MECS 570-1: Organizational Economics I

Michael Powell

Organizational Structure I: Span of Control, Knowledge Hierarchies

Organizational Structure

Throughout the class, we have taken as given that an organization consists of multiple individuals, but we have said little about why they interact together or why a single individual cannot do everything. For example, in our discussion of incentive conflicts, we took as given that there was an agent who was able to exert effort in production and a principal who was unable to do so. In our discussion of firm boundaries, we took as given that one party was able to make a type of specific investment that the other party was unable to do so. Garicano and Van Zandt (2013) argue that, "If the mythical unbounded rational, all-capable owner-manager-entrepreneur-manager existed, there would be no organizational processes to talk about, no need to delegate the coordination in organizations to several or many agents, and therefore no organization structure other than a center and everyone else." In other words, fundamentally, bounded rationality (of some sort) is why there are returns to specialization, it is why these returns to specialization are not unlimited, and it is why there is scope for organizations to serve, as Arrow (1974) points out, "as a means of achieving the benefits of collective action in situations in which the price system fails." (p. 33).

In this note, we will explore a couple classes of models that take aspects of bounded rationality as a key limitation that organizations have to contend with and which organizations are specifically designed to address. Before we do so, I will first describe a simplified version of the Lucas (1978) span-of-control model, which in some sense takes a form of bounded rationality as given and thinks about its aggregate implications for the economy. This model is in some ways a building block for the models that will follow.

Span-of-Control Model

If we are to understand why different firms of different size and productivity coexist in equilibrium, we need a model in which firm sizes and firms' production decisions are determined equilibrium. I will begin by describing a simplified version of the canonical Lucas (1978) "span of control" model in which people in the economy choose whether to be workers or to become entrepreneurs who employ workers. Some people are better at managing others, and these are the people who will, in equilibrium, opt to become entrepreneurs. Better managers optimally oversee more workers (the "span of control" effect), but there are diminishing marginal returns, so that it is not an equilibrium for there to only be a single firm. The underlying organizational source of diminishing marginal returns to management is unmodeled, and Lucas describes the model as providing not "serious organization theory, but perhaps some insights into why organization theory matters economically."

Description There is a unit mass of agents in the economy. Agents differ in their ability to oversee the work of others. Denote this ability by φ, and let $\Gamma(\varphi)$ denote its distribution function, which we will take to be continuous. Each agent chooses whether to be a worker and receive wage w or to become an entrepreneur and receive the profit associated with the enterprise. Output generated by an entrepreneur depends on her managerial ability (φ) and on the number of workers (n) she employs:

$$y = \varphi n^{\theta},$$

where $\theta < 1$ is a parameter that captures in a reduced-form way the organizational diseconomies of scale.

A **competitive equilibrium** is a wage w^*, a labor-demand function $n^*(\varphi)$, and an occupational choice function $d^*(\varphi) \in \{0, 1\}$ specifying which subset of agents become entrepreneurs, how many workers each firm hires, and a wage at which labor demand equals

labor supply.

Suppose an agent of ability φ chooses to become an entrepreneur ($d = 1$). Her labor demand solves

$$\max_n \varphi n^\theta - wn$$

or $n^*(\varphi, w) = (\varphi\theta/w)^{1/(1-\theta)}$, and her associated profits are

$$\pi(\varphi) = \varphi n^*(\varphi, w)^\theta - wn^*(\varphi, w) = (1-\theta)\varphi^{\frac{1}{1-\theta}}\left(\frac{\theta}{w}\right)^{\frac{\theta}{1-\theta}}.$$

An agent will therefore choose to become an entrepreneur if $\pi(\varphi) \geq w$. Since $\pi(\varphi)$ is increasing in φ, there will be some cutoff $\varphi^*(w)$ such that all agents with ability $\varphi \geq \varphi^*(w)$ will become entrepreneurs and all those with ability $\varphi < \varphi^*(w)$ will choose to be workers. Equilibrium wages, w^*, therefore solve

$$\int_{\varphi^*(w^*)}^{\infty} n^*(\varphi, w^*)\, d\Gamma(\varphi) = \Gamma(\varphi^*(w^*)),$$

where the expression on the left-hand side is aggregate labor demand at wages w^*, and the right-hand side is labor supply—the mass of agents who choose to be workers at wage w^*.

