Broad Terms and Organizational Codes*

Joel Sobel†

March 18, 2016

Abstract

This paper studies organizations in which participants have common preferences but communication is costly. In this model, use words to describe similar information, but need not use precise words for frequent events and vague words for unusual ones. The model identifies a source of communication failure across units. It provides an argument for giving the best-informed agents decision-making authority.

Journal of Economic Literature Classification Numbers: D23, D83, L23; Keywords: communication, organizational behavior, authority

*I worked on the project while visiting the Department of Economics at Erasmus University, CREED at the University of Amsterdam, and Paris School of Economics. I thank these institutions and NSF for their support. I am grateful to Andrea Galeotti, Christian List, and Gilles Saint-Paul for providing references. I thank Andreas Blume, Wouter Dessein, Navin Kartik, Niko Matouschek, Andrea Prat, referees, and seminar participants for comments.

†University of California, San Diego

1 Introduction

Effective communication requires conversational partners to coordinate on a language. The nature of the language determines what information individuals share. This paper studies the structure of optimal languages in a setting where communication is costly and investigates the interplay between structure of organization and language. Inspired by Arrow [2], Crémer, Garicano, and Prat [7] (CGP) present a team-theoretic model of communication in organizations that delivers several sharp results. This paper reexamines the core questions raised by CGP in a slightly different model.

In the basic model of CGP, there are a finite set of states of the world. One agent (salesman) must describe the state of the world to another agent (engineer). The salesman communicates to the engineer. CGP model the complexity of communication by restricting what the salesman can say to a finite set of words. If the number of words is less than the number of states, then, inevitably, the salesman must sometimes use the same word to describe different states. When this happens, the engineer must pay a cost to learn the true state. CGP assume that this cost is an increasing function of the number of states represented by a word. In this setting, a language is a partition of the states. Associated with each language is an expected cost. The common objective of the salesman and engineer is to minimize expected costs. CGP describe interesting properties of optimal languages and the implications of the costly communication on organizational structure. One feature of natural language is the **interval property**: broad terms lump together similar items. If one has a limited vocabulary to describe a scalar state variable, it is common for the language to partition states into intervals. Course grades lump together students who receive similar exam scores. Cities may be classified according to their size (small, medium, large). The climate might be hot or cold. Color terms typically group together colors with similar characteristics. One cannot investigate whether optimal languages pool similar states in the basic model of CGP because the paper imposes no structure on states. CGP generalize their model to allow the state space to be a real interval. This provides a natural notion of similarity. For this model, CGP demonstrate that if words are constrained to be intervals, then the main qualitative properties of their general analysis continue to hold.[1]

[1] Gärdenfors [10] presents a model of conceptual spaces that describes how to form categories. The book contains examples of different objects that are lumped together by

What CGP's analysis fails to do is provide an explanation for why optimal languages should pool similar items. Investigating this issue is the starting point for my analysis. In order to investigate the question, I derive the objective function from more basic assumptions. Specifically, I assume that after the engineer hears the report of the salesman, she makes a decision and then incurs a loss that depends on the decision and the state of nature. This approach leads to an alternative form of loss function: Losses depend on the set of states lumped together not just the cardinality of the set. In this setting, there are simple conditions under which optimal languages are interval partitions. Hence, in contrast to the model of CGP, the interval property of optimal languages is a conclusion.

CGP show that optimal languages in their model satisfy interesting properties. First, the most likely states are pooled in more precise words. Second, precise words are the most common. The first property is only compatible with the property that words describe similar states if similar states arise with similar probabilities. Hence CGP's property does not hold in my model without additional assumptions. The second result generally does not hold in my model. CGP also show that in their model a problem becomes easier (in the sense that the cost of the optimal language decreases) if probability shifts from unlikely to likely states. In my model, lower costs are associated with more concentrated distributions. In Section 4 I describe these results in detail.

CGP extend their model to discuss efficient organization of firms when information is diffuse and costly to transmit. They demonstrate how decreasing costs of communication create incentives to integrate separate units and maintain fewer layers in organizational hierarchies. Section 5 shows that these insights persist under the model of this paper. Explicitly modeling decision making makes it possible to connect the observations of CGP to findings in the delegation literature. I do this in Section 6. In Section 6 I relate the multi-agent version of my model to the frameworks introduced by Alonso, Dessein, and Matouschek [1] and Rantakari [13] to study optimal assignment of decision authority in organizations. Organizational design is important in their models because agents have different preferences. My

similar properties and describes when representations involve lumping together similar objects. Motivated by models of rational inattention, Saint-Paul [14] provides a model of a coarse language. His paper derives a general version of the interval property similar to Proposition 3. I derived my result independently of Saint-Paul, but his result takes precedence.

analysis revisits the problem under the assumption that it is complexity of communication that determines optimal design of organizations.

Blume and Board [5], Blume and Board [4], and Blume **??** study communication games in which full communication may be impossible even though agents have identical preferences. In these papers, what stands prevents full communication is failure to have common knowledge about the language competence of players. For example, in this environment, the Receiver may not know whether the Sender sends a message because it accurately describes the state or because the Sender does not know the message that corresponds to the true state. These papers identify a fundamental reason for communication failure. The perspective of my paper is different because one of the purposes of training in firms is to educate employees on the organizational code. One cannot guarantee that training will make the language common knowledge, but it suggests that the forces that may lead to communication failure in the models of Blume and Board are less central in organizations.

2 The Model

The state of the world is an element $\theta \in [0,1]$. $F(\cdot)$ is the cumulative distribution function of θ. $F(\cdot)$ is strictly increasing on $(0,1)$ and has positive density $f(\cdot)$. One agent, the salesman, observes θ. The salesman selects a word from a finite set, $W_1, \dots, W_K$. The other agent, the engineer, observes the word selected by the salesman, W_k, but not the state of the world. The engineer must take an action $a \in \mathbb{R}$. There is a loss function $L : \mathbb{R} \times [0,1] \to \mathbb{R}$; $L(a,\theta)$ is the loss associated with taking action a in state θ. Assume $L(\cdot)$ is twice continuously differentiable and strictly convex in a. These assumptions are primarily for convenience. They permit one to use calculus to describe optima and guarantee that there is a unique optimal response to any word.

The joint objective of the salesman and engineer is to minimize the expected losses. They do so by coordinating on an optimal language.

An optimal language $\{W_1^*, \dots, W_K^*\}$ must solve:

$$\min \sum_{i=1}^{K} \int_{W_i} L(a_i^*, \theta)\, dF(\theta) \tag{1}$$

where

$$a_i^* \ solves \ \min_{a \in A} \int_{W_i} L(a,\theta)\, dF(\theta). \tag{2}$$

Further, it must be the case that

$$W_i \subset \{\theta : L(a_i,\theta) = \min_j L(a_j,\theta)\}. \tag{3}$$

That is, informed agents will always use a word that induces a minimum-cost action. If (3) failed, then losses would go down if one merely rearranged the states assigned to each word without changing the responses to each word.[2]

A potential generalization would be to permit the team to choose probabilistic languages. That is, the language could be thought of as a function $\sigma : \{1,\dots,K\} \times [0,1] \to [0,1]$ such that $\sum_{k=1}^{K} \sigma(k,\theta) = 1$ for all θ. The interpretation is that $\sigma(k,\theta)$ is the probability of using the k^{th} message (word) given state θ. In the general formulation, σ gives rise to posterior beliefs $\mu_1,\dots,\mu_K$ where

$$\mu_i(\theta \mid k) = \frac{\sigma(k,\theta) f(\theta)}{\int \sigma(k,\theta') dF(\theta')}. \tag{4}$$

In the formulation, I assume $\sigma(\cdot)$ is either zero or one so that words partition the state space. Restricting to partitions is without loss of generality in the sense that the value of the problem in (1) is equal to the value if probabilistic languages were permitted. I prove this claim in the appendix.

