MECS 570-1: Organizational Economics I

Michael Powell

Organizational Structure: Knowledge Hierarchies Cont., Monitoring Hierarchies

Knowledge Hierarchies in Equilibrium

In this note, we continue our discussion of knowledge hierarchies to examine some of the implications of the Garicano (2000) model for the evolution of income inequality in the United States since the 1970's. Between 1974 and 1988, we saw an increase in both the 90/50 gap (the ratio of the log hourly wage for workers at the 90th percentile of the wage distribution to those at the median) and the 50/10 gap, two commonly used measures of wage inequality. But beginning in the late 1980's, the 50/10 gap began to decline, while the 90/50 gap continued to increase. This change in the late 1980's is difficult to reconcile with the two leading explanations for the increase in income inequality in the United States post-war period.

One source of wage inequality is the "superstar effect," which is present in Lucas (1978): higher-ability workers choose to become managers, and higher-ability managers manage larger firms and earn higher profits. Recall that in the Lucas (1978) model, the production technology is given by $y = \varphi n^{\theta}$, where $\theta < 1$ is a parameter that captures organizational diseconomies of scale. If we think of innovation in communication technology as reducing θ, the superstar effect becomes stronger: with a lower θ, better managers can manage larger firms, which means that for a given population of workers, a smaller share of them will become managers, and the returns to ability will increase. This reduction in θ will also have the effect of raising labor demand for a given wage level for production workers and will therefore lead to increased production-worker wages. Innovation in communication technology therefore leads to increased inequality at the top of the distribution, higher wages at the bottom, but no change in inequality at the bottom. This model can be augmented to allow

production workers of ability φ to supply, say, φ efficiency units of labor. If wages are the price for efficiency units of labor, there would be inequality at the bottom of the distribution, since different workers supply different numbers of efficiency units of labor. Innovation in communication technology would lead to an increase in the price of efficiency units of labor and therefore would lead to an increase in wage inequality at the bottom of the distribution, which is consistent with the trend from the mid-1970s to the late 1980s, but not with the trend following the late 1980s.

Explanations for the evolution of inequality based on this "superstar effect" (see, for example, Gabaix and Landier (2008)) at their core predict a fanning out of the wage distribution, but this is not what we have seen in recent years, where we have seen a hollowing out of the middle of the income distribution and a rise in the top and the bottom. Garicano and Rossi-Hansberg (2006) argue that taking organizational structure seriously leads to the phenomenon they refer to as the "shadow of the superstars," and advances in communication technology have made this phenomenon more important in recent years, providing an explanation for these recent trends.

Model Description Following a simplified version of Garicano and Rossi-Hansberg (2006), assume individuals are endowed with overlapping knowledge sets $[0, z]$. We can characterize this knowledge set entirely by its upper bound z, which we will take to be the individual's type. Assume the distribution of problems is uniform on $[0, 1]$, and let $\phi(z)$ denote the distribution of skill in the population. As in Lucas (1978), each worker chooses either to work in a firm or to work as an entrepreneur, and if she chooses to be an entrepreneur, she can either be an independent entrepreneur, or she can form a hierarchy and choose how many workers to employ. Assume that hierarchies have only two layers consisting of a single manager and a mass n of production workers.

As in Garicano (2000), a production worker receives one problem and is able to solve a fraction z of them. If he works in a hierarchy, and he is unable to solve a problem on

his own, he can refer the problem to his manager, who can assess $1/h$ referred problems. An **organization** then consists of a vector $g = (n, z_m, z_p)$, where n denotes the number of workers, z_m denotes the skill of its manager, and z_p denotes the skill level of its production workers. The manager of an organization will be referred $n(1 - z_p)$ problems and can solve them all as long as she has enough time to do so, or if $hn(1 - z_p) \leq 1$. A manager who hires more knowledgeable workers (i.e., hires workers with higher values of z_p) can hire more of them and still have enough time to assess the problems they refer, and a more-knowledgeable manager will be able to solve a larger fraction of these referred problems, so there is a natural complementarity between manager knowledge, worker knowledge, and the manager's span of control.

The price of hiring a worker of type z_p is $w_p(z_p)$. This production-worker wage function $w_p(z_p)$ will be endogenous to the equilibrium. Given wage function $w_p(z_p)$, a manager with skill z_m who hires n workers of skill z_p will produce output nz_m and incur a wage bill of $nw_p(z_p)$.

A **competitive equilibrium** is a **production-worker wage function** $w_p^*(z)$, which specifies the wage a firm must pay to hire a worker of knowledge z, a **labor-demand function** $n^*(z)$, which specifies the mass of production workers an entrepreneur of knowledge z hires, an **occupation-choice function** $d^*(z)$, which specifies for a worker of knowledge z, whether he becomes a production worker, an independent entrepreneur, or a manager, and an **assignment function** $m^*(z)$, which denotes the skill of the manager that a worker of knowledge z is matched to if he chooses to be a production worker.

