting from $w_L(x_L) = x_L$ by:*

$$\forall l \in \{0, ..., L-1\}, \quad w_l(x_l) = \frac{1}{N^l_{l+1\to 0}(x_l)} \int_{z_l^L}^{x_l} w_{l+1}\left(m_l(u)\right) \frac{dN^l_{l+1\to 0}}{dx_l}(u)du + \frac{N^l_{l+1\to 0}(z_l^L)}{N^l_{l+1\to 0}(x_l)} w_l(z_l^L).$$

The initial conditions for these differential equations in integral form are solution of:

$$\forall l \in \{0, ..., L-1\}, \quad R_{l+1}(z^L_{l+1}) = R_l(z^L_{l+1}).$$

Proof. Managers with skill x_{l+1} maximize their revenue $R_{l+1}(x_{l+1})$:

$$R_{l+1}(x_{l+1}) = \max_{x_l} \left\{ N^l_{1+1\to 0}(x_l) \left(w_{l+1}(x_{l+1}) - w_l(x_l)\right) \right\}$$
$$\text{with} \quad N^l_{l+1\to 0}(.) = N^l_{l+1\to l}(.) * N^l_{l\to l-1}(.) * ... * N^l_{1\to 0}(.).$$

It must therefore be that, for all $l \in \{0, ..., L-1\}$:

$$m_l^{-1}(x_{l+1}) = \arg\max_{x_l} \left(w_{l+1}(x_{l+1}) - w_l(x_l)\right) N^l_{l+1\to 0}(x_l)$$
$$\Rightarrow \quad \frac{d}{dx_l}\left(w_l(x_l) N^l_{l+1\to 0}(x_l)\right) = w_{l+1}\left(m_l(x_l)\right) \frac{dN^l_{l+1\to 0}}{dx_l}(x_l)$$
$$\Rightarrow \quad w_l(x_l) N^l_{l+1\to 0}(x_l) - w_l(z_l^L) N^l_{l+1\to 0}(z_l^L) = \int_{z_l^L}^{x_l} w_{l+1}\left(m_l(u)\right) \frac{dN^l_{l+1\to 0}}{dx_l}(u)du.$$

The L indifference conditions for agents with skills z_l^L with $l \in \{0, ..., L-1\}$, which are initial conditions for the wage functions, are then written as follows:

$$\forall l \in \{0, ..., L-1\}, \quad R_{l+1}(z^L_{l+1}) = R_l(z^L_{l+1})$$

By reverse induction, this defined the whole sequence of wage functions $w_l(.)$ for all $l \in \{0, ..., L-1\}$, starting from the known $w_L(x_L) = x_L$. □

Of particular interest is the theoretical distribution of top managers' wages, both because it does not depend on the underlying distribution of skills as for the distribution of span of control, and because inequality issues have gained some prominence in the public debate recently. We are therefore interested in the distribution of $R_L(x_L)$, which obtains rather straightforwardly from Lemma 7.

Proposition 6 (Top Labor Income Distribution)**.** *The incomes at the top are given as a function of the skill of top managers x_L by:*

$$R_L(x_L) = R_L\left(z_L^L\right) + \int_{z_L^L}^{x_L} N^L_{L\to 0}(u)du.$$

The incomes at the top are therefore arbitrarily close to a Pareto distribution:

$$R_{top}(x_L) \simeq \text{Pareto}\left[\frac{(1+\rho)(2+\rho)^L}{\rho(2+\rho)^L - (1+\rho)^{L+1}}\right].$$

In fact, more generally, total returns of managers are given as:

$$R_{l+1}(x_{l+1}) = R_{l+1}\left(z_{l+1}^L\right) + \int_{z_{l+1}^L}^{x_{l+1}} w'_{l+1}(u) N_{l+1\to 0}^{l+1}(u) du.$$

One can get a justification for this formula through a very straightforward envelope condition. Since, again, managers of type $l+1$ optimize over their expected revenue $R_{l+1}(x_{l+1})$:

$$\max_{x_l} R_{l+1}(x_{l+1}) = \left(w_{l+1}(x_{l+1}) - w_l(x_l)\right) N_{l+1\to 0}^{l}(x_l).$$

We have, at the optimum:

$$\frac{\partial R_{l+1}(x_{l+1})}{x_{l+1}} = w'_{l+1}(x_{l+1}) N_{l+1\to 0}^{l}(x_l) = w'_{l+1}(x_{l+1}) N_{l+1\to 0}^{l+1}(x_{l+1}).$$

Integrating straightforwardly gives the result. And Proposition 6 is just an application of this result for $l = L-1$, in which case $w'_L(x_L) = F'(x_L) = 1$ by definition. One may however feel uncomfortable with the successive changes of variables, as well as wonder whether an envelope theorem can so directly be applied in this sequential maximization problem, so I provide a lengthier, though perhaps less problematic, proof in Appendix B.4.

Polynomial Density Functions. Assume that the distribution of skills is taken in the family of polynomial functions defined above $g(x) = (\rho+1)(1-x)^\rho/(1-\Delta)^\rho$. From the formula in Proposition 6:

$$R_L(x_L) = R_L\left(z_L^L\right) + \int_{z_L^L}^{x_L} N_{L\to 0}^{L}(u) du.$$

From the fact that:

$$N_{L\to 0}^{L}(x_L) \simeq Pareto(\alpha_L) \qquad \text{with} \quad \alpha_L = \left(\frac{1}{1+\rho}\frac{(2+\rho)^L}{(2+\rho)^L - (1+\rho)^L}\right)$$

From straightforward integration, and in the case where α_L, we have that:

$$R_L(x_L) \simeq Pareto\left(\frac{1}{\frac{1}{\alpha_L} - 1}\right).$$

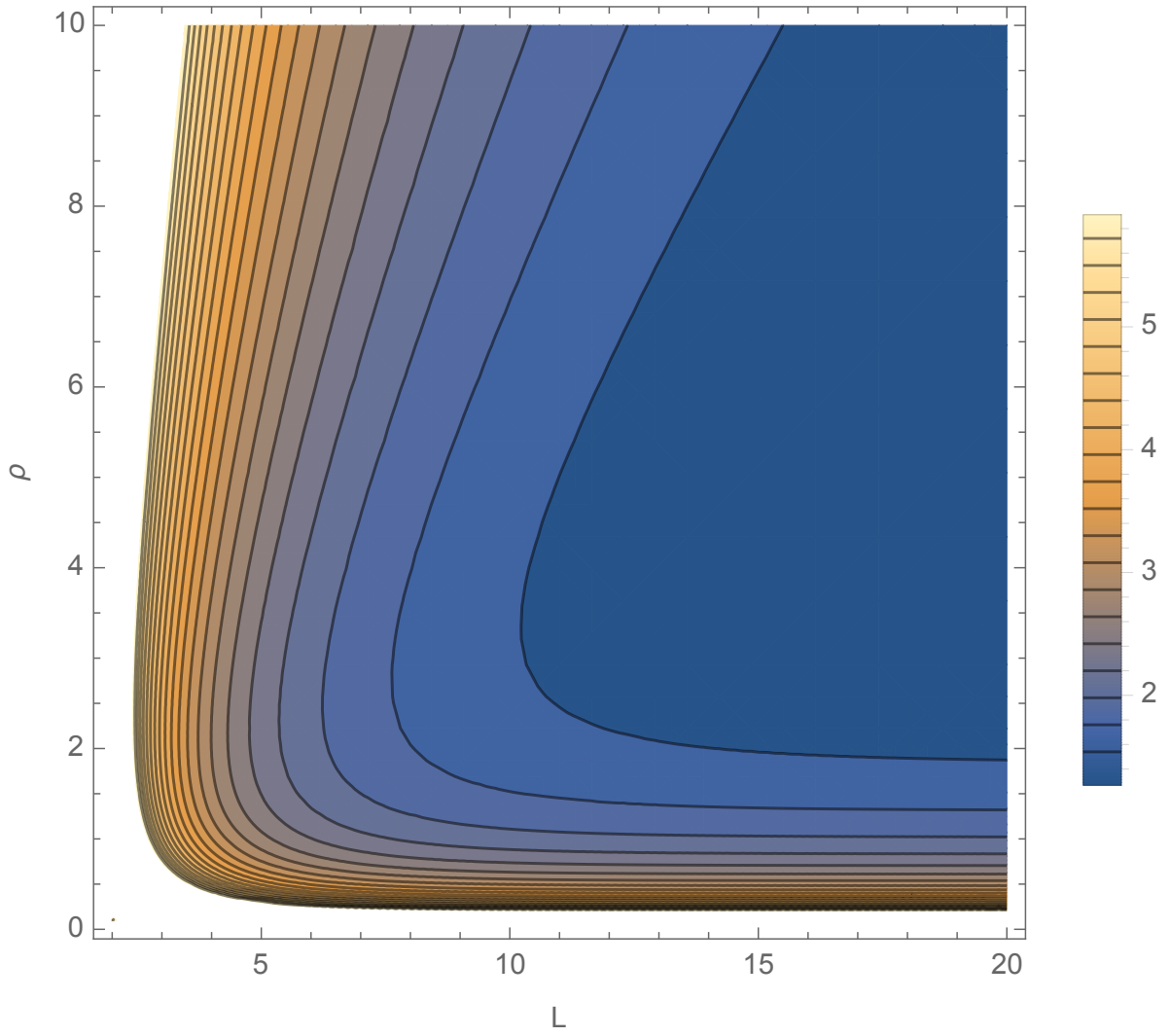
Finally, the top income distribution as measured in the data is arbitrarily close to a Pareto in the upper tail with:

$$R_{top}(x_L) \simeq Pareto\left[\frac{(1+\rho)(2+\rho)^L}{\rho(2+\rho)^L - (1+\rho)^{L+1}}\right].$$

The Pareto Coefficients for Top Incomes therefore vary as a function of ρ and L, in a manner illustrated on Figure 22. Note that those Pareto coefficients are comprised between 1 and 3 for plausible values of the parameters.

General Distribution of Skills. Again, these cases can be generalized to the case of sufficiently regular functions, which can be approximated around one by one of the above polynomials as a Taylor expansion.

Figure 22: Theoretical Top Labor Incomes Pareto Coefficients, as a function of ρ and L



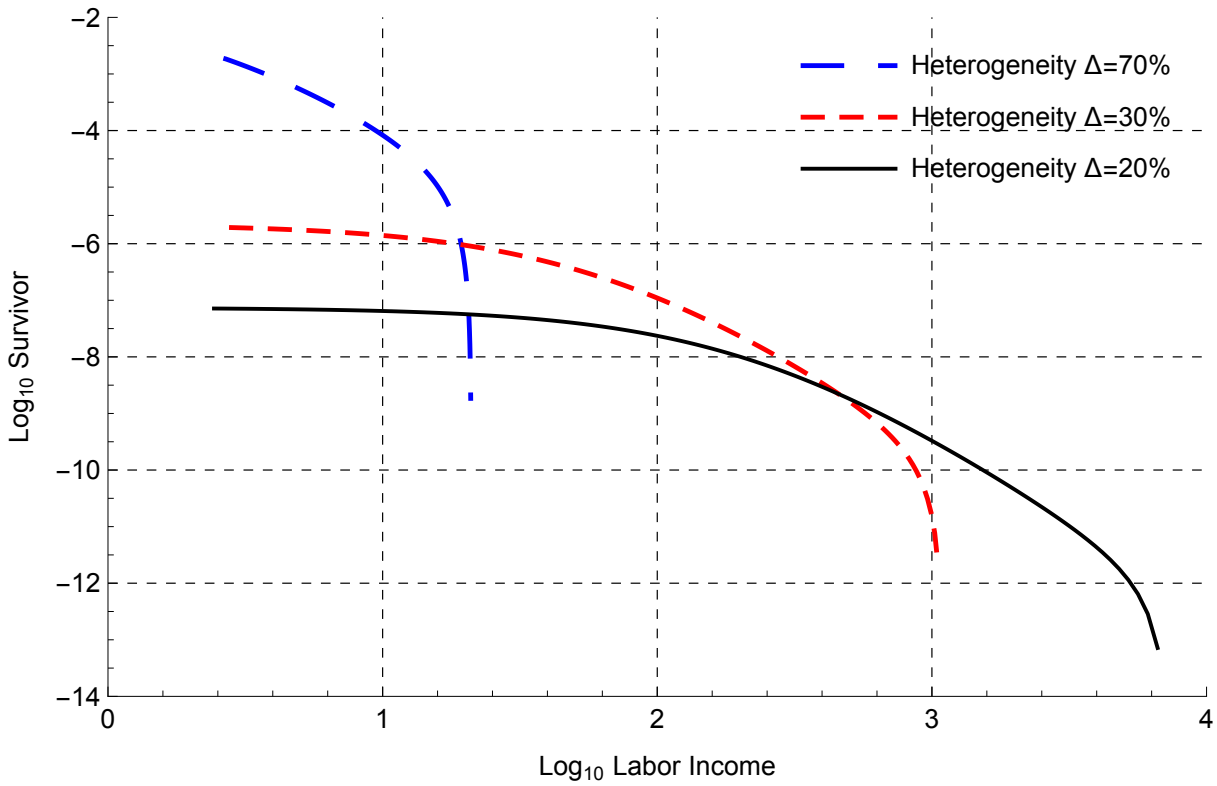
Note: This chart represents the Pareto Coefficients for Top Incomes as a function of the race between technology and education, embodied in the density of highest skills at the top, and the equilibrium number of layers of hierarchical organization. Darker colors correspond to lower Pareto coefficients for Top Incomes (converging to 1), so a relatively fatter tail and more inequality. The red dot represents the maximum level of inequality attainable for $\rho \in [0, 20]$ and $L \in [1, 20]$, with a Pareto coefficient equal to 1.26.