This model makes predictions about who will become an entrepreneur, and it has predictions about the distribution of wages and earning as well as firm size. The diminishing returns to the "span of control" effect are a key variable determining the model's predictions, and the model says little about what governs them. Moreover, the model has some predictions about the evolution of wage inequality that are counter to what we have observed over the post-war period in the United States. These two issues are addressed in the models that we will be examining next.

Knowledge Hierarchies

Garicano (2000) points out that knowledge is an important input into the production process that many of our existing models ignore. Knowledge, the ability to solve problems that naturally arise in the production process, is embodied in individuals who have limited time to work, and this is fundamentally why there are returns from specialization and why organization is important. As Demsetz (1988) points out, "those who are to produce on the basis of this knowledge, but not be possessed of it themselves, must have their activities directed by those who possess (more of) the knowledge. Direction substitutes for education (that is, for the transfer of knowledge)." Education is costly, so an organization will want to specialize the possession of non-routine knowledge in a few and allow production workers to ask for direction when they need help on such non-routine problems. The organizational-design question under the Garicano (2000) view is then: how should a firm organize the acquisition, use, and communication of knowledge in order to economize on the scarce time of its workers and leverage scarce knowledge?

Description A unit mass of workers interact together to produce output, and problems $z \in [0, \infty)$ arise during the production process. Workers are segmented into L distinct classes, with a fraction $\beta_i \geq 0$ in class $i \in \{1, \ldots, L\}$ so that $\sum_i \beta_i = 1$. Workers in class i spend a unit of time producing (t_i^p) or helping others $\left(t_i^h\right)$ so that $t_i^p + t_i^h \leq 1$ and $t_i^p, t_i^h \geq 0$, and they possess knowledge set $A_i \subset [0, \infty)$, which costs the organization $c\mu(A_i)$, where $\mu(A_i)$ is the Lebesgue measure of the set A_i. The knowledge sets of workers in different classes can in principle overlap.

Each unit of time a worker spends producing generates a problem z, which is drawn according to distribution function $F(z)$ with density $f(z)$. Without loss of generality, we can order the problems so that $f'(z) < 0$. If a worker of class i encounters a problem $z \in A_i$, she solves the problem and produces one unit of output. If $z \notin A_i$, she can refer the problem to someone else in the organization. If she does not know the solution to the problem, she

does not know who else in the organization might know how to solve the problem. That is, workers "don't know what they don't know."

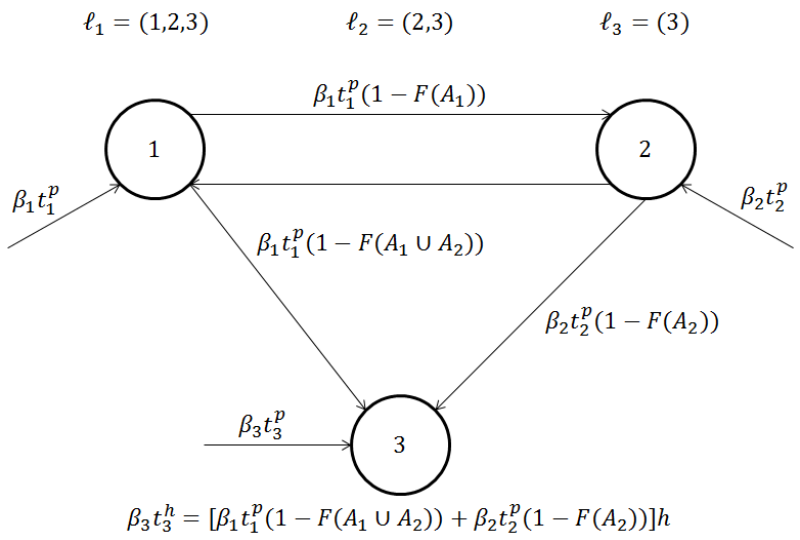
Each unit of time a worker spends helping allows her to process $1/h$ referred problems, where $h < 1$ represents the "communication costs" the helper incurs learning about and assessing the problem. She incurs these costs even when she does not know the answer.