CGP introduce a related model. In their setting, they assume that there exists a function $D : [0,1] \to \mathbb{R}$. $D(l)$ that is the cost of decoding a word that has (Lebesgue) measure l.[3] A CGP optimal language $\{W_1^*,\dots,W_K^*\}$ must solve:

$$\min \sum_{i=1}^{K} p(W_i) D(l_i) \tag{5}$$

[2]There are several papers in the engineering literature that study this problem. Kieffer [12], for example, presents conditions describes how to characterize solutions to (1) and gives sufficient conditions for the solution to be unique. Related problems arise in cheap-talk models of communication, where differences in preferences between the salesman (Sender) and engineer (Receiver) lead to constraints on the number of words used in equilibrium. Chen and Gordon [6] and Szalay [15] study models of this kind.

[3]CGP concentrate on a model in which the set of states is finite. In this case, $D(\cdot)$ is a function of the number of states pooled together in a single word. They do consider a variation of the model in which the state space is the unit interval. When they analyze this variation, they assume $D(\cdot)$ is linear.

where $p(W_i) = \int_{W_i} dF(\theta)$ is the probability of word W_i and l_i is the Lebesgue measure of W_i. Since $\int_{W_i} D(l_i)\, dF(\theta)$ is equal to the probability of W_i, the objective is to minimize the expected costs of decoding.

CGP and I study the same basic problem: The optimal communication structure in a setting where information is distributed, participants have common preferences, and perfect communication is not feasible. Both approaches model complexity by imposing an exogenous limit on the number of different things that the informed player can say (or the uninformed player can understand). Both approaches assume that the organization can coordinate on an optimal language given the complexity constraint. The approaches differ because they assume different objectives. The costs in CGP are a function of the "size" of the words. The costs in my model are a measure of how difficult it is to make a decision with limited information. The two approaches are alternative reduced forms of a model in which the objective is to minimize expected costs and costs are an arbitrary function of the set of states in a word. CGP explain that their formulation applies to a setting in which the engineer must learn the state of nature and the cost of identifying the state given a vague statement by the salesman is proportional to the number of sets consistent with the statement. My model applies to a setting in which the engineer must make a decision and the quality of the decision depends on the quality of the information provided by the salesman.[4]

I conclude this section with a technical note. There is a general model that captures both CGP's formulation and mine. In this model the optimal language $\{W_1^*, \ldots, W_K^*\}$ must solve:

$$\min \sum_{i=1}^{K} p(W_i) M(W_i) \tag{6}$$

where $M(W)$ the expected cost given word W. Expression (6) describes the problem of minimizing a general expected cost. In CGP's formulation,

[4]CGP describe their model as if there are two ways to get the information: either the salesman supplies it or the engineer figures it out. The salesman is unable to supply full information due to constraints on the available language. The engineer has a cost that depends on the number of states. I make identical assumptions about what the salesman knows and how much he can say, but assume that the engineer must make a decision based on information available to her. This assumption is compatible with the interpretation that the engineer must acquire information to make a decision, but typically does not give rise to a cost function that depends only on the number of states in each word.

the cost given W is a function of the length of W: $M(W) = D(l(W))$; in my formulation, the cost given W is the expected loss given W: $M(W) = \min_a \int_W L(a,\theta)\, dF(\theta)/p(W)$. It is straightforward to see – and subsequent results will demonstrate – that my formulation is not a special case of CGP's formulation. It is also the case that CGP's formulation is not a special case of mine. The next proposition formalizes this insight. For any loss function L, let G be a probability distribution over states (to be though of as a posterior distribution given a word) and define $M^*(G) = \min \int L(a,\theta)\, dG(\theta)$ to be the expected cost.

Proposition 1 *Let $L : \mathbb{R} \times [0,1] \to \mathbb{R}$ be a continuous function such that $M^*(G)$ exists for all nonatomic distributions on $[0,1]$. If $M^*(G)$ depends only on the support of G, then for all measurable G' and G'', $M^*(G') = M^*(G'')$.*

The Appendix contains proofs for this and subsequent propositions.

The proposition states that if the expected cost derived in a loss-minimizing formulation depends only on the support of the word, then the expected loss function is constant. CGP's loss function depends only on the length of the support of W. Hence the proposition implies that the only such loss function must be constant. So no non-trivial CGP loss function can be described using my model.

3 The Interval Property

In my model, language establishes a relationship between words and actions. A partition of the state space into words, $W_1, \ldots, W_K$, induces actions $\{a_1, \ldots, a_K\}$, where a_i is the response to word W_i. I seek conditions under which the optimal language groups similar states together. The next definition formalizes this idea.

Definition 1 *The language $\{W_1, \ldots, W_K\}$ is an* **interval partition** *of $[0,1]$ if there exists $\theta_0 \leq \theta_{i-1} \leq \theta_i \leq \ldots \leq \theta_K = 1$ such that $W_i = [\theta_{i-1}, \theta_i)$ of $0 < i < K$ and $W_K = [\theta_{K-1}, \theta_K]$.*[5]

It is straightforward to show under maintained assumptions, all words in optimal languages will be non empty so that $\theta_{i-1} < \theta_i$.

[5] Since the density $f(\cdot)$ is positive, it does not matter which words contain the endpoints of intervals.

A sufficient condition for the optimal language to group similar states together is for the loss function to satisfy the **single-crossing property**: if $a'' > a$" and $\theta'' > \theta'$, then

$$L(a'', \theta') \leq (<) L(a', \theta') \implies L(a'', \theta') \leq (<) L(a', \theta').$$

Proposition 2 *If L satisfies the single-crossing property, then there exists an optimal language that is an interval partition.*

The proposition suggests that there may be optimal languages that do not form an interval partition. This is true because one can change the assignment of states to words on a set of measure zero without increasing losses. The optimal language must be an interval partition if one requires the salesperson to pick the optimal word for all states.

I state the result when the state of the world is an element of the real line, but the argument generalizes to higher dimensions. That is, if the decision maker's objective is to minimize expected losses and the loss function is well behaved, then words will consist of "similar" objects. To make the statement precise, let A and Θ be compact, convex subsets of Euclidian spaces. Define the function $N(x, x', \theta) \equiv L(x, \theta) - L(x', \theta)$.

Proposition 3 *If $\{(\theta : N(x, x', \theta) > 0\}$ and $\{(\theta : N(x, x', \theta) \geq 0\}$ are convex for all x and x', then there exists an optimal language $\{W_1, \ldots, W_K\}$ such that W_i is convex for each i.*

Say that the function $N(x, x', \theta)$ is monotonic on rays if $N(x, x', \theta_0 + \lambda\theta)$ is a monotonic function of the scalar variable λ for all θ_0, θ, x, and x'. It is straightforward to verify that the conditions in Proposition 3 hold if $N(\cdot)$ is monotonic on rays.[6] In the one-dimensional case, sub or supermodularity of $L(\cdot)$ implies that N is monotonic on rays. In higher dimensions, the convexity assumptions are restrictive, but hold in the leading case where $L(a, \theta)$ is the Euclidean distance between a and θ.