3.2 Comparative Statics

Three comparative statics are of interest and can all lead to an increase in top income inequality, or rise of the top 1% income share.

Decrease in communication costs. The comparative statics exercises involving a decrease in communication costs are rather straightforward. As in Garicano and Rossi-Hansberg (2006), decreases in communication costs lead to larger and more heterogenous span of control for man-

Figure 23: Negative Relationship between Heterogeneity in Agents' Skills and Top Labor Incomes Inequality, $h = 99\%$, $\rho = 1$



agers, and therefore rising income inequality. (in fact, it leads also to rising wage polarization) This corresponds to Figure 32 in Appendix E.

Through comparative statics on parameter h, the model can therefore give some flesh to the skill-biased technical change hypothesis, presenting an alternative to a macroeconomic model with skilled and unskilled labor appearing as imperfect substitutes in a neoclassical production function (Katz and Murphy (1992)). It can also explain why a macroeconomic analysis leads to attribute a large share of skill-biased technological change to a rise in capital equipment, if the latter is complementary with skilled labor as in Krusell et al. (2000). Indeed, decreases in communication technology both appear as more capital equipment in firms that now use information technology more intensively, and lead skilled individuals to manage more workers through larger firms.

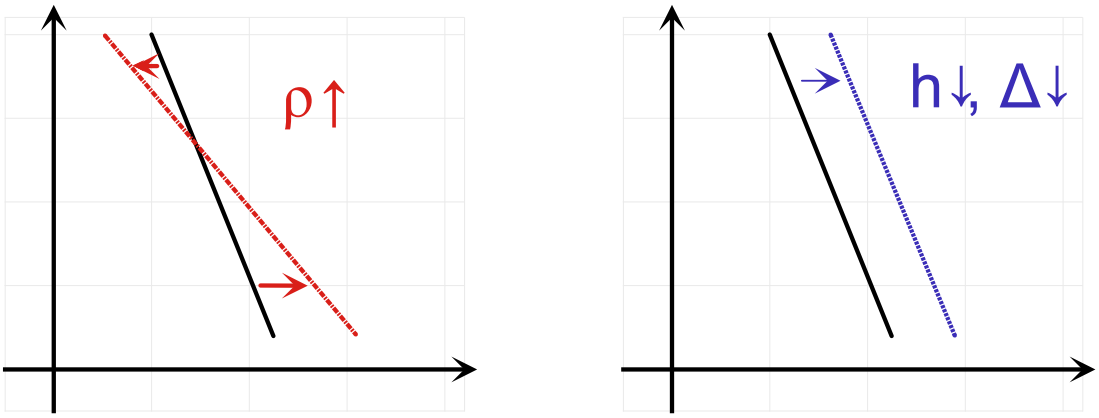
Increase in ρ. Comparative statics with respect to ρ are a little bit less straightforward. In particular, interesting and not completely intuitive is that when $\rho = 0$, that is when the distribution is bounded away from zero for higher skills, income inequality is at a low. One can see that straightforwardly on the top incomes distribution formula in Proposition 6. Integrating the Pareto distribution with a coefficient higher than 1 leads to a distribution that is no longer Pareto, as the integral of $N^L_{L\to 0}(.)$ converges. The intuition is that managers then compete to

attract the best workers, and they end up giving almost all the surplus from the match to them. In contrast, for higher values of ρ, that is when talented individual are in relatively scarce supply, everything happens "as if" they had some market power: the distribution of the surplus shifts towards the managers. This force is actually stronger than a countervailing force: because there are fewer good managers, the workers also are relatively less competent, and so the span of control of managers is lower. These comparative statics are illustrated on Figure 33 in Appendix E.

Decrease in heterogeneity. Just as the span of control distribution is more heterogenous when agents' skills are more similar (Δ decreases), the comparative statics of the labor income distribution with respect to heterogeneity are, again, quite counterintuitive. The intuition again is that when heterogeneity decreases, workers that managers hire are relatively more competent and can almost answer all questions by themselves, so that they allow managers to achieve a higher span of control ((Figure 23).

To summarize, decrease in communication costs and heterogeneity both lead to a shift of the wage distribution to the right on a log wage-log survivor Pareto plot, as is shown on the bottom-right panel of Figure 24. In contrast, a higher race between education and technology leads to an increase in the Pareto coefficient, as is shown on the bottom-left panel. This allows to potentially disentangle these two effects from micro-data.

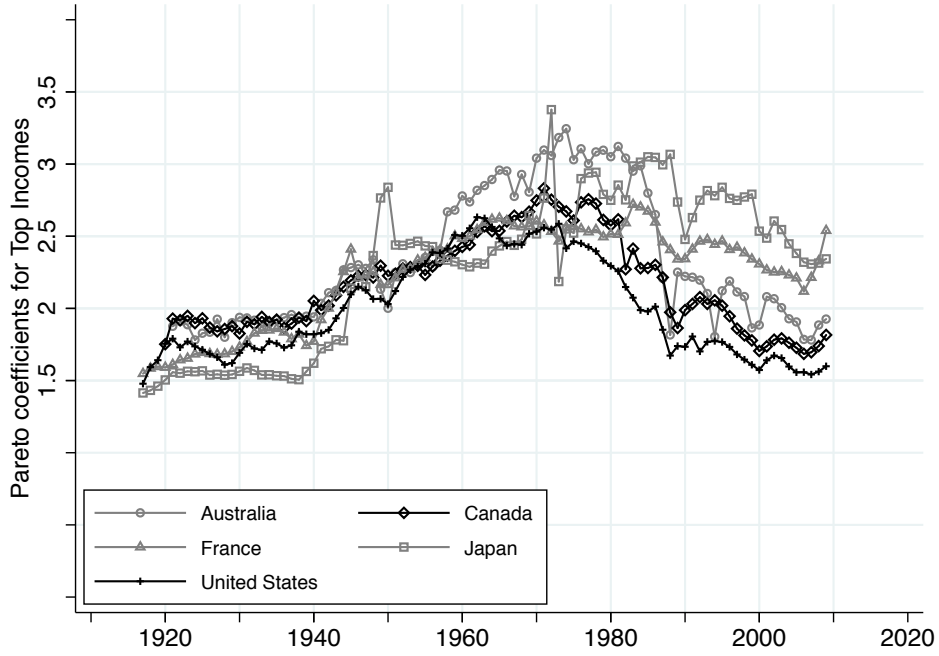
Figure 24: What Caused the Increase in Inequality ? A Structural Approach



3.3 Evidence [Preliminary]

As section 3.2 has shown, one would ideally need micro-level data in order to assess the different factors having contributed to the recent rise in labor income inequality. Preliminary evidence seems to suggest that Pareto coefficients for top incomes have lowered, pointing to a race between education and technology (see the US series on Figure 25). The Survey of Consumer Finances data on Figures 29(a) and 29(b) seems to point to the same tendency. This assessment however

Figure 25: Empirical Top Incomes Pareto Coefficients in a Sample of Advanced Economies (1920-2010).



Note: Source: World Top Income Database. This plot shows the evolution of Pareto-Lorenz coefficients α, as taken from the World Top Income Database (Atkinson et al. (2011), see also Alvaredo et al. (2013)). The sample is restricted to the countries having the longest series on record.

remains preliminary at this stage. Note that the top 1% income share, on which data is far more reliable, is not a sufficient statistic to investigate this issue. In the model, it is consistent with a decrease in communication costs, a larger race between education and technology than before, or (a new insight) a decrease in average worker skill heterogeneity.

4 Discussion

Section 2 has shown how to understand Zipf's law from a production hierarchies model with potentially arbitrarily small differences in skills between agents, and without any functional form assumption. However, another leading explanation for Zipf's law is a dynamic stochastic process of firm growth with statistical frictions. I discuss the relationship of this paper to the random growth literature in Section 4.1.

Section 3 has shown how to derive Pareto distributions for top labor incomes in the same unifying model. It is however useful to discuss the advantages of explaining Zipf's law for firms at the same time as examining wage inequality. In Section 4.2, I therefore compare the above approach to the literature taking the firm size distribution as exogenous, notably Terviö (2008) and Gabaix and Landier (2008).

4.1 Relationship to the Random Growth literature

The leading explanation given for the existence of Zipf's law is based on a random growth model with statistical frictions. In its simplest version, the model assumes Gibrat's law and a reflecting barrier. Zipf's law is then understood as the stationary distribution of this process. Luttmer (2010) is a recent survey of this approach. However, this explanation comes with a number of well-known puzzles, both empirical and theoretical.

Empirics. First, it has been known for a long time, at least since Mansfield (1962), that Gibrat's law for firm growth does not hold so well in the data, and that variance in growth rates is actually a decreasing function of size. More recent evidence in Rossi-Hansberg and Wright (2007) shows also that both the mean and the variance and firm's growth rates decrease with size.

Second, Cabral and Mata (2003) present even more puzzling evidence, if one has in mind that a random growth process is at work in the economy. They show on Portuguese micro level data that the distribution of firm sizes is already very skewed to the right at the time of birth. More importantly, the fact that small firms conditional on survival grow at a faster rate than large firms has been argued to be consistent with Gibrat's law if there is a lot of firm exit (for example, because firms gradually learn how productive they are, as in Jovanovic (1982)). However, Cabral and Mata (2003) show that selection only accounts for a very small fraction of the evolution of firm sizes.

Third, the overall firm size distribution presents at least three deviations from Zipf's law. The Pareto coefficient is approximately 1.059 in the United States. The firm size distribution is bounded unlike the Pareto distribution (and moreover has a thinner tail in the very upper tail than the Pareto distribution). And Pareto is a good approximation only for the upper tail of the distribution, not at all for small firms which are better described by the lognormal. In other words, Zipf is a good fit only for firms which have between 50 and $100,000$ employees. These deviations are all visible on Figure 30 taken from Axtell (2001).

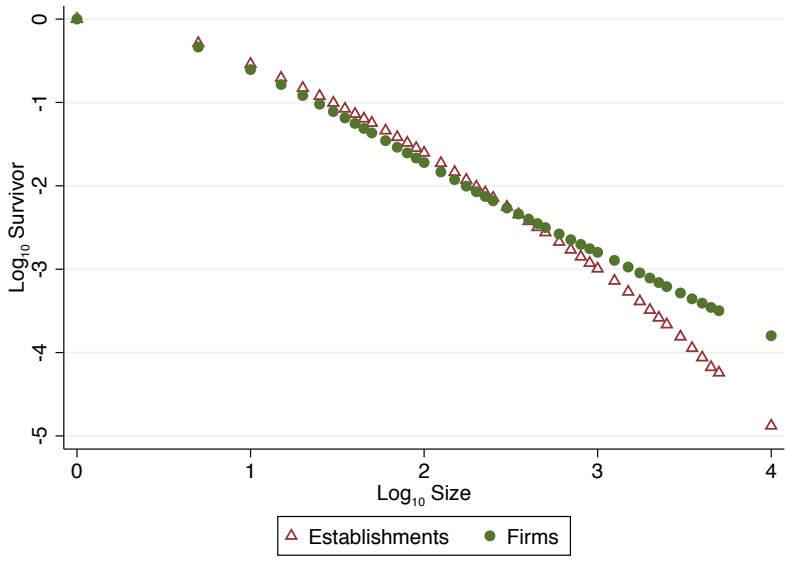
Perhaps the most compelling evidence in favor of the mechanism presented in this paper is the evidence presented in Section 2.7, concerning the size distribution of span of control at more disaggregated levels of the firm's hierarchy. This evidence and in particular the Pareto distribution with a coefficient equal to two shown on Figure 1 has, to the best of my knowledge, no counterpart in random growth theory.