The Program Given production-worker wage function $w_p(z_p)$, a manager with skill z_m will solve

$$w_m(z_m) \equiv \max_{z_p, n} (z_m - w_p(z_p))\, n = \max_{z_p} \frac{z_m - w_p(z_p)}{h(1 - z_p)},$$

where I substituted the manager's time constraint, holding with equality (which, under an equilibrium occupation-choice function, it will). A worker who chooses to be an independent

entrepreneur will produce one unit of problems and be able to solve a fraction z of them and will therefore receive a total payoff of $w_I(z) = z$. A worker with knowledge z therefore has to choose whether to be a production worker, an independent entrepreneur, or a manager, and therefore solves

$$w(z) = \max\{w_p(z), w_I(z), w_m(z)\}.$$

The objective will be to characterize the function $w(z)$ that arises in equilibrium and to describe how it changes in response to an increase in communication technology (a reduction in h).

To do so, first note that the production-worker wage slope at z_p has to satisfy the first-order condition for the firm that employs production workers with knowledge z_p:

$$\frac{-h(1-z_p)w_p'(z_p) + (z_m - w_p(z_p))h}{(h(1-z_p))^2} = 0$$

or $w_p'(z_p) = (z_m - w(z_p))/(1-z_p)$. For those workers who choose to become production workers, it must be the case that $w_p(z_p) > z_p$, and therefore

$$w_p'(z_p) = \frac{z_m - w(z_p)}{1-z_p} < \frac{z_m - z_p}{1-z_p} < 1.$$

For those workers who choose to become independent entrepreneurs, clearly, $w_I'(z) = 1$. Finally, by the envelope theorem, for workers who choose to become managers,

$$w_m'(z_m) = \frac{1}{h(1-z_p)} > 1.$$

We therefore have $w_m'(z) > w_I'(z) > w_p'(z)$, which implies that there will be two cutoffs, z^* and z^{**}, such that workers with $z \in [0, z^*]$ will choose to become production workers, workers with $z \in (z^*, z^{**}]$ will choose to become independent entrepreneurs, and workers with $z \in (z^{**}, 1]$ will choose to become managers. The marginal worker $z = z^*$ is indifferent between being a production worker and being an independent entrepreneur, so $w_p(z^*) = z^*$,

and the marginal worker $z = z^{**}$ is indifferent between being an independent entrepreneur and being a manager, so $w_m(z^{**}) = z^{**}$, and clearly $w_m(1) = 1/h$. These conditions pin down w^* and z^*.

The equilibrium assignment function $m^*(z)$ is pinned down by the market-clearing condition. First, there will be positive sorting, so that $m^*(z)$ is increasing in z. Next, note that the labor market must clear for production workers of knowledge z for all $z \leq z^*$. Labor supply for workers of knowledge z is $\phi(z)$, and labor demand for workers of knowledge z is $n(m^*(z))\phi(m^*(z))$. The labor-market clearing for production workers with knowledge $z \leq z_p$ can therefore be written as

$$\int_0^{z_p} \phi(z)\, dz = \int_{z^{**}}^{m^*(z_p)} n(m^*(z_p))\, \phi(m^*(z_p))\, dz,$$

which we can differentiate with respect to z_p to get

$$m^{*\prime}(z_p) = \frac{1}{n(m^*(z_p))} \frac{\phi(z_p)}{\phi(m^*(z_p))} = h(1 - z_p) \frac{\phi(z_p)}{\phi(m^*(z_p))}.$$

If ϕ is the uniform distribution, then this condition is simply $m^{*\prime}(z_p) = h(1 - z_p)$ for all $z_p \in [0, z^*]$. This condition, along with the facts that $m^*(0) = z^{**}$ and $m^*(z^*) = 1$, pin down m^* and z^{**}.

As in Lucas (1978), this model features the superstar effect: wages for managers satisfy $w_m'(z_m) = 1/(h(1 - z_p))$ so that higher-ability managers receive higher wages, and it is convex in z_m, since $m^*(z)$ is increasing. Moreover, the slope and convexity of this wage function is higher when h is lower, so innovations in communication technology can lead to increased inequality at the top of the distribution.