Versions of Proposition 3 appear is the literature. Jäger, Koch-Metzger, and Riedel [11] study this model under the assumption that $L(a, \theta)$ is of the form $l(\|a - \theta\|)$, where $\| \cdot \|$ is a norm on $\mathbb{R}^n$ and $l(\cdot)$ is a continuous, convex, and strictly increasing function. They show that the conclusion

[6]In non-degenerate cases, $\{(\theta : N(x, x', \theta) \geq 0\}$ is the closure of $\{(\theta : N(x, x', \theta) > 0\}$ so that convexity of $\{(\theta : N(x, x', \theta) > 0\}$ implies convexity of $\{(\theta : N(x, x', \theta) \geq 0\}$.

of Proposition 3 holds for these preferences under the added condition that $\{(\theta : N(x, x', \theta) = 0\}$ is a set of measure zero. Saint-Paul [14] proves a similar result (he states the "monotonicity on rays" condition in a differential form).

It is straightforward to see that the optimal language in CGP will not satisfy the interval property in general. The cost minimizing language tends to place the most likely states together into precise words. If the most likely states are extreme, then in the optimal language, words will not be intervals. For a formal example, suppose that $K = 2$ and the prior distribution of states is symmetric around .5 and that $\delta \in (0, .5)$ and $\epsilon > 0$ are such that the probability of states in $[\delta, 1-\delta]$ is less than ϵ. The cost of having W_1 represent $[\delta, 1-\delta]$ and W_2 the complement, is

$$\lambda(W_1)Pr(\theta \in W_1) + \lambda(W_2)Pr(\theta \in W_2) \leq (1-2\delta)\epsilon + 2\delta, \tag{7}$$

while the cost of an interval partition is at least the cost associated with the partition element that contains .5, which is at least .25. Since (7) can be made arbitrarily close to zero, the interval property does not hold.

The example is not special. In fact, if the density is not monotonic, then there exists K such that the cost-minimizing CGP language is not an interval partition. If K is sufficiently large, there will exist a pair of words such that one of the words contains a positive measure of states that are both more and less likely than states in another word. It can be shown that it is possible to construct a new, lower cost, language in which the likely states are pooled together (so that the interval property fails).

On the other hand, the CGP cost-minimizing language has the interval property when the density of states is monotonic.

Proposition 4 *If the density is monotonic, the CGP cost-minimizing language $\{W_1, \ldots, W_K\}$ can be taken to be an interval partition. Furthermore, words are monotonic in length. That is, if $W_i = [\theta_{i-1}, \theta_i)$ for $\theta_{i-1} < \theta_i$, then $\theta_i - \theta_{i-1}$ is monotonic in i.*

As I discuss in the next section, CGP demonstrate that in their model more common states are pooled together. More common states are adjacent precisely when the density is monotonic.

There is a rich literature that compares color terms in different languages (see Berlin and Kay [3]). This literature provides strong evidence that color terms associate similar colors.[7] Gärdenfors [10] describes many other natural situations in which the interval property (and higher dimensional variations) holds.

Language does contain words for "extreme" states. One can say that on outcome is an outlier or that a result is n standard deviations from the mean, but these statements are probably more useful when combined with knowledge that indicates the direction of the outcome.[8] I am not aware of any study that associates the interval property with monotonicity of the density, however. I conclude that to the extent that the interval property is a natural property of real-world communication, it follows for reasons that are not included in the CGP model. My formulation provides a basis for the interval property under a condition that is commonly assumed in economic applications.

I have shown that in my model the interval property arises as a consequence of assumptions about the loss function. An alternative point of view is that the property arises as a consequence of cognitive limitations of the salesman. If the salesman is unable to distinguish between two sets, then he would be unable to separate the sets into two different words. One way to formalize this approach would be to assume that the salesman can only observe certain kinds of subsets of the state space and words must be elements of this set. This approach provides a partial justification for CGP's restriction that words must be intervals. If a salesman can only describe intervals, then the interval property must hold. This restriction is not appropriate, however, if one assumes that the set of describable subsets of states is a $\sigma-$algebra. If the salesman can identify intervals and unions of intervals (coining a word for states that are "either high or low" for example), then the interval property

[7] Berlin and Kay [3] provide evidence that color terms evolve in common ways independent of cultural or linguistic context. The data suggest that color terms enter a language in a well defined order. For example, languages that have a word for "green" will also have a word for "red." For the purposes of this paper, the relevance of the work on basic color terms is that the one can construct a lower-dimensional model of colors with the property that color terms typically describe a convex subset of this space. In particular, color terms lump together tones with similar frequencies.

[8] The interval property is not universal. One class of exceptions are "periodic" properties: people with birthday in January (independent of the year). Even in cases like this one might conjecture that payoff-relevant versions of states are pooled together, so that the interval property holds in a reduced-form version of the model. Alternatively, in the birthdate example items are pooled together in a convex set when dates are described by a pair consisting of the year and month of birth.

will fail.

4 Other Properties of the Models

In this section I discuss the extent to which properties of optimal codes in CGP carry over to my model.

4.1 Precise Words

The optimal language in CGP must satisfy two properties:

1. more likely states are pooled together in more precise words; and
2. more precise words are used more frequently.

Stated in terms of interval partitions, these properties are

CGP-1 if $\theta \in W_i$, $\theta' \in W_j$, and $f(\theta_i) < f(\theta_j)$, then $\theta_i - \theta_{i-1} > \theta_j - \theta_{j-1}$; and

CGP-2 if $\theta_i - \theta_{i-1} > \theta_j - \theta_{j-1}$, then $F(\theta_i) - F(\theta_{i-1}) < F(\theta_j) - F(\theta_{j-1})$

These properties do not hold without additional assumptions in my model. Even if the density on the states of the world is monotonic, it may be the case that smaller words contain unlikely states. The reason for this is clear: In my formulation, losses are not directly associated with the length of the word. Instead, losses depend on the value of the information provided by the word.

Example 1 *Suppose that the state θ is uniformly distributed on $[0,1]$, that the action a is an element of $[0,1]$, and that the loss function is $L(a,\theta) = (a - \theta - .5)^2$. A simple computation identifies the optimal two-word language as $\{[0,1/3),[1/3,1]\}$. That is, although the states are equally likely it is optimal to use one word (referring to higher states) more common. The optimal language is not symmetric because information about the high states is not as valuable as information about low states. This follows because actions are bounded by 1, so that it is always optimal to take the action $a = 1$ when it is known that $\theta \geq .5$. Starting from equally precise words, making the lower word more precise has a first-order benefit, but making the higher word less precise has no first-order cost. The qualitative feature of this example (low word less likely that high word) would remain even if the density was strictly increasing.*

CGP's first result holds under some conditions. Assume that the loss function $L(a, \theta)$ can be written $g(a - \theta)$ where $g(\cdot)$ is a differentiable, strictly convex function satisfying $g(0) = 0$. (Functions of the form $g(x) = |x|^{\alpha}$ for $\alpha > 1$ satisfy these conditions.) If the density of states is monotonic, then in the optimal language less frequent states are pooled in larger intervals.

Proposition 5 *Assume that $L(a, \theta) \equiv g(a - \theta)$ where $g(\cdot)$ is a differentiable, strictly convex, symmetric function minimized at 0. If the density of states is monotonic, then in the optimal language bigger intervals are associated with less frequent words.*

Here is a sketch of the proof of Proposition 5. Suppose $W_i = [\theta_{i-1}, \theta_i)$ and a_i is the action taken in response to W_i. When the density is decreasing, low states in an interval are more likely than high states. When losses are a convex function of the distance between the action and the true state, this causes the optimal action to be below the midpoint of the interval. The first-order condition for Problem 1 is $g(\theta_i - a_i) = g(a_{i+1} - \theta_i)$ so that the state that divides two adjacent words is equidistant to the two adjacent actions, since both actions must be below the (uniform) average state in a word, W_{i+1} must pool together more states that W_i.