But this new micro-data evidence should in fact not come as a surprise to the researcher working on firm size's distributions. It has long been known that the distribution of establishments was less fat tailed than that of firms. For example, Luttmer (2010) writes in his review of the random growth literature: "In the United States, the right tail of the size distribution of establishments is noticeably thinner than that of firms."[12] The model presented in this paper

[12]Similarly, Rossi-Hansberg and Wright (2007) write: "It is worth noting that the size distribution of enterprises is much closer to the Pareto, especially if we focus attention on enterprises with between 50 and 10,000 employees.

is perfectly suited to address this particular issue, as it is precisely the way in which Zipf's law is generated, through the multiplication of Pareto distributions with a higher tail coefficient. Figure 26 illustrates the fact that establishment sizes is also Pareto distributed in the upper tail in the United States, albeit with a higher tail coefficient, through publicly available US Census data. To the best of my knowledge, no random growth mechanism to date is able to explain simultaneously the distribution of team sizes, establishment sizes and firm sizes. This is perhaps the main supporting evidence for the model.

Figure 26: US Firm and Establishment Sizes



Note: Source: Census Bureau, Statistics of US Businesses, 1990.

Theory. The main puzzle to an economist observing firm size distributions is that although employment at individual firms appears to be non stationary, the distribution of firm employment on the contrary is stationary: it always follows (at least in the tail) Zipf's law. This is what may have led researchers to write dynamic models of firm growth and to look for stationary distributions of these processes. Under Gibrat's law, a scale-independent process, the resulting stationary size distribution is also scale independent, so it has to be a Pareto distribution. But in these theories, stationarity (or balanced growth, in the case of a growing economy) is more of an assumption than a conclusion. In contrast, a static model generating Zipf's law gives a very straightforward intuition for why the firm size distribution appears stationary: regardless of the distribution of productivities and of communication costs (which may well evolve), Zipf's law arises in the upper tail, through a static mechanism. In this theory, stationarity therefore is not an assumption, but a result.

The differences between the size distributions for establishments and enterprises may shed light on the forces that determine the boundaries of the firm. Our theory focuses, however, on the technology of a single production unit and does not address questions of ownership or control."

A second assumption needed for random growth processes to generate Zipf's law, is that they follow Gibrat's law, or have a scale independence property. In the interpretation where organization capital is replicated from existing capital, this scale independence results from the assumed firm growth process. Ultimately, scale independence in the stationary distribution arises from this scale independence in the process, which is assumed. In Section 2.4, I have laid out the intuition for where scale independence originates from in this model. In have in particular explained why it does not rely on any functional form underlying the economy's primitives. I should add that in a dynamic extension of this model, firm growth would likely seem to be scale independent, although the way in which the number of layers is endogenously determined could lead to discontinuities in the growth process, which I leave to future research. Again, scale independence is a result in the model, not an assumption.

4.2 Relationship to Terviö (2008) and Gabaix and Landier (2008)

In many ways, the results on Pareto distributions for top labor incomes have a flavor of the results in Terviö (2008) and Gabaix and Landier (2008). And indeed, I show in Appendix B.5 how to map the discrete model in Gabaix and Landier (2008), where everything is expressed as a function of the rank of CEOs in the talent distribution, into the one presented here. There are however many differences between these two papers and this one, even to understand the top labor income distribution.

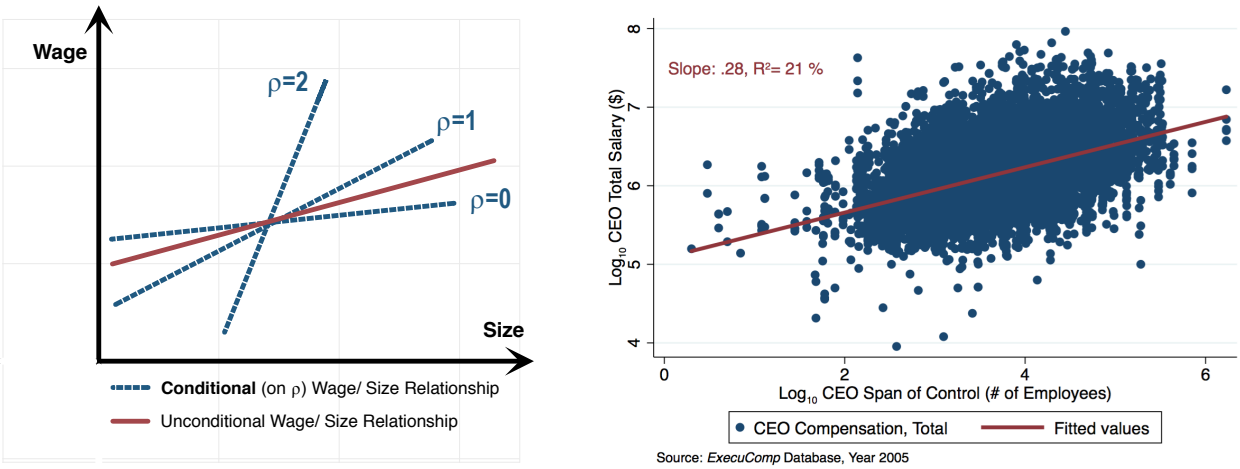
The most important of them is that Zipf's law is taken as exogenous in their model. This is an important difference theoretically because ultimately, all the heterogeneity in wages that Gabaix and Landier (2008) obtain comes from the heterogeneity in span of control, which remains unexplained. From a positive perspective, if one wants to understand the ultimate source of wage inequality, and why this it takes the form of a Pareto distribution, the origins of Zipf's law need to be understood. To know why Rosen (1982)'s economics of superstars apply to CEOs, one needs to understand why some firms are so large to allow some people to express their small differences in talents, and why the distribution of the size of these firms is Pareto.

Moreover, Gabaix and Landier (2008) postulate a linear interaction between managers' talents and the size of a firm, which the production hierarchies model allows to microfound. This linear interaction comes from the fact that talent is defined as a probability of being able to solve problems, and that the number of problems that are solved by the whole organization is given by the number of employees in the firm. The model also applies very naturally to explain the whole wage distribution, not just that of CEOs, and to explain in particular why other executives than CEOs can also enjoy high levels of compensation. Another advantage is that firm size is an equilibrium outcome, so that no CEO has an incentive can expand the size of his firm in equilibrium because of the time constraint.

Perhaps more importantly, a big difference between the two models concerns the relationship between size and pay, which they name Roberts' law, after Roberts (1956). In Gabaix and

Landier (2008), there is a one-way relationship between the two: the most talented managers run the largest firms, and the relationship between the two is a constant. Even with multiple sectors, with different values of ρ, and assuming that CEOs could not move between sectors, there would be no a priori reason to expect the firm size distributions' to be different, and so this relationship would still hold. In this model however, the relationship between size and pay is not so automatic. Just as in Gabaix and Landier (2008), it is really those sectors with the highest ρ, or where talented managers are in scarcer supply, which feature the largest correlation between pay and size. The reason is that when talented individuals are not so rare, firms have less incentives to hire the really best CEOs, and so pay is less correlated to size. But in contrast to Gabaix and Landier (2008), higher ρ also means lower firm sizes on average (and even, lower number of layers, so higher Pareto coefficient): this is because workers are really less skilled on average, and so do not allow managers to achieve as high a span of control. The left panel of Figure 27 shows that theoretically, the largest wages do not then necessarily correspond to the largest firms. This non systematic correlation between pay and size is a new qualitative insight, which seems to hold true in the data, as shown for ExecuComp in year 2005 on the right panel of Figure 27. More concretely, one can think of high ρ firms as firms where the hardest problems cannot be solved by many agents, so relatively "high-tech" (or "new economy") companies. Those companies are not necessarily the largest, because managers hire workers who are not very skilled in that technology, but nevertheless these are the sectors where the pay-size sensitivity is the highest.

Figure 27: Roberts' Law, in Theory (Left) and Practice (Right).



Endogenizing Zipf's law and top managers' wages at the same time also allows to get a better understanding of the comparative statics exercices: the increase in the size of firms between 1980-2003 is something that Gabaix and Landier (2008) need to take as given. The model presented in this paper in contrast allows to trace that increase back to a decrease in the costs of Information Technology, or to a decrease in overall skill heterogeneity (see Section 3.2). In

particular, neglecting the effects of heterogeneity on span of control would lead one to conclude through Gabaix and Landier (2008) that less heterogeneity in skills overall leads to less inequality, because potential CEOs have more similar talents, so the market gives them more similar wages. However, a stronger, countervailing force is that the best of them now enjoy even greater span of control, which tends to increase inequality. I showed in Section 3.2 and Figure 23 in particular, that the latter force dominates the former. A reduced form analysis a la Gabaix and Landier (2008) therefore misses this key novel qualitative insight: inequality increases when heterogeneity decreases.

Finally, having a microfounded model of both firm sizes' and labor incomes' distributions can be used to investigate other issues, such as that of optimal labor income taxation. In Gabaix and Landier (2008), CEO wages are indeed pure economic rents, which could as a consequence very well be taxed away. In contrast, the model presented here extends very naturally to a Mirrlees (1971) setting where the units of time (labor supply) which an agent supplies are endogenous. Taxing away CEO wages could then lead them to supply a lower number of hours, which would lead to a reduction of the equilibrium size of firms, and entail potentially large efficiency costs. The model therefore provides an alternative to assuming a Pareto distribution of skills, which is often done in the optimal labor income taxation literature (Diamond (1998), Saez (2001)). Whether or not that affects optimal labor income tax formula is an interesting avenue for future research.

5 Concluding Remarks

In this paper, I have developed a new static, microfounded, hierarchy-based model of Zipf's law for firm size's distribution. Functional form assumptions are very commonly based on the observed Zipf's law for firm sizes, such as the distribution of productivities in the heterogenous firms literature. My model suggests on the contrary that no mechanic link between the two exists, but that Zipf's law arises regardless of the underlying distribution of productivities. Even the comparative statics exercises can get it wrong: heterogeneity in firm sizes rises, instead of decreases, when productivity heterogeneity decreases. As already emphasized in the literature review, the implications of the finding therefore span many different parts of the economics literature, from the magnitude of misallocations stemming from size-dependent regulations to the sources of high labor income inequality. In particular, the fact that heterogeneity can very well be bounded, even infinitesimal, and yet yield what looks like an unbounded heterogeneity in outcomes, suggest that more parsimonious structural models, in the sense of incorporating less ex-ante heterogeneity on primitives, can be written to explain the same observed phenomena.

Unfortunately, the model is very much a first pass in many respects. Its most important limitation is perhaps that skills are assumed to be a given in the model, and that all agents work the same, so that unequal wages only result from innate abilities. I conjecture that the span of control distributions would remain the same in a larger class of extended models with education and elastic labor supply, as differences in education abilities are in many ways ex-post isomorphic

to ex-ante differences in skills, and labor supply decisions would only amplify the forces at hand in the paper. Yet for other issues, this simplification is clearly a problem. Consider optimal taxation. If abilities were all innate, a Rawlsian planner would be able to redistribute all unequal labor income in equal proportions at no efficiency cost. Yet, skills are not entirely innate, and doing so would discourage both skill investment through education as well as more labor supply from agents at the top of the distribution; because the model is purely competitive, prices provide exactly the right incentives to educate themselves more, or to work more.[13] Therefore, in its current form, the model remains essentially silent on the key efficiency-equity tradeoffs. Not completely, though: since optimal taxation models are very often calibrated using an underlying Pareto distribution in agents' productivity (see, for example, Diamond (1998), Saez (2001), Tuomala (1990)), the model holds promises for the structural estimation of optimal income tax rates. As already stressed, CEO wages are in contrast pure economic rents from the partial equilibrium analysis in Terviö (2008) and Gabaix and Landier (2008), and could therefore be taxed away at no efficiency cost. In contrast, labor income taxation would have an effect on the size of firms in a model with elastic labor supply, and so on the number of employees benefiting from a manager's superior skills.