A **referral pattern** ℓ is, for each class i, an ordered set of classes ℓ_i that i can refer to, where $\ell_i(1) = i$ for all i (so that each class can solve any problem it originates), and $\ell_i(n)$ is the nth class that i can refer to. In other words, a worker first tries to solve any problem herself, then she refers it to $\ell_i(2)$, $\ell_i(3)$, and so on. We will say that $j \prec_k i$ if class j precedes group i in the referral pattern for group k. An **organization** is a vector $g = (L, \beta, A, t, \ell)$, which specifies a **number of classes** L, a **class assignment** $\beta = (\beta_1, \ldots, \beta_L)$ with $\beta_i \geq 0$, $\sum_{i=1}^{L} \beta_i = 1$, a **knowledge assignment** $A = (A_1, \ldots, A_L)$, a **time allocation** $t = (t_1, \ldots, t_L)$, where $t_i = \left(t_i^h, t_i^p\right)$ with $t_i^h, t_i^p \geq 0$, $t_i^h + t_i^p \leq 1$, and a referral pattern ℓ.

The problem the organizational designer faces is to choose an organization to maximize the firm's net output. To determine the firm's net output, first note if class i spends enough time helping the classes that refer to it, we will have

$$\underbrace{\beta_i t_i^h}_{\text{time helping}} = \underbrace{\sum_{k:i\in\ell_k} \beta_k t_k^p \left[1 - F\left(\bigcup_{j\prec_k i} A_j\right)\right]}_{\text{problems referred to } i} \underbrace{h}_{\text{time per referral}}.$$

The following figure depicts a sample organization.



In this organization, there are three classes that each spend some time in production, so problems flow into each class. Class 1 is able to refer problems to class 2 and then to class 3, class 2 is able to refer problems to class 3, and class 3 is unable to refer any problems. Class 2 must spend $\beta_2 t_2^h = \beta_1 t_1^p \left(1 - F\left(A_1\right)\right) h$ to process all the problems that class 1 refers to it, and class 3 must spend

$$\beta_3 t_3^h = \left[\beta_1 t_1^p \left(1 - F\left(A_1 \cup A_2\right)\right) + \beta_2 t_2^p \left(1 - F\left(A_2\right)\right)\right] h$$

to process all the problems that classes 1 and 2 refer to it.

The **net output of class i** is equal to the mass of problems it originates times the probability that someone in its referral pattern knows the solution to those problems minus the costs of class i's knowledge. Total net output is the sum of the net output of all L classes:

$$y = \sum_{i=1}^{L} \left(\beta_i t_i^p F\left(\bigcup_{k \in \ell_i} A_k \right) - c\beta_i \mu\left(A_i\right) \right).$$

The **Coasian program** is therefore

$$\max_{L,\beta,A,t,\ell} y \text{ subject to } t_i^h + t_i^p \leq 1, \sum_{i=1}^{L} \beta_i = 1.$$

This problem is not a well-behaved convex programming problem, but several variational arguments can be used to pin down the properties of its solution. First, for any knowledge assignment A, it turns out that $t_i^p = 1$ for some class i, which we will without loss of generality set to $i = 1$, and $t_i^h = 1$ for all others. In other words, in any optimal organization, each worker uses all of her time, one class of workers specializes entirely in production, and the remaining workers specialize in solving problems the production workers refer to them. Any optimal organization has this feature, because if one class produces a higher net output than another, we can always move some of the workers from the less-productive class to the more-productive class and adjust helping times so as to maintain the high-productivity class's net output. Doing so will reduce the amount of time other classes spend producing, and their net output will fall. This perturbation is always feasible as long as multiple groups are production.