The assumption that the density is monotonic is a strong condition, but Proposition 4 demonstrates that it is necessary. On the other hand, a local version of the result holds. Suppose that the density of states in monotonic on an interval $[a, b]$ and W_1 and W_2 are adjacent words contained in this interval ($W_1 = [a_1, b_1], W_2 = [a_2, b_2]$, with $a \le a_1$; $b_1 = a_2$ and $b_2 \le b$). In the optimal language $b_1 - a_1 \ge b_2 - a_2$ if the density decreases and $b_2 - a_2 \ge b_1 - a_1$ if the density increases.

Example 1 fails to satisfy the conditions of Proposition 5 because it restricts the set of actions to $[0, 1]$. If agents could take any real action, then it is valuable to distinguish between states known to be greater than one half. The optimal two-word language would partition states into those less than one half and those greater than one half.

Proposition 5 provides conditions under which CGP-1 holds in my model. I cannot provide conditions under which CGP-2 holds. One can show that CGP-2 need not hold even in restricted examples in which $L(a, \theta) = -(a - \theta)^2$ and $F(\theta) = \theta^{\alpha}$.

4.2 Complexity

I conclude this section with a discussion of how the cost of the optimal language varies with the prior.

CGP present a nice result on the complexity of decision problems. Fix the state space and the cardinality of the language. Associated with each probability distribution over states, $F(\cdot)$, one can compute the minimum cost of associated with the optimal language, $V^*(F)$. CGP give a simple comparative-statics property of $V^*(\cdot)$. Assume that one can order the states so that higher states are more likely. This ordering is without loss of generality in a model with a finite number of states. Interpret $F(\theta)$ as the probability that the state is less than or equal to θ. The cumulative distribution function $\tilde{F}$ dominates F (in the sense of first-order stochastic dominance) if $F(\theta) \geq \tilde{F}(\theta)$ for all θ. CGP prove that if $\tilde{F}$ dominates F, then the problem associate with F is more complicated than the problem associated with $\tilde{F}$ in the sense that $V(F) \geq V(\tilde{F})$. That is, when probability shifts from less likely to more likely events, then the cost associated with bounded rationality decreases. The intuition for this result is that the optimal language under F uses more precise words for more likely events. Hence, losses are lower when likely events happen. Shifting probability to more likely events reduces the losses associated with this language.

The complexity result does not hold in my model. To see this, consider a case in which $K = 1$, $L(a, \theta) = (a - \theta)^2$, and $F(\cdot)$ is the uniform distribution on $[0, 1]$. The optimal action will be $a = .5$. Shifting probability mass from a small interval around .5 to an interval near 1 will lead to a distribution that stochastically dominates $F(\cdot)$, but it will place weight from states for which $a = .5$ is a good decision to one in which it is a bad decision. Intuitively, in my model complexity is more closely related to the dispersion of F. One can generalize the example to arbitrary K. The solution to my problem will induce K words (W_i, $i = 1, \ldots, K$) and K actions (a_i best responses to W_i). Any shift in the prior distribution that moves probability mass in an interval W_i closer to a_i will lower the cost of the problem. It will always be possible to find such a shift that leads to a first-order stochastically dominating shift (in either direction).

The next result describes how to rank distributions in my model. Given a distribution function $F(\cdot)$ defined on $[0, 1]$, $\tilde{F}$ is a **compression** of F if there exists a $\rho \in (0, 1)$ such that $\tilde{F}(\rho\theta) \equiv F(\theta)$ for all $\theta \in [0, 1]$. Plainly, $\tilde{F}(\cdot)$ is a distribution function on $[0, \rho]$.

Proposition 6 *Assume that* $L(a,\theta) \equiv g(|a-\theta|)$ *where* $g(\cdot)$ *is strictly increasing on* $(0,\infty)$*. If* $\tilde{F}$ *is a compression of* F*, then the losses associated with* $\tilde{F}$ *are less than the losses associated with* F*.*

Proposition 6 states that compressing the state space reduces losses. The assumptions in the proposition are strong, but I could not prove a more general statement. The intuition for the result is clear. Compressing the state space reduces losses.

5 Communication in Organizations

The basic model assumes that there are just two agents: a single salesman and a single engineer, but the model has implications for more complicated interactions. CGP enrich their model by assuming that there is more than one distribution that generates problems. The simplest extension is to assume that there are two salesman-engineer pairs, with the pairs distinguished only by the prior distribution over states. I will call a salesman-engineer pair a **division**. If the divisions operate in isolation and both face identical complexity restrictions (limits on the cardinality of language), then it will generally be optimal to use different languages for the each division. CGP assume that there are gains associated with integrating the pairs.[9] Costs arise because the integrated organization is still constrained to use a language with a fixed set of words. Losses associated with coarse communication will be greater in the integrated organization.

This section reviews CGP's model of communication in organizations in order to observe that the qualitative properties identified in CGP do not depend on their specification of complexity costs. These properties will hold for my formulation of costs and for more general formulations. Hence the message of this section is different from that of the previous sections. Up until now, I have demonstrated that the properties of CGP-optimal codes generally do not hold if one wants to be able to deduce the interval property. This section argues that important implications for communication in organizations deduced by CGP are consistent with the interval property.

[9] CGP provide one specific motivation for this assumption. In a situation in which customers arrive at appropriate salesman but depart without making a purchase if the salesman is occupied, integration allows the organization to increase the total number of customers served because with positive probability a customer can be diverted to the other salesman when her preferred salesman is occupied.

5.1 Integration versus Separation

It is straightforward to summarize CGP's insight in precise terms and demonstrate that it does not depend on the way in which they formalize communication cost. They assume that all customers generate revenue (normalized to) one and the cost to serve a customer is β times the complexity cost. They denote this cost $D^*(\cdot)$. $D^*(\cdot)$ is a function of the probability distribution over states (and K, which is held fixed). They assume that there are two populations of customers, v_i is the fraction of type i customers and $f_i(\cdot)$ is the probability distribution of states for type i customers. They compare profits in a firm in which a common language allows both engineers to deal with all customers (organization C for communication) to one in which one salesman-engineer pair i have their own $K-$word language and deal exclusively with type i customers (organization NC for no communication). They assume that the form of the organization determines the expected number of customers so that $q^C > q^{NC}$, where q^C denotes the expected number of customers with communication and q^{NC} the expected number of customers without communication. Given this formulation, CGP deduce that profits are equal to

$$q^{NC}(1 - \beta\,(v_1 D^*(f_1) + v_2 D^*(f_2))) \tag{8}$$

without communication and to

$$q^C(1 - \beta D^*(v_1 f_1 + v_2 f_2)) \tag{9}$$

with communication. CGP use (8) and (9) to deduce when integration is more profitable than separation. Given these formulas, their conclusions are straightforward: Increases in the gains from communication (as measured by q^C/q^{NC}) or reductions in losses associated with complexity (as measured by β), make integration more profitable relative to separation.

The only property of the optimal complexity cost function used in the analysis is concavity: $D^*(v_1 f_1 + v_2 f_2)) \geq v_1 D^*(f_1) + v_2 D^*(f_2)$. Concavity also holds if complexity costs are derived using my formulation,[10] so their qualitative properties also hold in my model.

[10] Concavity follows because it is feasible for both pairs to use the same language without communication as they would optimally choose under communication.