The model presented in this paper attributes all heterogeneity in firm sizes to the heterogeneity in skills of the workers and managers they employ, which is consistent with some empirical evidence attributing most wage differentials to person effects instead of firm effects (for example Abowd et al. (1999)).[14] Even though the model really matches the data very well, attributing all heterogeneity to agents and none to firms may be too extreme an assumption. An interesting extension of the model would therefore be to allow for firm specific heterogeneity in excess of managers' respective talents. Along these lines, an important factor that the model has neglected is firm's capital, including management capital. In the context of the model, one would model firm heterogeneity as different values for communication costs h, which can proxy for how much a firm has invested in better management tools, or extend the model to one of moral hazard between workers and entrepreneurs along the lines of Bloom et al. (2012).

The fact that the skills distribution is exogenously fixed is not only an issue for normative economics, but also for a more thorough positive investigation. This is arguably a good assumption in the short run, for example after the arrival of a new technology; or to capture the fact that some skills cannot be taught, or that some are in any case always better at learning than others. If one however allows for homogenous learning ability, then the model could help answer a number of fascinating questions. In particular, it is likely that the heterogeneity in skills generating a very unequal income distribution does not forever remain as the incentives to invest in skills are then very high. A dynamic version of the model would therefore allow to

[13] This would likely be reinforced in the presence of learning externalities as in human capital accumulation models.

[14] Note moreover that in the model developed in Section 3.1, and using Abowd et al. (1999)'s methodology, the econometrician would attribute to firm effects the fact that a worker has moved to a firm with more productive colleagues, because of complementarities.

investigate the "race between technological development and education" which may have been the cause of the recent rise in inequality (Tinbergen (1974), Katz and Murphy (1992)). On a same line of reasoning, a second limitation of the model is that the good being produced is itself exogenous. If one allows for innovation, and in particular for the fact that at the beginning of the product cycle only few entrepreneurs are able to design it, then the question would be to study the incentives to spread around this knowledge, to increase span of control (which increases when the skills of lower ranking agents increases); and the countervailing forces to retain knowledge, as the wage distribution is all the more rewarding to entrepreneurs that there are relatively few of them.

On a more general note, the model confirms that competitive forces, to generate the maximum level of complementarities, can lead to very high level of heterogeneity in outcomes (in this model, firm employment, and labor incomes) from an arbitrarily small level of heterogeneity on primitives. In Geerolf (2013), I have shown that this was true also for a model of collateralized lending with heterogenous beliefs and endogenous margins. In that case, very optimistic investors were able to manage a lot more capital than less optimistic ones, again according to a Pareto distribution, because this arrangement was constrained optimal and maximized ex-ante welfare. A natural conjecture is that Pareto distributions for allocations in fact obtain for a more general class of assignment models with complementarities. This is left to future research.

References

Abowd, J. M., Kramarz, F. and Margolis, D. N. (1999), 'High Wage Workers and High Wage Firms', *Econometrica: Journal of the Econometric Society* **67**(2), 251–333.

Acemoglu, D. and Autor, D. (2011), 'Skills, Tasks and Technologies: Implications for Employment and Earnings', *Handbook of Labor Economics* **4**(11), 1043–1171.

Alvaredo, F., Atkinson, A. B., Piketty, T. and Saez, E. (2013), 'The Top 1 Percent in International and Historical Perspective', *Journal of Economic Perspectives* **27**(3), 3–20.

Antràs, P., Garicano, L. and Rossi-Hansberg, E. (2006), 'Offshoring in a Knowledge Economy', *Quarterly Journal of Economics* **121**(1), 31–77.

Atkinson, A., Piketty, T. and Saez, E. (2011), 'Top Incomes in the Long Run of History', *Journal of Economic Literature* (49), 3–71.

Axtell, R. (2001), 'Zipf Distribution of US Firm Sizes', *Science* **293**, 1818–1821.

Bloom, N., Sadun, R. and Reenen, J. V. (2012), 'The Organization of Firms Across Countries', *Quarterly Journal of Economics* **127**(4), 1663–1705.

Cabral, L. and Mata, J. (2003), 'On the Evolution of the Firm Size Distribution: Facts and Theory', *American Economic Review* **93**(4), 1075–1090.

Caliendo, L., Monte, F. and Rossi-Hansberg, E. (2014), 'The Anatomy of French Production Hierarchies', *Journal of Political Economy* (forthcoming).

Caliendo, L. and Rossi-Hansberg, E. (2012), 'The Impact of Trade on Organization and Productivity', *Quarterly Journal of Economics* (3), 1393–1467.

Champernowne, D. (1953), 'A Model of Income Distribution', *The Economic Journal* **63**(250), 318–351.

Diamond, P. (1998), 'Optimal Income Taxation : An Example with a U-Shaped Pattern of Optimal Marginal Tax Rates', *American Economic Review* **88**(1), 83–95.

Eeckhout, J. (2004), 'Gibrat's Law for (All) Cities', *American Economic Review* **94**(5), 1429–1451.

Eeckhout, J. and Kircher, P. (2012), 'Assortative Matching With Large Firms: Span of Control over More versus Better Workers', *Working Paper* (February 2012).

Gabaix, X. (1999), 'Zipf's Law for Cities: an Explanation', *Quarterly Journal of Economics* **114**(3), 739–767.

Gabaix, X. (2009), 'Power Laws in Economics and Finance', *Annual Review of Economics* **1**(1), 255–293.

Gabaix, X. (2014), 'Power Laws in Economics : An Introduction', *prepared for the Journal of Economic Perspectives* (1-21), http://pages.stern.nyu.edu/xgabaix/papers/pl–jep.pdf.

Gabaix, X. and Landier, A. (2008), 'Why Has CEO Pay Increased so Much?', *Quarterly Journal of Economics* **1**(February), 49–100.

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

Garicano, L. and Rossi-Hansberg, E. (2004), 'Inequality and the Organization of Knowledge', *American Economic Review, Papers and Proceedings* **94**(2), 197–202.

Garicano, L. and Rossi-Hansberg, E. (2006), 'Organization and Inequality in a Knowledge Economy', *Quarterly Journal of Economics* **121**(4), 1383–1435.

Geerolf, F. (2013), A Theory of Power Law Distributions for the Returns to Capital and of the Credit Spread Puzzle.

Gibrat, R. (1931), *Les Inégalités Economiques; Applications: aux inégalités des richesses, à la concentration des entreprises, aux populations des villes, aux statistiques des familles, etc., d'une loi nouvelle, la loi de l'effet proportionnel*, Librairie du Recueil Sirey, Paris.

Head, K., Mayer, T. and Thoenig, M. (2014), 'Welfare and Trade without Pareto', *American Economic Review, Papers and Proceedings* **104**(5), 310–316.

Hopenhayn, H. and Rogerson, R. (1993), 'Job Turnover and Policy Evaluation: A General Equilibrium Analysis', *Journal of Political Economy* **101**(5), 915–938.

Hsieh, C. and Klenow, P. (2009), 'Misallocation and Manufacturing TFP in China and India', *Quarterly Journal of Economics* **124**(4), 1403–1448.

Jones, C. I. and Kim, J. (2014), 'A Schumpeterian Model of Top Income Inequality'.

Jovanovic, B. (1982), 'Selection and the Evolution of Industry', *Econometrica: Journal of the Econometric Society* **50**(3), 649–670.

Katz, L. and Murphy, K. (1992), 'Changes in Relative Wages, 1963-1987 : Supply and Demand Factors', *Quarterly Journal of Economics* **107**(1), 35–78.

Kesten, H. (1973), 'Random Difference Equations and Renewal Theory for Products of Random Matrices', *Acta Mathematica* **131**(1), 207–248.

Kremer, M. (1993), 'The O-Ring Theory of Economic Development', *The Quarterly Journal of Economics* **108**(3), 551–575.

Kremer, M. and Maskin, E. (1996), 'Wage Inequality and Segregation by Skill', *NBER Working Paper 5718* .

Krusell, P., Ohanian, L. E., Rios-Rull, J.-V. and Violante, G. L. (2000), 'Capital-Skill Complementarity and Inequality: A Macroeconomic Analysis', *Econometrica, Journal of the Econometric Society* **68**(5), 1029–1053.

Levy, M. (2009), 'Gibrat's Law for (All) Cities: Comment', *American Economic Review* **99**(4), 1672–1675.

Liegey, M. (2014), 'Search Frictions, the Use of Knowledge and the Labor Market', *Toulouse School of Economics, mimeo* .

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

Luttmer, E. (2007), 'Selection, Growth, and the Size Distribution of Firms', *Quarterly Journal of Economics* (August), 1103–1144.

Luttmer, E. G. (2010), 'Models of Growth and Firm Heterogeneity', *Annual Review of Economics* **2**(1), 547–576.

Mansfield, E. (1962), 'Entry, Gibrat's Law, Innovation, and the Growth of Firms', *American Economic Review* **151**(3712), 867–8.

Melitz, M. (2003), 'The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity', *Econometrica: Journal of the Econometric Society* **71**(6), 1695–1725.

Mincer, J. (1958), 'Investment in Human Capital and Personal Income Distribution', *The Journal of Political Economy* **66**(4), 281–302.

Mirrlees, J. (1971), 'An Exploration in the Theory of Optimum Income Taxation', *The Review of Economic Studies* **38**(2), 175–208.

Pareto, V. (1896), La Courbe de la Répartition de la Richesse, *in* 'Ecrits sur la Courbe de la Répartition de la Richesse', pp. 1–15.

Pigou, A. C. (1920), *The Economics of Welfare*, Macmillian and Co, Limited.

Piketty, T. (2014), *Capital in the Twenty-first Century*, Harvard University Press, Cambridge.

Piketty, T. and Saez, E. (2003), 'Income Inequality in the United States, 1913-1998', *Quarterly Journal of Economics* **118**(1), 1–39.

Radner, R. (1992), 'Hierarchy: The Economics of Managing', *Journal of Economic Literature* **30**(3), 1382–1415.

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

Roberts, D. (1956), 'A General Theory of Executive Compensation Based on Statistically Tested Propositions', *Quarterly Journal of Economics* **70**(2), 270–294.

Rosen, S. (1974), 'Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition', *Journal of Political Economy* **82**(1), 34–55.

Rosen, S. (1981), 'The Economics of Superstars', *American Economic Review* **71**(11), 845–858.

Rosen, S. (1982), 'Authority, Control, and the Distribution of Earnings', *The Bell Journal of Economics* **13**, 311–323.

Rossi-Hansberg, E. and Wright, M. (2007), 'Establishment Size Dynamics in the Aggregate Economy', *American Economic Review* **97**(5), 1640–1666.

Saez, E. (2001), 'Using Elasticities to Derive Optimal Income Tax Rates', *Review of Economic Studies* **68**(1), 205–229.

Sattinger, M. (1975), 'Comparative Advantage and the Distributions of Earnings and Abilities', *Econometrica: Journal of the Econometric Society* **43**(3), 455–468.

Simon, H. (1955), 'On a Class of Skew Distribution Functions', *Biometrika* **42**(3), 425–440.

Simon, H. and Bonini, C. (1958), 'The Size Distribution of Business Firms', *American Economic Review* **48**(4), 607–617.

Sutton, J. (1997), 'Gibrat's Legacy', *Journal of Economic Literature* **35**(March), 40–59.

Terviö, M. (2008), 'The Difference that CEOs Make: An Assignment Model Approach', *American Economic Review* **2**(98), 642–668.

Teulings, C. (1995), 'The Wage Distribution in a Model of the Assignment of Skills to Jobs', *Journal of Political Economy* **103**(2), 280–315.

Tinbergen, J. (1956), 'On the Theory of Income Distribution', *Weltwirtschaftliches Archiv* **77**, 155–175.

Tinbergen, J. (1974), 'Substitution of Graduate by Other Labour', *Kyklos* .

Tuomala, M. (1990), *Optimal Income Tax and Redistribution*, Oxford Clarendon Press.

Zipf, G. K. (1949), *Human Behavior and the Principle of Least Effort*, Cambridge, UK.

A Computational Appendix

A.1 Calculating the Cutoffs $\{z_l^L\}_{l=1}^L$ and the Matching Functions $\{m_l(.)\}_{l=0}^{L-1}$

From market clearing for time, the differential equations for the matching functions are:

$$\forall l \in \{0, ..., L-1\}, \quad \forall x \in [z_l^L, z_{l+1}^L], \quad m_l'(x)g(m_l(x)) = h(1-x)g(x) \quad \text{and} \quad m_l(z_l^L) = z_{l+1}^L.$$

Integrating, this is equivalent to:

$$\forall l \in \{0, ..., L-1\}, \quad \forall x \in [z_l^L, z_{l+1}^L], \quad m_l(x) = G^{-1}\left[G(z_{l+1}^L) + \int_{z_l^L}^x h(1-u)g(u)du\right].$$

The computational procedure follows the uniqueness proof. Taking z_1^L as given, all the remaining $\{z_l^L\}_{l=2}^L$ can be expressed recursively as a function of z_1^L using the above equations as well as:

$$\forall l \in \{0, ..., L-1\}, \quad z_{l+1}^L = m_l\left(z_{l+1}^L\right).$$

In particular, this gives a final value for $z_{L+1}^L(z_1^L)$ as a function of z_1^L, which defines z_1^L implicitly through:

$$z_{L+1}^L(z_1^L) = 1.$$

The matching functions and cutoffs are then expressed as a function of z_1^L, and all the primitives in the model.