The second property of the solution is that the measure of any overlap between two knowledge sets is zero: $\mu(A_i \cap A_j) = 0$. The reason for this is that the knowledge of "problem solvers" never gets used if it is known by the producer class, and it never gets used if it is known by an earlier class in the producer class's list. So whenever $\mu(A_i \cap A_j) > 0$ for some $i \prec_1 j$, we can let $\tilde{A}_j = A_j \backslash A_i$. Under this perturbation, the same problems are solved by the organization, but at a lower cost, since the costs of the higher class's knowledge set is lower.

Next, any optimal organization will feature $A_1 = [0, z_1]$, $\ell_1 = (1, \ldots, L)$, and $A_i = [z_{i-1}, z_i]$ with $z_i > z_{i-1}$. Production workers will learn to solve the most common ("routine") problems, and problem solvers learn the exceptions. Moreover, the workers in the higher classes learn to solve more unusual problems. To see why this is true, suppose class i knows

$[z, z+\varepsilon]$ and $j \prec_1 i$ knows $[z', z'+\varepsilon]$, where $f(z) > f(z')$. Then we can swap these two intervals for a small mass of each class of workers. Doing so will keep the learning costs the same. Production will be the same, since the total amount of knowledge is unchanged. But the time spent communicating problems goes down, since those earlier in the referral pattern are now less likely to confront a problem they do not know. As a result, some of the time freed up from the higher class can then be reallocated to the producer class, increasing overall output. A similar argument guarantees that there will be no gaps in knowledge between the classes (i.e., A_i and A_{i+1} overlap at exactly one point). Garicano describes this property as "management by exception" and highlights that it allows specialization in knowledge to be attained while minimizing communication costs.

Finally, any optimal organization necessarily has a pyramidal structure: if $L \geq 2$, then $\beta_1 > \beta_2 > \cdots > \beta_L$. The reason for this is that the total time spent helping by class i is $\beta_i = [1 - F(z_{i-1})] h\beta_1$, and the total time spent helping by class $i+1$ is $\beta_{i+1} = [1 - F(z_i)] h\beta_1$. Since $z_i > z_{i-1}$, $\beta_{i+1} < \beta_i$.

The Coasian program therefore becomes

$$\max_{L, z_1, \ldots, z_L, \beta} F(z_L) \beta_1 - \sum_{i=1}^{L} c\beta_i (z_i - z_{i-1}),$$

where $z_0 = 0$, subject to

$$\beta_i = [1 - F(z_{i-1})] h\beta_1 \text{ for } i > 1.$$

This problem can be solved for specific distributional assumptions. For example, it can be solved explicitly if the distribution of problems is exponential, so that $f(z) = e^{-\lambda z}$. In the solution to this problem, production workers know how to solve problems in an interval of length Z_w^*, and all problem solvers know how to solve problems in an interval of length Z_s^*. If we define the **span of control** at layer i as $s_i = \beta_i / \beta_{i+1}$, we can say something about the comparative statics of the optimal organization.

First, if communication costs fall (i.e., if h goes down) because of improvements in com-

munication technology, then Z_s^* increases, Z_w^* falls, and the span of control increases at each layer. That is, improvements in communication technology lead to flatter organizations with less-knowledgeable production workers. If communication becomes cheaper, relying on problem solvers is "cheaper," so it is optimal for each production worker to acquire less knowledge, and each problem solver can communicate solutions to a larger team, so the span of control of problem solvers increases.

If the cost of acquiring knowledge falls (i.e., c decreases) because of improvements in information technology, then Z_s^*, Z_w^*, and s_i all increase. Improvements in information technology therefore also lead to flatter organizations with more knowledgeable helpers, but they also lead to an increase in the knowledge possessed by production workers.