5.2 Hierarchy

CGP also consider a hierarchical form of organization under which the firm can serve q^{NC} customers, but salesman-engineer pairs use different languages. Engineer i can serve a customer handled by Salesman j, $j \neq i$ because the hierarchical organization hires a costly translator. The translator can turn any message from the salesman into a word in the (appropriate) engineer's language. Denoting the cost of hiring a translator as μ, the profits from a hierarchical organization are

$$q^C \left(1 - \beta \left(\tilde{v}_1 D^*(\tilde{f}_1) + \tilde{v}_2 D^*(\tilde{f}_2)\right)\right) - \mu \tag{10}$$

where $\tilde{v}_1 = v_1\phi + v_2(1-\phi)$ and $\tilde{v}_2 = 1 - \tilde{v}_1$, while and

$$\tilde{f}_1 = \frac{v_1\phi f_1 + v_2(1-\phi)f_2}{v_1} \text{ and } \tilde{f}_2 = \frac{v_1(1-\phi)f_1 + v_2\phi f_2}{v_2}$$

where $\phi \in (0,1)$ is an exogenous parameter than determines how effectively the translator can allocate customers from one salesman to the corresponding engineer (when $\phi = 1$, Engineer 1 sees only customers from Salesman 1).

Hierarchical organizations serve the same number of customers as integrated organizations. Relative to integrated organizations, they have a cost and a benefit. The cost is that translation services are costly. The benefit is that customers are partially sorted (so that different engineers can use somewhat specialized languages). Relative to separated organizations, hierarchies have the advantage of serving more customers and the disadvantage of incurring higher complexity costs (because languages are less specialized in hierarchical organizations than in separated organization). Once again, the qualitative features depend on concavity of D^*, but not in the precise model that generates D^*.[11]

5.3 Comparison

CGP (page 373) provide this summary of their contribution:

> A broader organizational scope allows for more synergies to be captured, but reduces within-unit efficiency, since it requires a

[11] CGP understand the importance of concavity of D^*. They identify the property in Lemma 1 of their paper (page 389). My point is only to emphasize that concavity of D^* does not depend on the specific form of the loss function that gives rise to costs.

> more generic language. A manager working as specialized translator may also be used to achieve between-unit coordination while maintaining separate languages. Our theory reconciles two recent well-documented phenomena within organizations: the recent increase in information centralization and the reduction in hierarchical centralization.

In this section I explain that the increase in informational centralization and the reduction in hierarchical centralization follow from a property of their model, concavity of the optimal loss function, that does not depend on their precise formulation of losses. If one interprets lower cost as lower weight on the loss function (reduction is β), then integration becomes more attractive because the value of having specialized language goes down. CGP associate reductions in hierarchical centralization with less reliance on translation. The property that makes reductions in β favor integration over hierarchical organization is concavity of D^*.

The model of translation raises issues that neither CGP nor I pursue. These considerations probably do not influence the qualitative conclusion that hierarchical organizations are optimal in "intermediate" situations (in terms of q^C/q^{NC} and β), but may be important in other ways. First, one might expect that the cost of the translator depends on the context. The closer are the distributions f_1 and f_2, the easier it is to carry out the translation. Second, translation may have qualitatively different properties depending on whether the interval property holds. In CGP's model, one would expect the optimal K word language for two different salesman-engineer pairs to be completely different. When $K = 2$ and the state space contains N elements, for example, there are $2^N - 1$ qualitatively different languages and no natural way that a word in one language corresponds to a word in another language. Translation seems relatively difficult. On the other hand, if the states are ordered and the optimal language satisfies the interval property, then when $K = 2$ there are $N - 1$ different languages. All of these languages classify some states as "low" and the others as "high." So on one hand translation in this environment seems easier (because there are fewer possible optimal codes), but on the other hand when both languages satisfy the interval property, mistakes in translation seem possible. That is, when languages satisfy the interval property, individuals can get confused about the meaning of "low," but in the general case, words from different languages could be completely non comparable. This could lead to more failures to communicate, but fewer misunderstandings.

6 Coordination with Costly Communication

Section 5 applies a model of costly communication to the study of organizations. It points out that the qualitative conclusions of CGP hold for more general models of the complexity of communication. Since my model contains an explicit model of decision making, one can use it to study how to allocate decision-making authority within an organization. I illustrate this possibility using a variation of a model studied by Alonso, Dessein, and Matouschek [1] (ADM) and Rantakari [13] (R).

6.1 A Model of Decision Making in Organizations

There are two agents, $i = 1$ and 2. Each agent has private information: Agent i observes θ_i. ADM and R assume that the θ_i are independently distributed on an interval symmetric around 0. (I only require that the distributions be independently distributed and I denote the support by $[\underline{\theta}_i, \overline{\theta}_i]$.) The organization must take two decisions, a_i, for $i = 1$ and 2. Profit to Agent i is $\pi_i = K_i - (a_i - \theta_i)^2 - \delta(a_1 - a_2)^2$, where K_i is a positive constant.[12] ADM study two institutions for making decisions in the organization. When decision making is **centralized**, the agents observe their private information and simultaneously send a cheap-talk message to a center. That is, they send a message $s_i(\theta_i) = m_i$; the messages partition the sets into words of the form $\{\theta_i : s_i(\theta_i) = m_i\}$. The center then selects actions to maximize $\pi_1 + \pi_2$. When decision making is **decentralized**, the agents observe their private information and simultaneously exchange cheap-talk messages. Agent i then selects a_i to maximize $\lambda\pi_i + (1 - \lambda)\pi_j$. In addition to these institutions, R studies two additional organizations. Under **partial delegation**, the center makes the decision of Agent i while Agent j, $j \neq i$ makes the decision for division j. Under **directional authority**, one of the divisions makes both actions. ADM assume that Agent i seeks to maximizes a weighted average of π_i and π_j, leading to the objective function $\lambda\pi_i + (1 - \lambda)\pi_j$, where $\lambda \geq .5$, while the center maximizes $\pi_1 + \pi_2$.

Centralized decision making has two potential advantages. First, since the center makes both decisions, coordination is simple. Second, since the

[12] Rantakari permits δ to depend on i.

conflict of interest between an agent and the center is not as great as the conflict between the two agents, centralized communication is more effective than decentralized communication.[13] Decentralized decision making has the advantage that Agent i has perfect information about θ_i. One might conjecture that since an increase in δ increases the importance of coordination, increasing δ will increase the value of centralized decision making relative to decentralized decision making. ADM and R show that this intuition is not correct. ADM point out that increases in δ reduce the conflict between the agents (since they both care more about coordination), which in turn leads to improved direct communication. Partial delegation and directional authority have some of the strengths and some of the weaknesses of decentralization and centralization. R shows that there exist preference parameters under which any one of the four organizations may be optimal.

I modify ADM and R's problem in two ways. First, I assume that $\lambda = .5$. This eliminates the conflict of interest between the agents.[14] Hence the common loss function is

$$.5(a_1 - \theta_1)^2 + .5(a_2 - \theta_2)^2 + \delta(\theta_1 - \theta_2)^2. \tag{11}$$

Second, I assume that complexity costs limit communication. Following CGP, I assume that there are only a finite number of messages available to the organization. In this setting, centralization still has the advantage of permitting coordinated decision making. Decentralization still has the advantage of allowing an agent to make a better decision. Decentralization will clearly be optimal when $\delta = 0$. In this case, communication is unnecessary and decentralized decision making leads to the first-best option. When there is no conflict of interest, centralization and partial delegation are dominated by directional authority. Consider the optimal language under either centralization or partial delegation. The same language is feasible when under directional authority. Concentrating authority for both decisions in the hands of one of the agent's (in the case of centralization) or the agent who makes

[13]This assertion hides subtleties. The preferences of an individual agent are more similar to those of the center than to those of the other agent because the center's preferences are the average of the preferences of the two agents. ADM demonstrate that there always exists an equilibrium with centralized communication that is "more informative" under centralized decision than the "most informative" equilibrium with decentralized communication. A more informative distribution of θ_i (in this context) is one that leads to a lower expected value of $(a_i - \theta_i)^2$.