Polynomial Functions. In the case of the family of polynomial functions, these are even expressed in closed form as a function of z_1^L because:

$$\begin{aligned}\int_{1-\Delta}^x (1-u)g(u)du &= \frac{\rho+1}{\Delta^{\rho+1}}\int_{1-\Delta}^x (1-u)^{\rho+1}du \\ &= \frac{\rho+1}{\Delta^{\rho+1}}\left[-\frac{(1-u)^{\rho+2}}{\rho+2}\right]_{1-\Delta}^x \\ H(x) \equiv \int_{1-\Delta}^x (1-u)g(u)du &= \frac{\rho+1}{\rho+2}\Delta\left[1 - \frac{(1-x)^{\rho+2}}{\Delta^{\rho+2}}\right].\end{aligned}$$

The inverse of this function is also expressed in closed form:

$$H^{-1}(x) = 1 - \Delta\left(1 - \frac{\rho+2}{\Delta(\rho+1)}\right)^{\frac{1}{\rho+2}}.$$

Finally, g and G are also closed form expressions as well as $G^{-1}(.)$:

$$G^{-1}(x) = 1 - \Delta(1-x)^{\frac{1}{\rho+1}}.$$

Span of control distributions are similarly expressed in closed form in that case, as the inverse of the matching functions are given by:

$$\begin{aligned}G(x) - G\left(z_l^L\right) &= h\left(H\left(m_{l-1}^{-1}(x)\right) - H\left(z_{l-1}^L\right)\right) \\ \Rightarrow \quad m_{l-1}^{-1}(x) &= H^{-1}\left[H\left(z_{l-1}^L\right) + \frac{G(x) - G\left(z_l^L\right)}{h}\right].\end{aligned}$$

Computationally speaking, solving the competitive equilibrium completely thus only relies on finding one scalar $\{z_1^L\}$, which determine the allocations completely. Allocations, matching functions and span of control distributions are even given in closed form in the case of polynomial functions.

A.2 Calculating Initial Conditions of the Wage Functions $\{a_l\}_{l=0}^{L-1}$, and the Return Functions $\{R_l\}_{l=0}^{L}$

Denote the initial conditions for the wage functions as $\{a_l\}_{l=0}^{L-1}$, that is:

$$\forall l \in \{0, ..., L-1\}, \quad a_l = w_l\left(z_l^L\right),$$

The wage functions are entirely defined recursively in integral form as a function of $\{a_l\}_{l=0}^{L-1}$ as well as $\{z_l^L\}_{l=1}^{L}$ through:

$$\forall x \in [z_l^L, z_{l+1}^L], \quad w_l(x_l) = \frac{1}{N_{l+1\to 0}^l(x_l)} \int_{z_l^L}^{x_l} w_{l+1}\left(m_l(u)\right) \frac{dN_{l+1\to 0}^l}{dx_l}(u)du + \frac{N_{l+1\to 0}^l(z_l^L)}{N_{l+1\to 0}^l(x_l)} w_l(z_l^L)$$
$$\text{s.t.} \quad w_l(z_l^L) = a_l.$$

For symmetry the wage function of manager of type L is here defined as $w_L(x_l) = F(x_l) = x_l$. By backward recursion, starting from given by the identity function, one finds the whole sequence of wage functions $\{w_l(.)\}_{l=0}^{L-1}$ as a function of the $\{a_l\}_{l=0}^{L-1}$. The $\{a_l\}_{l=0}^{L-1}$ are in turn given by the $1 + L - 1 = L$ indifference equations:

$$R_1(z_1^L) = w_0(z_1^L),$$
$$\forall l \in \{1, ..., L-1\}, \quad R_{l+1}(z_{l+1}^L) = R_l(z_{l+1}^L).$$

Denote by $\{b_l\}_{l=0}^{L-1}$ the following quantities:

$$\forall l \in \{0, ..., L-1\}, \quad b_l = w_l\left(z_{l+1}^L\right).$$

Then the L previous equations simplify as follows:

$$R_1(z_1^L) = w_0(z_1^L) \quad \Rightarrow \quad b_0 = (a_1 - a_0) N_{1\to 0}^0\left(z_0^L\right).$$
$$\forall l \in \{1, ..., L-1\}, \quad R_{l+1}(z_{l+1}^L) = R_l(z_{l+1}^L)$$
$$\Rightarrow \quad \left(w_{l+1}\left(z_{l+1}^L\right) - w_l\left(z_l^L\right)\right) N_{l+1\to 0}^l(z_l^L) = \left(w_l\left(z_{l+1}^L\right) - w_{l-1}\left(z_l^L\right)\right) N_{l\to 0}^{l-1}(z_l^L)$$
$$\Rightarrow \quad (a_{l+1} - a_l)\, N_{l+1\to 0}^l(z_l^L) = (b_l - b_{l-1})\, N_{l\to 0}^{l-1}(z_l^L),$$

Note that for symmetry, I have defined a_L as the initial condition of the hypothetical wage function of managers of type L, equal to $F(.)$ so that $a_L = w_L(z_L^L) = z_L^L$.

The numbers $\{b_l\}_{l=0}^{L-1}$ are given as a function of $\{a_l\}_{l=0}^{L-1}$ by using the above differential equations between $[z_l^L, z_{l+1}^L]$ so that:

$$\forall l \in \{0, ..., L-1\}, \quad N_{l+1\to 0}^l(z_{l+1}^L) b_l = \int_{z_l^L}^{z_{l+1}^L} w_{l+1}\left(m_l(u)\right) \frac{dN_{l+1\to 0}^l}{dx_l}(u)du + N_{l+1\to 0}^l(z_l^L) a_l.$$

By recursion:

$$N_{L\to 0}^{L-1}\left(z_L^L\right) b_{L-1} = \int_{z_{L-1}^L}^{z_L^L} m_{L-1}(t)\frac{dN_{L\to 0}^{L-1}}{dx_{L-1}}(t)dt + N_{L\to 0}^{L-1}(z_{L-1}^L)a_{L-1}$$

$$N_{L-1\to 0}^{L-2}\left(z_{L-1}^L\right) b_{L-2} = \int_{z_{L-2}^L}^{z_{L-1}^L} w_{L-1}\left(m_{L-2}(v)\right)\frac{dN_{L-1\to 0}^{L-2}}{dx_{L-2}}(v)dv + N_{L-1\to 0}^{L-2}(z_{L-2}^L)a_{L-2}$$

With:

$$w_{L-1}\left(m_{L-2}(v)\right) = \frac{1}{N_{L\to 0}^{L-1}\left(m_{L-2}(v)\right)}\int_{z_{L-1}^L}^{m_{L-2}(v)} m_{L-1}(t)\frac{dN_{L\to 0}^{L-1}}{dx_{L-1}}(t)dt + \frac{1}{N_{L\to 0}^{L-1}\left(m_{L-2}(v)\right)}N_{L\to 0}^{L-1}(z_{L-1}^L)a_{L-1}.$$

Therefore:

$$N_{L-1\to 0}^{L-2}\left(z_{L-1}^L\right) b_{L-2} = \int_{z_{L-2}^L}^{z_{L-1}^L}\frac{1}{N_{L\to 0}^{L-1}\left(m_{L-2}(v)\right)}\int_{z_{L-1}^L}^{m_{L-2}(v)} m_{L-1}(t)\frac{dN_{L\to 0}^{L-1}}{dx_{L-1}}(t)dt\frac{dN_{L-1\to 0}^{L-2}}{dx_{L-2}}(v)dv + ...$$
$$... \left[\int_{z_{L-2}^L}^{z_{L-1}^L}\frac{1}{N_{L\to 0}^{L-1}\left(m_{L-2}(v)\right)}\frac{dN_{L-1\to 0}^{L-2}}{dx_{L-2}}(v)dv\right] N_{L\to 0}^{L-1}(z_{L-1}^L)a_{L-1} + N_{L-1\to 0}^{L-2}(z_{L-2}^L)a_{L-2}.$$

Finally:

$$N_{L-2\to 0}^{L-3}\left(z_{L-2}^L\right) b_{L-3} = \int_{z_{L-3}^L}^{z_{L-2}^L} w_{L-2}\left(m_{L-3}(y)\right)\frac{dN_{L-2\to 0}^{L-3}}{dx_{L-3}}(y)dy + N_{L-2\to 0}^{L-3}(z_{L-3}^L)a_{L-3}.$$

With:

$$w_{L-2}\left(m_{L-3}(y)\right) = \frac{1}{N_{L-1\to 0}^{L-2}\left(m_{L-3}(y)\right)}\int_{z_{L-2}^L}^{m_{L-3}(y)} w_{L-1}\left(m_{L-2}(v)\right)\frac{dN_{L-1\to 0}^{L-2}}{dx_{L-2}}(v)dv + ...$$
$$...\frac{1}{N_{L-1\to 0}^{L-2}\left(m_{L-3}(y)\right)}N_{L-1\to 0}^{L-2}(z_{L-2}^L)a_{L-2}.$$

Therefore:

$$N_{L-2\to 0}^{L-3}\left(z_{L-2}^L\right) b_{L-3} = \int_{z_{L-3}^L}^{z_{L-2}^L}\frac{1}{N_{L-1\to 0}^{L-2}\left(m_{L-3}(y)\right)}\int_{z_{L-2}^L}^{m_{L-3}(y)} w_{L-1}\left(m_{L-2}(v)\right)\frac{dN_{L-1\to 0}^{L-2}}{dx_{L-2}}(v)dv...$$
$$...\frac{dN_{L-2\to 0}^{L-3}}{dx_{L-3}}(y)dy + \left[\int_{z_{L-3}^L}^{z_{L-2}^L}\frac{1}{N_{L-1\to 0}^{L-2}\left(m_{L-3}(y)\right)}\frac{dN_{L-2\to 0}^{L-3}}{dx_{L-3}}(y)dy\right] N_{L-1\to 0}^{L-2}(z_{L-2}^L)a_{L-2} + N_{L-2\to 0}^{L-3}(z_{L-3}^L)a_{L-3}.$$

Finally:

$$N_{L-2\to 0}^{L-3}\left(z_{L-2}^L\right) b_{L-3} = \int_{z_{L-3}^L}^{z_{L-2}^L}\frac{1}{N_{L-1\to 0}^{L-2}\left(m_{L-3}(y)\right)}\int_{z_{L-2}^L}^{m_{L-3}(y)}\frac{1}{N_{L\to 0}^{L-1}\left(m_{L-2}(v)\right)}...$$
$$...\int_{z_{L-1}^L}^{m_{L-2}(v)} m_{L-1}(t)\frac{dN_{L\to 0}^{L-1}}{dx_{L-1}}(t)dt\frac{dN_{L-1\to 0}^{L-2}}{dx_{L-2}}(v)dv\frac{dN_{L-2\to 0}^{L-3}}{dx_{L-3}}(y)dy+$$
$$+\left[\int_{z_{L-3}^L}^{z_{L-2}^L}\frac{1}{N_{L-1\to 0}^{L-2}\left(m_{L-3}(y)\right)}\int_{z_{L-2}^L}^{m_{L-3}(y)}\frac{1}{N_{L\to 0}^{L-1}\left(m_{L-2}(v)\right)}\frac{dN_{L-1\to 0}^{L-2}}{dx_{L-2}}(v)dv\frac{dN_{L-2\to 0}^{L-3}}{dx_{L-3}}(y)dy\right] N_{L\to 0}^{L-1}(z_{L-1}^L)a_{L-1}$$
$$\left[\int_{z_{L-3}^L}^{z_{L-2}^L}\frac{1}{N_{L-1\to 0}^{L-2}\left(m_{L-3}(y)\right)}\frac{dN_{L-2\to 0}^{L-3}}{dx_{L-3}}(y)dy\right] N_{L-1\to 0}^{L-2}(z_{L-2}^L)a_{L-2} + N_{L-2\to 0}^{L-3}(z_{L-3}^L)a_{L-3}.$$