In addition to this superstar effect, this model features what Garicano and Rossi-Hansberg (2006) refer to as the "shadow of the superstars" effect: improvements in communication technology improve managers' ability to leverage their knowledge, so more-knowledgeable managers will manage larger teams. The threshold for being a manager therefore in-

creases, which reduces the earnings for those that would have been managers for higher values of h. Moreover, among production workers, the slope of the matching function $m^{*\prime}(z_p) = h(1 - z_p)$ declines when there is better communication technology (since production workers will now be employed by a smaller mass of managers), which implies a reduction in the convexity of the production-worker wage function, which satisfies $w_p'(z_p) = (m(z_p) - w(z_p)) / (1 - z_p)$. Improvements in communication technology therefore raise wages at the top, reduce wages in the middle, and increase wages at the bottom.

Garicano and Rossi-Hansberg (2006) is motivated by compelling and, at a high level, puzzling facts about the evolution of income inequality in the United States. And it provides an organizationally based explanation that does a better job of matching the facts than other leading explanations. For example, one of the leading explanations for increased inequality in recent years is the skill-biased technological change hypothesis in which the decline in the cost of capital equipment has decreased the price of routine tasks that are highly substitutable with this form of capital. In contrast, analytic and manual tasks are less substitutable, so the improvement in technology leads to higher demand and higher employment of workers performing those tasks. While this explanation can account for the observed trends, it is less obvious why this would lead to a decline in the middle class rather than a decline in the lower end of the income distribution. Explanations based on superstar effects at their core, predict a fanning out of the income distribution, which is not consistent with recent trends.

Monitoring Hierarchies

Any theory of optimal firm size has to provide an answer to the *replication question*: "Why can we not simply double all the firm's inputs and double the resulting output?" Fundamentally, the answer has to be that there is some sort of fixed factor of production *at the firm level* that is not replicable, and while Penrose (1959) argues that this factor must be related in some way to what managers do, it is not obvious exactly what it is about what

they do that makes it fixed in nature (i.e., if one manager's ability to coordinate activities is fixed, why not hire a second manager?), "Whether managerial diseconomies will cause long-run increasing costs [requires that] management... be treated as a 'fixed factor' and the nature of the 'fixity' must be identified with respect to the nature of the managerial task of 'coordination.' This identification has never been satisfactorily accomplished." (p. 12)

Vertical Control Loss

Williamson's (1967) answer is that even if one were to double the number of managers in a firm in order to get double their coordination efforts, someone would have to coordinate *their* activities as well, and therefore coordination activities at the highest level necessarily must be embodied within a single individual. He describes a theory in which a firm consists of a layer of workers and a hierarchy of monitors. The top-level manager supervises a layer of subordinates who each supervise a layer of subordinates, and so on until we get to the bottom layer of the firm, which consists of production workers. Production workers produce one unit of output each, but some of this output gets lost for each layer in the organization, a reduced-form way to capture communication losses and agency costs, and to capture Williamson's idea that "The larger and more authoritarian the organization, the better the chance that its top decision-makers will be operating in purely imaginary worlds." (p. 123)

Model Description In the simplest version of Williamson's model (due to Mookherjee (2013)), each manager has an exogenously specified span of control s, so if there are $N+1$ layers $i \in \{0, 1, \ldots, N\}$, there are s^i employees in layer i and therefore s^N production workers who each produce α^N units of output, where $\alpha < 1$ represents the fraction of output that gets lost for each layer of the organization. This parameter, referred to as the **vertical control loss**, is a reduced-form way to capture communication losses and agency costs.

Wages for production workers are w, and wages for employees in layer i are $\beta^{N-i}w$, where $\beta > 1$ represents the additional wages that have to be paid for employees higher in

the organization. Assume $\alpha s > 1$ and $s > \beta$. An **organization** is fully characterized by a number of layers N.

The Program The firm chooses the number of layers of subordinates, N, to solve

$$\max_N \pi(N, \alpha)$$

where

$$\pi(N, \alpha) = (\alpha s)^N - w \sum_{i=0}^{N} \beta^{N-i} s^i = (\alpha s)^N - w \frac{s^{N+1} - \beta^{N+1}}{s - \beta}.$$

The objective function satisfies increasing differences in (N, α), so N^* is increasing in α (i.e., the optimal number of layers is higher if less information is lost between successive layers or if there are lower agency costs). Under some parameter restrictions, there is an interior solution as long as $\alpha < 1$ ($N^* \to \infty$ as $\alpha \to 1$), so this model pins down the optimal number of layers and hence the optimal number of workers in the firm.

The paper provides an answer to the question of why there are organizational diminishing returns to scale: activities within the firm must be coordinated, and the highest-level coordination must occur within the single individual who occupies the top position. The model's main results, however, require some pretty stringent parameter restrictions, since the firm's revenues, $(\alpha s)^N$ are convex in N (since $\alpha s > 1$). For there to be an interior solution, it has to be the case that the firm's costs, $w\left(s^{N+1} - \beta^{N+1}\right)/(s - \beta)$ are, in some sense, even more convex in N. Moreover, while the paper does answer the initial question, it is silent both on why there is vertical control loss and what determines it, as well as on why wages progress in a proportional way through the hierarchy.