[14]ADM's analysis includes the case $\lambda = .5$.

the decision for his own division (in the case of partial delegation) cannot raise the expected losses faced by the organization. Since there is no conflict of interest, the decision maker learns as much as the center would about the other division's information. On the other hand, partial delegation permits the decision maker to use the information about his own state and allows decisions to be coordinated (since the same agent takes both decisions).

In the ADM-R model, what limits communication is conflict of interest, not complexity cost. If $\delta = 0$, nothing would prevent the agents from fully sharing their information. There would be no role for centralization. ADM-R study communication equilibria when there is a conflict of interest between the informed agents. Notice that in this case, there will also be a conflict of interest between an informed agent and the central planner. Intuitively, this conflict is smaller than the conflict between the agents. ADM formalize this intuition and demonstrate that there can be more effective communication between agents and the center than directly between agents in equilibrium. Perhaps surprisingly, however, they show that as coordination increases in importance (δ increases), decentralized communication leads to a higher payoff for the organization than centralized communication. The intuition for this result is that when agents care more about coordination, they are more willing to communicate directly and are also inclined to coordinate their actions even when they have the autonomy to choose their actions independently.

In this section, I characterize the optimal language and demonstrate when it is advantageous to grant all decision making authority to one of the agents. I show that if one does give decision making authority to a single informed agent, it is better to give the authority to the agent with more precise prior information.

6.2 Characterization of Optimal Organizations

Consider the case where a single decision maker makes both decisions. A straightforward computation (included in ADM and R) establishes that conditioned on the messages m_1 and m_2, the decision rule satisfies:

$$a_i^C = \gamma^C E[\theta_i \mid m_i] + (1 - \gamma^C) E[\theta_j \mid m_j], \tag{12}$$

where

$$\gamma^C = \frac{1 + 2\delta}{1 + 4\delta}.$$

From (12) a computation leads to an expression for the common objective function. I include the details in the appendix and summarize the result in the following lemma.

Lemma 1 *Given message rules s_1, s_2, let $K_i(m_i) = \{\theta_i : s_i(\theta_i) = m_i\}$. The expected value of the organization can be written*

$$c_R(R_1^C + R_2^C) + c_H H, \tag{13}$$

where, for $i, j = 1$ and 2, and $j \neq i$,

$$R_i^C \equiv \sum_{i=1}^{K} \int_{K_i(m_i)} (E[\theta_i \mid m_i] - \theta_i)^2 \, dF_i(\theta_i)$$

and

$$H \equiv \int \int (\theta_1 - \theta_2)^2 \, dF_1(\theta_i) \, dF_2(\theta_2),$$

and $c_R = .5\left((\gamma^C)^2 + (1-\gamma^C)^2\right) + \delta\left(2\gamma^C - 1\right)^2$, $c_H = (1-\gamma^C)^2 + \delta(2\gamma^C - 1)^2$.

R_1^C and R_2^C are residual variances. It is clear that $R_i^C \geq 0$. These quantities depend on the information structure and, intuitively, measure how much information is lost because of communication constraints. Notice that the residual variance is computed from the point of view of the decision maker. The quantity H measures the difference between the agents' information. It does not depend on the choice of code. Increases in this term make coordination more difficult under all organizations.

I have displayed the formula assuming that $\lambda = .5$, but one can carry out the computations for all λ. Hence an expression like (13) holds for the models of ADM and R.[15]

Proposition 7 *When agents have common loss functions, the optimal language for centralized decision making in the coordination model minimizes*

[15] See Proposition 4 in ADM and Proposition A3 in R for similar expressions. Both expressions require that actions are best responses (that is, (12) holds), but ADM-R's expression also uses properties of the equilibrium communication structure that I do not yet impose. Here the expressions in ADM-R are not identical to mine. The computation of equilibrium in ADM and R is different (and more difficult) than mine because when there is a conflict of interest, the equilibrium information structure will not generally be selected to minimize losses. Instead, it must satisfy constraints that reflect the incentives agents have to misrepresent their private information.

the sum of residual variance, $R_1^C + R_2^C$. Losses are lower under directional authority than when the center makes the decisions. If one agent's information is a compression of the other agent's information, losses are lower if the other agent makes the decision. Losses are decreasing in the number of messages and increasing in the expected difference in the agents' information (H).

Computing the optimal information structure is a straightforward optimization problem. Directional authority is superior to centralization because under directional authority, $R_i^C = 0$ when i is the decision maker. Since the loss function is quadratic, the optimal language will have the interval property. If the distribution of θ_j is uniform (as ADM and R assume) and authority is delegated to Agent i, then the optimal language divides the state space into K equal sized intervals.

When decision making is centralized and there are no conflicts of interest, it is best to have one of the agents (rather than a central location that has no private information) make both decisions. Let V_i^C minimize R_i^C over all K-word languages. Agent i should make decisions if $V_i^C > V_j^C$. The intuition is straightforward. The decision maker has complete information about her division, but noisy information about the other division.

A qualitatively similar computation makes it possible to characterize the optimal language when informed players first exchange messages and then make decisions. In this situation, both decisions are made with full information about one component of the state of nature, but they are not coordinated. This suggests that decentralization would be advantageous when using local information optimally is important (low δ) and centralization is superior when coordination is important. Again, the algebra confirms this intuition.

Let $a_i^D(\theta_i, m_j)$ be the decision of Agent i in the state θ_i when the message of the other agent is m_j.[16] Routine computations establish that (for $i = 1, 2$, $j \neq i$)

$$a_i^D(\theta_i, m_j) = \gamma^C \eta \theta_i + (1 - \gamma^C) E[\theta_j \mid m] + (1 - \gamma^C)(1 - \eta) E[\theta_i \mid m].$$

where, $\eta = (1 + 4\delta)/(1 + 2\delta)^2$ and, as before, $\gamma^C = (1 + 2\delta)/(1 + 4\delta)$.

[16] The decision rule a_i^D is, in principle, a function of both messages and θ_i, but since θ_i determines m_j I do not explicitly include the dependence on θ_j.

As in the case of centralization, it is routine to substitute these expressions into the objective function of the organization to obtain a formula for the organization's losses. The computation uses the independence of θ_i and θ_j and the fact that $E\left(E[\theta_i \mid m_i] - \theta_i\right) = 0$. After simplification, losses can be written:

$$d_R(R_1^D + R_2^D) + d_H H, \tag{14}$$

where, lett $d_R = .5\left((1-\gamma^C)^2 + (\gamma^C)^2\eta^2\right) + \delta(\gamma^C\eta - (1-\gamma^C))^2$ and $d_H = c_H$.

Proposition 8 *When agents have common loss functions, the optimal language for decentralized decision making minimizes the sum of residual variance, $R_1^D + R_2^D$. Losses are decreasing in the number of messages and increasing in the expected difference in the agents' information (H).*

Propositions 7 and 8 demonstrate that the organization's objective will be to minimize a sum of residual variances $(R_1 + R_2)$ and the expected difference between the two distributions (H). The H term does not depend on the form of communication. The formulas are consistent with a strong intuition. When communication is perfect, residual variances will be zero. There may still be losses because perfect coordination is infeasible. These losses should be independent of the form of the organization. This explains why $d_H = c_H$. It is straightforward to check that $0 \leq d_R \leq c_R$. Losses associated with limited information are smaller under decentralization because decision makers have more local information. The fact that $d_R \leq c_R$ does not imply that decentralization is necessarily superior to (the best form of) centralized decision making because under it is possible for $R_i^C < R_i^D$. This happens when decentralization degrades the information (about the other unit) available to the decision maker.