The first closest terms to the diagonal of $\{\lambda_i\}_{i<j=0}^{L-1}$ and the first terms of $\{\mu_l\}_{l=0}^{L-1}$ can therefore be written as follows:

$$\mu_{L-1} = \int_{z_{L-1}^L}^{z_L^L} m_{L-1}(t) \frac{dN_{L\to 0}^{L-1}}{dx_{L-1}}(t)dt$$

$$\mu_{L-2} = \int_{z_{L-2}^L}^{z_{L-1}^L} \frac{1}{N_{L\to 0}^{L-1}\left(m_{L-2}(v)\right)} \frac{dN_{L-1\to 0}^{L-2}}{dx_{L-2}}(v)dv \int_{z_{L-1}^L}^{m_{L-2}(v)} m_{L-1}(t) \frac{dN_{L\to 0}^{L-1}}{dx_{L-1}}(t)dt$$

$$\mu_{L-3} = \int_{z_{L-3}^L}^{z_{L-2}^L} \frac{1}{N_{L-1\to 0}^{L-2}\left(m_{L-3}(y)\right)} \int_{z_{L-2}^L}^{m_{L-3}(y)} \frac{1}{N_{L\to 0}^{L-1}\left(m_{L-2}(v)\right)} \cdots$$
$$\int_{z_{L-1}^L}^{m_{L-2}(v)} m_{L-1}(t) \frac{dN_{L\to 0}^{L-1}}{dx_{L-1}}(t)dt \frac{dN_{L-1\to 0}^{L-2}}{dx_{L-2}}(v)dv \frac{dN_{L-2\to 0}^{L-3}}{dx_{L-3}}(y)dy$$

$$\lambda_{L-2L-1} = \int_{z_{L-2}^L}^{z_{L-1}^L} \frac{1}{N_{L\to 0}^{L-1}\left(m_{L-2}(v)\right)} \frac{dN_{L-1\to 0}^{L-2}}{dx_{L-2}}(v)dv$$

$$\lambda_{L-3L-2} = \int_{z_{L-3}^L}^{z_{L-2}^L} \frac{1}{N_{L-1\to 0}^{L-2}\left(m_{L-3}(y)\right)} \frac{dN_{L-2\to 0}^{L-3}}{dx_{L-3}}(y)dy$$

$$\lambda_{L-3L-1} = \int_{z_{L-3}^L}^{z_{L-2}^L} \frac{1}{N_{L-1\to 0}^{L-2}\left(m_{L-3}(y)\right)} \int_{z_{L-2}^L}^{m_{L-3}(y)} \frac{1}{N_{L\to 0}^{L-1}\left(m_{L-2}(v)\right)} \frac{dN_{L-1\to 0}^{L-2}}{dx_{L-2}}(v)dv \frac{dN_{L-2\to 0}^{L-3}}{dx_{L-3}}(y)dy.$$

Example with $L = 4$. The above equations consist of L linear equations in $\{a_l\}_{l=0}^{L-1}$. Calculating the coefficients of these linear equations involves a computation of $L(L+1)/2$ integral quantities $\{\lambda_i\}_{i<j=0}^{L-1}$ and $\{\mu_l\}_{l=0}^{L-1}$, which give $\{b_l\}_{l=0}^{L-1}$ as a function of $\{a_l\}_{l=0}^{L-1}$ through:

$$\begin{bmatrix} N_{1\to 0}^0\left(z_1^L\right) b_0 \\ N_{2\to 0}^1\left(z_2^L\right) b_1 \\ N_{3\to 0}^2\left(z_3^L\right) b_2 \\ N_{4\to 0}^3\left(z_4^L\right) b_3 \end{bmatrix} = \begin{bmatrix} 1 & \lambda_{01} & \lambda_{02} & \lambda_{03} \\ 0 & 1 & \lambda_{12} & \lambda_{13} \\ 0 & 0 & 1 & \lambda_{23} \\ 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} N_{1\to 0}^0\left(z_0^L\right) a_0 \\ N_{2\to 0}^1\left(z_1^L\right) a_1 \\ N_{3\to 0}^2\left(z_2^L\right) a_2 \\ N_{4\to 0}^3\left(z_3^L\right) a_3 \end{bmatrix} + \begin{bmatrix} \mu_0 \\ \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}.$$

With the $\{\mu_l\}_{l=0}^{L-1}$ as follows:

$$\mu_0 = \iiint_{\substack{x_3\in[z_3,m_2(x_2)] \\ x_2\in[z_2,m_1(x_1)] \\ x_1\in[z_1,m_0(x_0)] \\ x_0\in[z_0,z_1]}} \frac{1}{N_{2\to 0}^1\left(m_0(x_0)\right)} \frac{dN_{1\to 0}^0}{dx_0}(x_0) \frac{1}{N_{3\to 0}^2\left(m_1(x_1)\right)} \frac{dN_{2\to 0}^1}{dx_1}(x_1) \cdots$$
$$\cdots \frac{1}{N_{4\to 0}^3\left(m_2(x_2)\right)} \frac{dN_{3\to 0}^2}{dx_2}(x_2) m_3(x_3) \frac{dN_{4\to 0}^3}{dx_3}(x_3) dx_3 dx_2 dx_1 dx_0$$

$$\mu_1 = \iiint_{\substack{x_3\in[z_3,m_2(x_2)] \\ x_2\in[z_2,m_1(x_1)] \\ x_1\in[z_1,z_2]}} \frac{1}{N_{3\to 0}^2\left(m_1(x_1)\right)} \frac{dN_{2\to 0}^1}{dx_1}(x_1) \frac{1}{N_{4\to 0}^3\left(m_2(x_2)\right)} \frac{dN_{3\to 0}^2}{dx_2}(x_2) m_3(x_3) \frac{dN_{4\to 0}^3}{dx_3}(x_3) dx_3 dx_2 dx_1$$

$$\mu_2 = \iint_{\substack{x_3\in[z_3,m_2(x_2)] \\ x_2\in[z_2,z_3]}} \frac{1}{N_{4\to 0}^3\left(m_2(x_2)\right)} \frac{dN_{3\to 0}^2}{dx_2}(x_2) m_3(x_3) \frac{dN_{4\to 0}^3}{dx_3}(x_3) dx_3 dx_2$$

$$\mu_3 = \int_{z_3}^{z_4} m_3(x_3) \frac{dN_{4\to 0}^3}{dx_3}(x_3) dx_3$$

And the closest terms to the diagonal ($j-i=1$) in $\{\lambda_i\}_{i<j=0}^{L-1}$ are given by:

$$\lambda_{01} = \int_{z_0}^{z_1} \frac{1}{N^1_{2\to 0}\left(m_0(x_0)\right)} \frac{dN^0_{1\to 0}}{dx_0}(x_0)dx_0 \qquad \lambda_{12} = \int_{z_1}^{z_2} \frac{1}{N^2_{3\to 0}\left(m_1(x_1)\right)} \frac{dN^1_{2\to 0}}{dx_1}(x_1)dx_1$$

$$\lambda_{23} = \int_{z_2}^{z_3} \frac{1}{N^3_{4\to 0}\left(m_2(x_2)\right)} \frac{dN^2_{3\to 0}}{dx_2}(x_2)dx_2$$

The next closest terms ($j-i=2$) are:

$$\lambda_{02} = \iint_{\substack{x_1\in[z_1,m_0(x_0)]\\ x_0\in[z_0,z_1]}} \frac{1}{N^1_{2\to 0}\left(m_0(x_0)\right)} \frac{dN^0_{1\to 0}}{dx_0}(x_0) \frac{1}{N^2_{3\to 0}\left(m_1(x_1)\right)} \frac{dN^1_{2\to 0}}{dx_1}(x_1)dx_1dx_0$$

$$\lambda_{13} = \iint_{\substack{x_2\in[z_2,m_1(x_1)]\\ x_1\in[z_1,z_2]}} \frac{1}{N^2_{3\to 0}\left(m_1(x_1)\right)} \frac{dN^1_{2\to 0}}{dx_1}(x_1) \frac{1}{N^3_{4\to 0}\left(m_2(x_2)\right)} \frac{dN^2_{3\to 0}}{dx_2}(x_2)dx_2dx_1$$

Finally, the term further away from the diagonal ($j-i=2$) is:

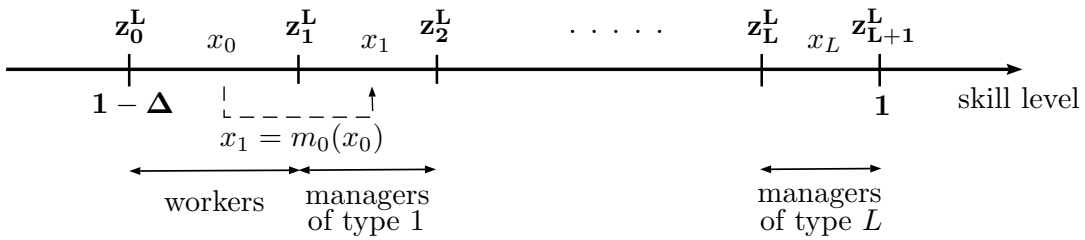
$$\lambda_{03} = \iiint_{\substack{x_2\in[z_2,m_1(x_1)]\\ x_1\in[z_1,m_0(x_0)]\\ x_0\in[z_0,z_1]}} \frac{1}{N^1_{2\to 0}\left(m_0(x_0)\right)} \frac{dN^0_{1\to 0}}{dx_0}(x_0) \frac{1}{N^2_{3\to 0}\left(m_1(x_1)\right)} \frac{dN^1_{2\to 0}}{dx_1}(x_1)...$$

$$...\frac{1}{N^3_{4\to 0}\left(m_2(x_2)\right)} \frac{dN^2_{3\to 0}}{dx_2}(x_2)dx_2dx_1dx_0.$$

Remember that for uniform distributions, the matching functions have closed form solutions. $\{a_l\}_{l=0}^{L-1}$ obtain then very easily from the preceding, also in closed form, apart from z_1^L. And wage functions can be calculated by inverse recursion beginning by $w_L(.) = F(.)$ through the previous expressions.

B Proofs

B.1 Proof of Lemma 1



To prove that there is no self-employment in equilibrium requires to use the notations which are later used to solve the model in the paper (reminded above). The fact that self-employment does not arise when communication costs are sufficiently low is quite intuitive, as h governs the strength of the complementarities in this model: when h is low, it becomes valuable for agents with even very close skills to form teams and work together. However, the behavior with respect to Δ is perhaps less intuitive.

Of course, there are some gains to forming teams of managers and workers: when a worker with low skill does not know a solution to a problem, he asks a manager with a higher skill which

increases the probability that the problem is going to be solved, and therefore output in the economy. However, there is also a cost: the managers spend some time away from production. For it to be profitable, the formation of a team of worker with skill x_0 together with managers of type l with type x_l it has to be that the first term is higher than the second. The gains are decomposed as follows:

$$\underbrace{x_L g(x_0)dx_0}_{\text{Output together}} - \underbrace{\sum_{l=0}^{L} x_l g(x_l)}_{\text{Output alone}} = \underbrace{(x_L - x_0)g(x_0)dx_0}_{\text{"Increased Productivity" of workers}} - \underbrace{\sum_{l=1}^{L} x_l g(x_l)}_{\text{"Lost production time" of managers}}.$$

Summing over the interval of workers, the total gains are therefore given by:

$$\int_{z_0^L}^{z_1^L} \left(m_{L-1} \circ m_{L-2} \circ ... \circ m_0(x_0) - x_0\right) g(x_0)dx_0 - \sum_{l=1}^{L} \int_{z_l^L}^{z_{l+1}^L} x_l g(x_l)dx_l.$$

I show later that when $\Delta \to 0$, $m_{L-1} \circ m_{L-2} \circ ... \circ m_0(.)$ converges to 1 in a Δ^2 term. The first term therefore is greater than a linear term in Δ, while the second is smaller than a quadratic term:

$$\sum_{l=1}^{L} \int_{z_l^L}^{z_{l+1}^L} x_l g(x_l)dx_l \leq \sum_{l=1}^{L} \int_{z_l^L}^{z_{l+1}^L} g(x_l)dx_l = 1 - G\left(z_1^L\right).$$

Intuitively, when heterogeneity goes to 0, there really does not need a lot of managers to increase the productivity of all workers. The lost production time is becoming infinitesimally small relative to the increased gains from trade.

$$\int_i t_W^i(x_l)di = 0 \quad \Rightarrow \quad g(m_l(x_l))dm_l(x_l) = h\left(1 - x_l\right) g(x_l)dx_l.$$
$$\Rightarrow \quad m_l'(x_l)g(m_l(x_l)) = h\left(1 - x_l\right) g(x_l). \qquad (M_l)$$

The positive sorting limiting condition writes, matching the less skilled and the more skilled:

$$m_l(z_l^L) = z_{l+1}^L$$
$$m_l(z_{l+1}^L) = z_{l+2}^L$$

Integrating the above differential equation for $m_l(.)$ between $[z_l^L, z_{l+1}^L]$, and using these two limiting conditions give the result.