Layers of Supervisors

Calvo and Wellisz (1978) put more structure on the Williamson (1967) model by explicitly introducing productive effort by production workers, monitoring effort by supervisors, and

optimal wage choices by the firm.

Model Description Suppose there are $N+1$ layers in the firm, there are M_i workers in layer $i \in \{0, 1, \ldots, N\}$, and layer $i = N$ represents production workers. Workers in layer i are paid a base wage of w_i, which may be docked for non-performance, as we will see below.

Suppose each worker in the firm has a utility function $w - c(e)$, where w is the worker's wage, $c(e)$ is the cost of effort, and $e \in [0, 1]$ represents the fraction of the week worked. Each unit of time spent working by a production worker generates one unit of output for the firm, so if all production workers exert effort e_N, the firm's total output will be $M_N e_N$. Supervisors exert effort to uniformly monitor their direct subordinates. If M_i supervisors at level i work e_i units of time each, they supervise each $i+1$-level employee an amount equal to $e_i M_i / M_{i+1}$. With probability $p_{i+1} = h(e_i M_i / M_{i+1})$, the firm observes e_{i+1} and reduces worker $i+1$'s wage from w_{i+1} to $e_{i+1} w_{i+1}$. With probability $1 - p_{i+1}$, the worker receives wage w_{i+1}.

If a worker exerts effort e and is monitored with probability p, his utility will be $pwe + (1-p)w - c(e)$, and he will choose $c'(e^*(p, w)) = pw$. If his outside option yields utility $\bar{u}$, he will accept employment at the firm if $pwe^* + (1-p)w - c(e^*) \geq \bar{u}$. An **organization** is a vector $g = (M, e, w, p)$, consisting of a number of workers at each level below the top, $M = (M_1, \ldots, M_N)$, an effort vector $e = (e_1, \ldots, e_N)$, a wage vector $w = (w_1, \ldots, w_N)$, and a monitoring probability vector $p = (p_1, \ldots, p_N)$, where $p_{i+1} = h(e_i M_i / M_{i+1})$.

The Program Suppose the firm has $N = 1$. The firm's problem is therefore to choose the number of production workers M_1 and wage w_1 to maximize

$$\pi_1^* = \max_{M_1, p, w, e} M_1 e - [pwe + (1-p)w] M_1,$$

subject to monitoring feasibility $p = h(1/M_1)$, incentive-compatibility $c'(e) = pw$, and individual rationality: $pwe + (1-p)w - c(e) \geq \bar{u}$. Since $p = h(1/M_1)$ is decreasing in M_1,

increasing M_1 reduces effort per worker fixing w. A larger firm must therefore either reduce worker effort or pay each worker more, which implies that there are decreasing returns to scale, holding the organizational structure fixed.

Another way the firm can expand is by increasing the number of layers. If the firm has $N+1$ layers, it chooses $g = (M, e, w, p)$ to solve

$$\pi_N^* = \max_{M,p,w,e} M_N e_N - \sum_{i=1}^{N} \left[p_i w_i e_i + (1 - p_i) w_i \right] M_i$$

subject to monitoring feasibility

$$p_i = h\left(e_{i-1} M_{i-1} / M_i \right),$$

incentive-compatibility $c'(e_i) = p_i w_i$, and individual rationality $p_i w_i e_i + (1 - p_i) w_i - c(e_i) \geq \bar{u}$. Adding an additional layer allows the firm to employ more production workers while maintaining a given effort level for those production workers, since it can increase $p_N = h\left(e_{N-1} M_{N-1} / M_N \right)$ by increasing the number of direct supervisors for production workers or getting them to work harder. Doing so is costly, though, since the firm has to pay and motivate the additional layer of supervisors.

The paper's main result is that as long as it is profitable to produce with a single layer rather for the entrepreneur to do all the production herself, $\pi_1^* > 1$, then firm size is unbounded, that is, $\lim_{N \to \infty} \pi_N^* = \infty$. The paper outlines a replication argument by showing that there is a way of expanding the number of production workers indefinitely by increasing the number of layers in the firm. It does so by finding an $N = 1$ arrangement that leads to positive profits for each production worker and uses it to construct an $N > 1$ level hierarchy in which wages, effort, and the span of control are constant across levels. If it was profitable to add the second layer, it will be profitable to add the third layer, and so on, so profits are unbounded along this path.

The negative conclusion of this paper is that Williamson's proposed fixed input, once mi-

crofounded, does not lead to bounded firm size. Qian (1994) shows that Calvo and Wellisz's result is also fragile along a couple dimensions, putting Williamson's explanation on foundations that are at best highly contingent on the underlying environment.