When δ is small (so that coordination of actions is not important), decentralization is superior to directional authority for a familiar reason. Decentralization can take advantage of all local information and when δ is small, there are no losses from failure to coordinate. When δ is large, directional authority can be superior. The reason for this is a bit more subtle. If Agent i is the decision maker under directional authority, then $R_i^D = 0$, while (because K is finite) $R_i^C > 0$. On the other hand, $R_j^C = R_j^D$. Hence decision makers will have better information about the other division under directional authority, leading to the possibility of better communication.

6.3 The Interpretation of Decision Authority

My analysis highlights the importance of directional authority. Directional authority is always superior to a centralized organization in which the decision maker must rely on signals to learn about both states. This observation is mathematically obvious; I have pointed out that it follows from simple dominance. It may not be economically relevant. Directional authority is not a common form of organization. I can think of two reasons for this. First, the superiority of directional authority to centralization depends on the assumption that there is no conflict of interest. When there is conflict, as Rantakari [13] shows, centralization may be superior. The possibility that centralization may improve communication relative to directional authority requires the existence of conflict. The second reason is that decision making is a specialized activity. It may be that only agents with special ability are capable of making decisions.

There are several ways in which one might model the possibility that decision-making authority is limited. One could assume that in each unit there is a separation between informed agents and decision makers. This corresponds to the division between salesmen and engineers in CGP. With this interpretation, one can compare a decentralized (or specialized) organization in which each unit uses an optimal language to determine its action, but there is no communication across divisions. This model creates a trade off between the efficiency gains of having specialized languages for the two units versus the costs associated with having poor information about the other unit. In such an environment, centralization is optimal when the two units face the same kind of problem (the distributions of θ_1 and θ_2 are similar) while decentralization is optimal when coordination is not important. When the problems are similar, the losses associated with using the same K-word language in two divisions is small. I do not include the analysis, which is similar to the construction in Section 6.2.[17]

Another possibility is to assume that agents can make local decisions, but not global ones. That is, Agent i can pick a_i but not a_j, for $j \neq i$. To make this specification non-trivial, one must also assume that the center is able to make both decisions. This interpretation is consistent with the interpretation centralization in ADM and is also consistent with real-world organizations. It is a consequence of the analysis in Section 6.2 that when

[17] Under decentralization, Division i's optimal action is a weighted average of the (unconditional) means of θ_1 and θ_2 and the mean of θ_i conditioned on the message received.

there are no conflicts of interest, any communication structure feasible for centralized decision making is feasible for decentralized decision making but (because of the existence of local information) any information structure will lead to lower residual variances under decentralization than under centralization. Since $c_R \geq d_R$, this means that decentralization must be at least as good as centralization. Intuitively, the advantage that centralization has is that it can use better information to coordinate decisions. This result has a substantively important implication: Complexity of communication (as I have modeled it) is not by itself sufficient to justify centralized decision-making authority. In order to emphasize that centralization (as opposed to delegated authority) is inferior to decentralization in my model, I state the observation as a proposition.

Proposition 9 *The expected losses under the optimal communication structure under decentralization are no greater than the losses under centralization.*

As ADM-R demonstrate, the conclusion of Proposition 9 need not hold when there are conflicts of interest between divisions. The conclusion also depends on the (implicit) assumption that the center has no advantage in implementing decisions or processing information.

6.4 Complexity versus Conflict

I have revisited questions raised by ADM-R in a setting where communication may be incomplete due to complexity rather than conflict of interest. On an analytical level, the problems are similar. Organizations have identical objective functions, but face different constraints: incentive constraints in ADM-R and complexity constraints in my model. Perhaps the biggest qualitative conclusion is that my model removes any rationale for centralized decision making when directed authority is feasible. It is plausible to attribute the use of centralized authority to the existence of different preferences.[18] A complete model would allow for both conflicts of interest and limitations of communication. I hope that it is constructive to study each factor separately.

[18]There are alternative explanations that do not require conflict of interest. For example, the ability to make division-specific decisions may be a specialized skill that is not available throughout the organization.

6.5 Comparison to CGP

I have described the optimal organization given that the firm is integrated. To compare the results to those in the previous section, one can also imagine that there is a decision whether to integrate. Separate divisions need not coordinate actions. Division i can take the action $a_i = \theta_i$ without communication. This would lead to a total profit of q^S. Profits for an integrated firm would be the expected value of

$$q^I \left(K - 5(a_1 - \theta_1)^2 - .5(a_2 - \theta_2)^2 + \delta(\theta_1 - \theta_2)^2 \right),$$

where the expectation is taken assuming the optimal organization. With this specification, it becomes clear that ADM-R and CGP present complementary analyses that illustrate the role of costly communication in the design of organizations. In CGP, increases in the relative number of customers one can serve or reductions in losses associated with complexity make integration more profitable relative to separation. Separation is attractive if the distribution of states across divisions is sufficiently different that it is useful to preserve separate languages. In ADM-R the barrier to integration is the need to coordinate actions across divisions. Integration is attractive if the need to coordinate is large (large δ), if integration generates relatively more customers (q^I/q^S is large), or if communication is not complex (K is large). Hence in both cases, integration is more attractive when, loosely, integration leads to greater demand for services, the divisions involved are more similar, and the complexity of communication is smaller.

6.6 Different Notions of Complexity

Dessein, Galeotti and Santos [8] study a team decision problem in which agents have private information about an idiosyncratic target. The team objective is to minimize a weighted sum of losses, where losses arise either if a team-member's action differs from her idiosyncratic target or if the actions of team members are not coordinated. This leads to an objective function similar to the one that arises in ADM's model without conflict of interest.[19] Like me, Dessein, Galeotti, and Santos study the impact of limited ability to communicate on outcomes. In their model, individuals have "local" information,

[19] Dessein, Galeotti, and Santos follow the formulation of Dessein and Santos [9] that gives rise to a slightly different objective function.

which they can share through communication. Communication is noisy, but allocating more resources to communication reduces noise. There is a fixed limit on the amount of time that can be devoted to communication. In this framework, Dessein, Galeotti, and Santos contrast balanced communication (when different agents communicate equally) and focused communication. In the two-agent model, they present conditions under which a leader makes decisions that are responsive to his local information and the follower is able to coordinate because the organization allocates its limited communication resources to enable the follower to learn the leader's local information. In the multi-agent model, they generalize this insight and show the exogenous formation of leaders (agents who communicate to many sources). The analyses carry a somewhat similar message: Limited communication may lead to organizations in which it is optimal to create asymmetries between agents. In my model, certain agents have extra authority to make decisions. In Dessein, Galeotti, and Santos, certain agents have access to more information. It would be straightforward to combine these models – either by changing the timing and allocation of decision rights in Dessein, Galeotti, and Santos or my changing the complexity costs in my model.[20]