Note that the stratification result obtains irrespective of the skill distribution. This is in contrast to Kremer and Maskin (1996) - see Garicano and Rossi-Hansberg (2006) for a discussion.

B.2 Proof of Lemma 2

[TO BE ADDED]

B.3 Self-Employment or No Self-Employment? Closed Form Expression for Assumption 2 (Uniform Density)

In the case where h or Δ are high enough, some agents remain self-employed. Self-employed agents have intermediary skills in equilibrium, because the gains from having a worker of skill x_0 and a manager of skill x_1 work together are given by what the two produce together minus what they would have produced by themselves, that is:

$$\frac{x_1}{h(1-x_0)} - x_1 - \frac{x_0}{h\,(1-x_0)} = \frac{1}{h}\left(1 - \frac{1-x_1}{1-x_0}\right) - x_1.$$

This is clearly decreasing when the skills of workers increase, so that it is better to match the managers with the relatively less productive workers.

Figure 28: With Self-Employment in Equilibrium, One Layer.

The notations for cutoffs are introduced on Figure 28. Now the matching function $m_0(.)$ is defined on $[1-\Delta, z_1^{**}]$. An important difference also is that in that case, allocations are not solved independently from agents' choices.

In the decentralized problem, the two differential equations for $m_0(.)$ and $w_0(.)$ do not change compared to the case of no-self employment. However I now have four equations, not three, determining two initial conditions as well as two cutoffs. They are given by matching of the less and more skilled of workers and team managers, as previously:

$$m_0(1-\Delta) = z_1^* \qquad m_0(z_1^{**}) = 1.$$

In addition, I now have two indifference equations between being a worker and self-employed with skills z_1^{**}, and being self-employed and a team manager with skills z_1^*:

$$w_0(z_1^{**}) = z_1^{**} \qquad z_1^* = R_1(z_1^*).$$

In the case of a uniform distribution of skills, the market clearing equation for skills valid on $(1-\Delta, z_1^{**})$, together with the terminal equation $m_0(z_1^{**}) = 1$, then gives:

$$\begin{aligned} m_0'(x_0) g\left(m_0(x_0)\right) = h\left(1-x_0\right) g(x_0) \quad &\Rightarrow \quad m_0'(x_0) = h(1-x_0) \\ \Rightarrow \quad m_0(x_0) &= \frac{1}{2}\left(-hx_0^2 + 2hx_0 + h\left(z_1^{**}\right)^2 - 2hz_1^{**} + 2\right). \end{aligned}$$

Inverting this expression, the inverse assignment function is therefore given by:

$$m_0^{-1}(x_1) = \frac{h - \sqrt{2h + h^2 - 2hx_1 - 2h^2 z_1^{**} + h^2 (z_1^{**})^2}}{h}.$$

In the case where self-employment arises in equilibrium, one must solve for the equilibrium wage function even to determine the spans of control of each team manager. One also uses $w_0(z_1^{**}) = z_1^{**} = z_1^{**}$ to integrate:

$$\begin{aligned}(1-x_0)\,w_0'(x_0) + x_0 w_0(x_0) &= x_0 m_0(x_0)\\ \Rightarrow \quad w_0(x_0) &= \frac{1}{2}\left(2x_0 + hx_0^2 - 2hx_0 z_1^{**} + h\,(z_1^{**})^2\right)\end{aligned}$$

Then using the two remaining $m_0(1-\Delta) = z_1^*$ and $z_1^* = R_1(z_1^*)$, and simple but lengthy algebra, one can express the cutoffs for occupational choice as a function of the heterogeneity parameter Δ and the helping time h:

$$z_1^* = -\frac{-2h + h^2\Delta + \sqrt{h^2\left(3 + h^2\Delta^2 - 2h(1+\Delta)\right)}}{h^2} \qquad z_1^{**} = \frac{-h + h^2 + \sqrt{h^2\left(3 + h^2\Delta^2 - 2h(1+\Delta)\right)}}{h^2}.$$

Replacing the cutoffs, one gets the assignment function as a function as these parameters as well:

$$m_0(x_0) = \frac{4h - 2h^2\Delta + h^3\left(-1+\Delta^2\right) - 2\sqrt{h^2\left(3 + h^2\Delta^2 - 2h(1+\Delta)\right)} + 2h^3x_0 - h^3x_0^2}{2h^2}.$$

The condition for there to be self-employment in equilibrium is that:

$$\begin{aligned}z_1^{**} < z_1^* \quad &\Leftrightarrow \quad \frac{-h + h^2 + \sqrt{h^2\left(3 + h^2\Delta^2 - 2h(1+\Delta)\right)}}{h^2} < -\frac{-2h + h^2\Delta + \sqrt{h^2\left(3 + h^2\Delta^2 - 2h(1+\Delta)\right)}}{h^2}\\ &\Leftrightarrow \quad 3 - h\Delta - h > 2\sqrt{3 + h^2\Delta^2 - 2h(1+\Delta)}\\ &\Leftrightarrow \quad (1 + 2\Delta - 3\Delta^2)h^2 + 2(1+\Delta)h - 3 > 0\\ &\Leftrightarrow \quad h > \frac{1 + \Delta - 2\sqrt{1 + 2\Delta - 2\Delta^2}}{-1 - 2\Delta + 3\Delta^2} \quad \text{since } h > 0.\end{aligned}$$

When the primitives of the model are such that this is verified, we are in the case where a non trivial measure of agents are self-employed. When it is not the case, then all agents either become managers or workers. The condition is illustrated graphically on Figure 7.

B.4 Alternative Proof of Proposition 6

In the main text, I provide a proof of Proposition 6 based on an envelope condition. Here is another proof, where the role of changes of variables perhaps appears more clearly.

Proof. The first result follows straightforwardly from the second, which is more general: set $l = L-1$ and use $w_L(x_L) = x_L$ so that $w_L'(x_L) = 1$ for all $x_L \in [z_L^L, 1]$. To prove the second result, we need to integrate result in Lemma 7 by parts:

$$\int_{z_l^L}^{x_l} w_{l+1}\left(m_l(u)\right)\frac{dN_{l+1\to 0}^l}{dx_l}(u)du = \left[w_{l+1}\left(m_l(u)\right)N_{l+1\to 0}^l(u)\right]_{z_l^L}^{x_l} - \int_{z_l^L}^{x_l} m_l'(u)w_{l+1}'\left(m_l(u)\right)N_{l+1\to 0}^l(u)du.$$

Using then the correspondance $N_{l+1\to 0}^l = N_{l+1\to 0}^{l+1}\circ m_l$, as well as the change in variable $v = m_l(u)$:

$$\int_{z_l^L}^{x_l} m_l'(u)w_{l+1}'\left(m_l(u)\right)N_{l+1\to 0}^l(u)du = \int_{z_{l+1}^L}^{x_{l+1}} w_{l+1}'(v)N_{l+1\to 0}^{l+1}(v)dv.$$

Now, replacing $w_l(x_l)$ in $R_{l+1}(x_{l+1})$:

$$\begin{aligned} R_{l+1}(x_{l+1}) &= (w_{l+1}(x_{l+1}) - w_l(x_l)) N^l_{l+1\to 0}(x_l) \\ &= w_{l+1}(x_{l+1}) N^l_{l+1\to 0}(x_l) - w_l(x_l) N^l_{l+1\to 0}(x_l) \end{aligned}$$

$$\begin{aligned} R_{l+1}(x_{l+1}) &= w_{l+1}(x_{l+1}) N^l_{l+1\to 0}(x_l) - \int_{z_l^L}^{x_l} w_{l+1}(m_l(u)) \frac{dN^l_{l+1\to 0}}{dx_l}(u)du - w_l(z_l^L)N^l_{l+1\to 0}(z_l^L) \\ &= w_{l+1}(x_{l+1}) N^l_{l+1\to 0}(x_l) - \left[w_{l+1}(m_l(u)) N^l_{l+1\to 0}(u)\right]_{z_l^L}^{x_l} + ... \\ &\quad ... \int_{z_{l+1}^L}^{x_{l+1}} w'_{l+1}(v)N^{l+1}_{l+1\to 0}(v)dv - w_l(z_l^L)N^l_{l+1\to 0}(z_l^L). \end{aligned}$$

Using the previous integration by parts, this simplifies into:

$$R_{l+1}(x_{l+1}) = R_{l+1}\left(z^L_{l+1}\right) + \int_{z^L_{l+1}}^{x_{l+1}} w'_{l+1}(u)N^{l+1}_{l+1\to 0}(u)du.$$

□

B.5 Comparison with Gabaix and Landier (2008)

In this Appendix, I show how to map the findings in Section 3 to that in Gabaix and Landier (2008), where the notations and the model are a bit different to the one I use. The key equation in Gabaix and Landier (2008) is the assignment equation (equation (5), pp. 57), that the marginal cost of a slightly better CEO, is given by the marginal benefit of this slightly better CEO:

$$w'(n) = CS(n)^\gamma T'(n),$$

where n denotes the rank of the CEO, $w(n)$ is his compensation, $S(n)$ the size of the firm he is running, $T(n)$ the talent of the n-th best CEO. In my model with a continuum of types, the skill of a CEO is denoted by variable $x \in [1-\Delta, 1]$, and what corresponds to the rank of a CEO is given by the cumulative distribution function for skills $G(x)$. We have:

$$1 - G(x) \sim \frac{n}{N},$$

as a correspondance, where N corresponds to the total number of firms (and of CEOs) in Gabaix and Landier (2008). This just comes from the general point that if $Y = F_X(X)$ then $Y \sim \mathbb{U}(0,1)$, which is also true for the survivor function $Y = 1 - F_X(X)$. Therefore:

$$1 - G(x) \sim \frac{(1-x)^{\rho+1}}{\Delta^{\rho+1}} \quad \Rightarrow \quad 1 - T(n) = \Delta\left(\frac{n}{N}\right)^{\frac{1}{\rho+1}} \quad \Rightarrow \quad T'(n) \sim -\left(\frac{n}{N}\right)^{\frac{1}{\rho+1}-1}.$$

The correspondance between the talent distribution in my model and to Gabaix and Landier (2008)'s where $T(n) \sim -n^{\beta-1}$ is therefore given by $\beta = 1/(1+\rho)$. Note also that the microfoundations of the model say that $\gamma = 1$, which is consistent with what they estimate empirically. They then go on to show that if $S(n) \sim n^{-\alpha}$ (with $\alpha = 1$), then $w(n) \sim n^{-(\alpha\gamma-\beta)}$. Conditional on having $\alpha = 1$, the model presented gets the same distribution of CEO's incomes.

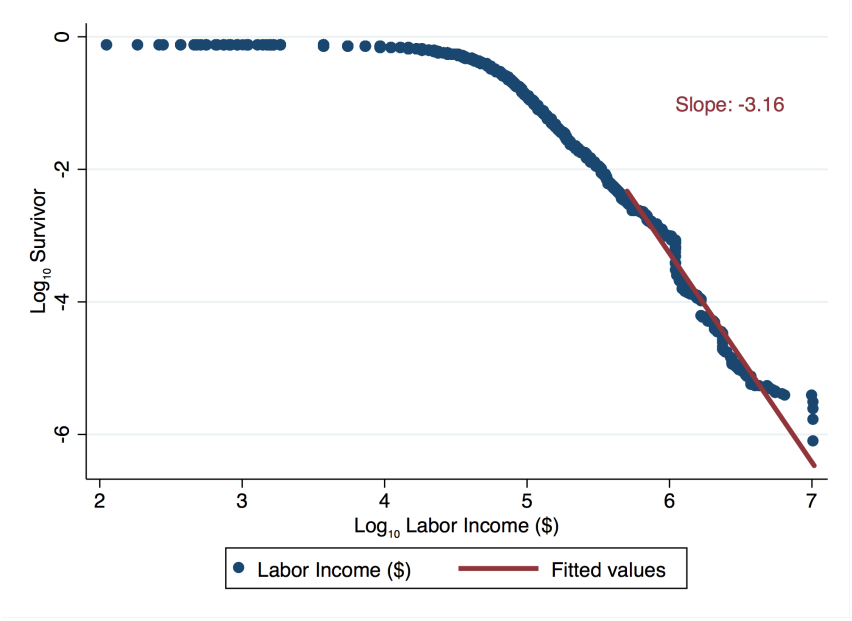
Section 4.2 explains what the advantages are of endogenizing Zipf's law for firms.

C Evidence

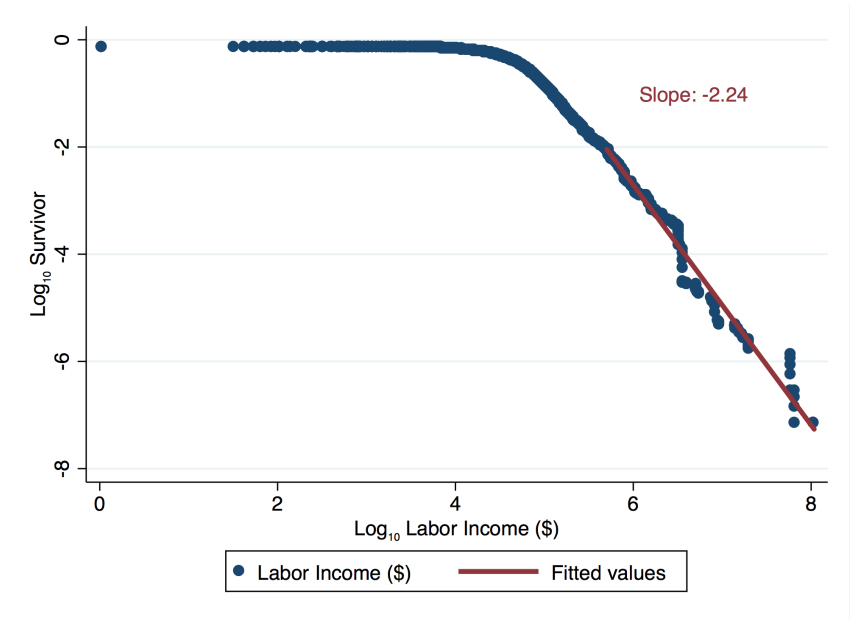
C.1 Wages - US Survey Data

Figure 29: Comparing the Labor Income Distribution in 1989 and 2010 in SCF

(a) 1989



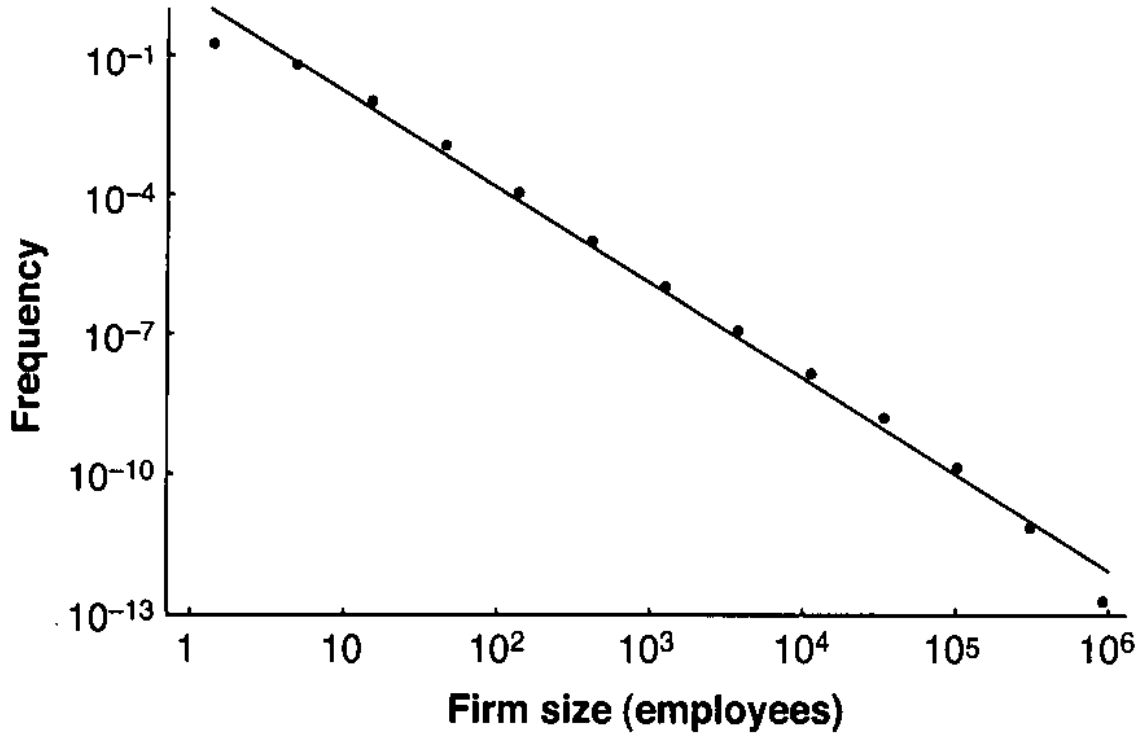
(b) 2010



Note: This graph must be interpreted with great caution, since surveys capture the top incomes very imperfectly.

C.2 Evidence from the Literature

Figure 30: Zipf's Law for US Firm Sizes, Log-Log Scale, Axtell (2001)



Note: In this figure, the density of firm sizes is represented. The estimated slope is 2.059 in the frequency domain, corresponding to a Pareto tail index of 1.059. The distribution of firms between 1 and 10 employees is more concave than Pareto, which has sparked a vigorous debate about whether the distribution was closer to a log-normal or to a Pareto (see Eeckhout (2004)). The static model I develop predicts a Zipf's law only in the upper tail (see Figure 18). Similarly, big firms are smaller than Zipf's law would predict, again a prediction of the model for non-zero heterogeneity.

D Linearly Increasing Distribution for Skills

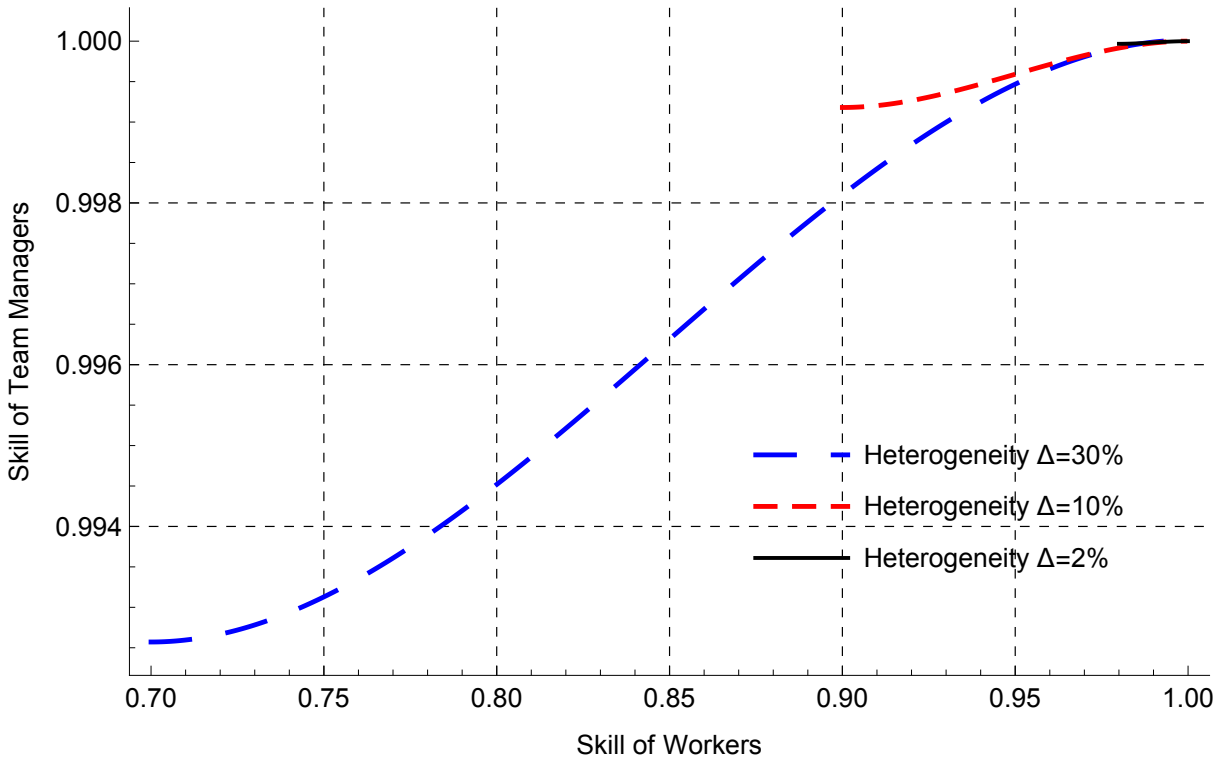
Because the results in the main paper hold for all sufficiently smooth distribution of skills, and that in particular the Pareto coefficients do not depend on them, it is useful to illustrate the equilibrium of the model with another example than the uniform distribution used throughout the main paper.

I illustrate this result using a linearly increasing distribution for the frequency of skills. In other words, relatively high skilled agents are relatively more numerous than relatively less skilled agents, so that:

$$G(z) = \frac{(z-(1-\Delta))^2}{\Delta^2}\mathbb{1}\left[1-\Delta,1\right](z) + \mathbb{1}\left[1,+\infty\right](z).$$

For example, one can see on Figure 31 some comparative statics with respect to heterogeneity.

Figure 31: Matching Function from Workers to Team Managers, $h = 70\%$ - Comparative Statics on Heterogeneity



Note: This figure gives the matching function from workers to team managers, with the assumption that the distribution of skills in the population is given by $G(z) = \frac{(z-(1-\Delta))^2}{\Delta^2}\mathbb{1}\left[1-\Delta,1\right](z) + \mathbb{1}\left[1,+\infty\right](z)$.

E Comparative Statics of the Top Labor Income Distribution

Figure 32: Communication Costs, $\Delta = 40\%$, $\rho = 1$

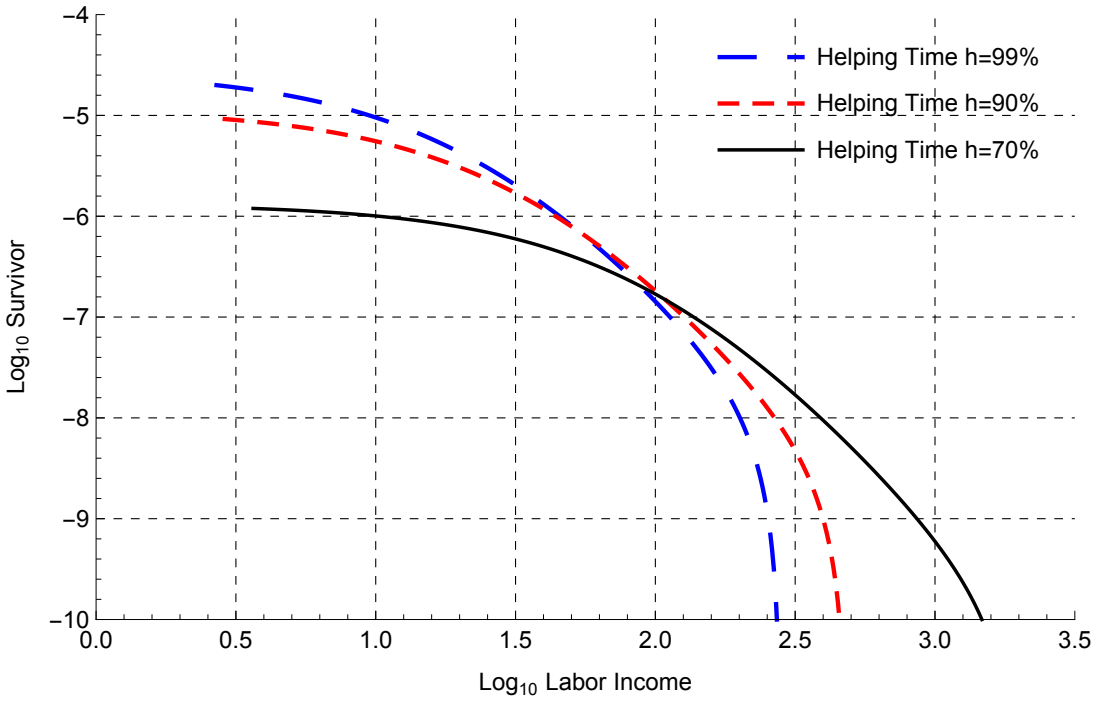


Figure 33: Scarcity of Managers ρ, $h = 90\%$, $\Delta = 10\%$