In this section I compared the performance of different organizational forms when communication is costly. It is worth noting that I chose a particular way to model costly communication. Following CGP, I assumed that what constrains communication is a finite set of words. It would be natural to investigate the implications of models in which it is expensive to produce words so that the bound K arises endogenously. In models with one division, it is straightforward to carry out this analysis. In organizations, one must decide who bears the cost of the new words. If it is an organizational decision and all agents have access to the same language, then there is no need to change the basic analysis. If instead, individuals can make investments, one might expect divisions that face more complicated problems to develop richer languages. This would create pressure for some divisions to be good at coding or sending messages while others are good at decoding or receiving. Even if one maintains the assumption that communication capacity is fixed, it is not clear that the number of words is the appropriate measure of complexity. Dessein, Galeotti, and Santos model limits on communication by the

[20]The natural way to do this is to assume that K bounds the number of distinct messages instead of the number of distinct words. With this modification it would be costly for two agents to use the same word. Following CGP, my model assumes that it is not costly for another agent to use an existing word.

amount of time an individual talks. One could also limit communication by the number of different actions that can be induced.[21]

6.7 Other Variations

ADM's conclusions about centralization depend strongly on the choice of objective function. ADM discuss the sensitivity of their conclusions to modeling assumptions in Section 7 of their paper and demonstrate that the arguments for decentralization are robust to variations in their assumptions. I mention a somewhat different example. Assume that the losses of Agent i are $(a_1 - \lambda\theta_i - (1-\lambda)\theta_j)^2 + (a_2 - \lambda\theta_i - (1-\lambda)\theta_j)^2$, while the center seeks to minimize the average of these losses. In this case, each agent wants both actions to be close to a target and the target is an average of θ_1 and θ_2. Agents differ because they have different targets. If $\lambda > .5$ an agent's own observation receives more weight in determining the target. If θ_i is uniformly distributed on $[-1, 1]$ and $\lambda = 1$, then decentralized decision making leads to the $a_i = \theta_i$. When decision making is centralized, there exists an equilibrium in which Agent i truthfully reports the sign of θ_i. This increases the center's payoff.

I did not optimize over all possible organizations forms. Following Rantakari, I emphasized the importance of directional authority in team problems. One could imagine situations in which communication is sequential and different divisions determine how much to say and who should act on the basis of past decision. For example, decentralized decision making could be the default organization, but a division could have the freedom to send a message indicating that additional consultation is needed. Decision makers would draw inferences from silence and costly deliberation could be reduced.[22]

[21] In the one division model, the number of words in the language is equal to the number of actions induced. When there are K words and two divisions, a centralized decision maker takes K^2 action and when there is decentralization, the action rule of at least one decision maker can take on a continuum of actions.

[22] Notice that if communication is free, alternative forms of organizational communication may be beneficial in ADM-R.

References

[1] Ricardo Alonso, Wouter Dessein, and Niko Matouschek. When does coordination require centralization? *American Economic Review*, 98(1):145–179, March 2008.

[2] Kenneth J. Arrow. *The Limits of Organization*. Norton, New York, NY, 1974.

[3] Brent Berlin and Paul Kay. *Basic Color Terms*. University of California Press, Berkeley and Los Angeles, 1969.

[4] Andreas Blume and Oliver Board. Common knowledge of language and communication success. Technical report, University of Arizona, May 2013.

[5] Andreas Blume and Oliver J. Board. Language barriers. *Econometrica*, 2013.

[6] Ying Chen and Sidartha Gordon. Information transmission in nested sender-receiver games. *Economic Theory*, forthcoming.

[7] Jacques Crémer, Luis Garicano, and Andrea Prat. Language and the theory of the firm. *The Quarterly Journal of Economics*, 122(1):373–407, 2007.

[8] Wouter Dessein, Andrea Galeotti, and Tano Santos. Rational inattention and organizational focus. Technical report, Columbia University, 2014.

[9] Wouter Dessein and Tano Santos. Adaptive organizations. *Journal of Political Economy*, 114(5):956–995, 2006.

[10] Peter Gärdenfors. *Conceptual Spaces*. MIT Press, Cambridge, Massachusetts, 2000.

[11] Gerhard Jäger, Lars Koch-Metzger, and Frank Riedel. Voronoi languages: Equilibria in cheap talk games with high-dimensional types and few signals. *Games and Economic Behavior*, 73:517–537, 2011.

[12] John Kieffer. Uniqueness of locally optimal quantizer for log-concave density and convex error weighting function. *Information Theory, IEEE Transactions on*, 29(1):42–47, 1983.

[13] Heikki Rantakari. Governing adaptation. *Review of Economic Studies*, 75:1257—1285, 2008.

[14] Gilles Saint-Paul. A "quantized" approach to rational inattention. Technical report, Paris School of Economics, 2014.

[15] Dezsö Szalay. Strategic information transmission and stochastic orders. Technical report, University of Bonn, September 2012.

Appendix

Proof of Claim that Deterministic Codes Minimize Costs. Given a solution to the optimization problem:

$$\min \sum_{i=1}^{K} \int L(a_i^*, \theta)\, d\mu(\theta \mid k) \tag{15}$$

where

$$a_i^* \ \textit{solves} \ \min_{a \in A} \int L(a, \theta)\, d\mu(\theta \mid k)$$

(and the $\mu(\cdot)$ are derived using Bayes's Rule (4) from a function σ), define a partition in which

$$W_i = \{\theta : L(a_i^*, \theta) \geq L(a_j^*, \theta) \text{ for all } j\} \cap \{\theta : L(a_i^*, \theta) > L(a_j^*, \theta) \text{ for all } j < i\}.$$

It is straightforward to confirm that the value of the problem (15) is

$$\sum_{i=1}^{K} \int_{W_i} L(a_i^*, \theta)\, dF(\theta),$$

so restricting to deterministic languages does not increases losses. ■

Proof of Proposition 1. Let H be a probability distribution on $[0, 1]$ with support $\mathcal{S}(H)$ that is absolutely continuous with respect to Lebesgue measure. Let $a^*(H)$ be an action that minimizes $\int L(a, \theta)\, dH(\theta)$. By assumption $a^*(H)$ exists.

I claim that $M^*(G') = M^*(G)$ for all G' with $\mathcal{S}(G) \subset \mathcal{S}(\mathcal{G}')$. First, I show that $M^*(G') \geq M^*(G)$. Observe that

$$M^*(G') = M^*((1-\delta)G + \delta G') \geq (1-\delta)M^*(G) + \delta M^*(G'),$$

where the first equation follows because G' and $(1-\delta)G + \delta G'$ have the same support and the inequality follows because $a^*((1-\delta)G + \delta G')$ is a feasible action for the minimization problem that defines $M^*(G)$ and $M^*(G')$. Taking the limit as δ approaches zero yields $M^*(G') \geq M^*(G)$. To see that $M^*(G') \leq M^*(G)$, note that

$$M^*(G') = M^*((1-\delta)G + \delta G') \leq (1-\delta) \int L(a^*(G), \theta)\, dG(\theta) + \delta \int L(a^*(G), \theta)\, dG'(\theta). \tag{16}$$

Since $M^*(G) = \int L(a^*(G), \theta)\, dG(\theta)$, taking the limit of (16) as δ approaches zero establishes $M^*(G') \geq M^*(G)$.

Finally observe that the result follows because if G' and G'' are any two distributions, $M^*(G') = M^*(.5G' + .5G'') = M^*(G'')$ because $\mathcal{S}(G'), \mathcal{S}(G'') \subset \mathcal{S}(.5G' + .5G'')$. ■

Proof of Proposition 2. If L is supermodular, then the optimal language much be an interval partition. To see this, observe that any language must induce actions $\{a_1, \ldots, a_K\}$. Order these actions so that $a_i < a_{i+1}$. It follows from supermodularity and (3) that $\{W_i\}$ constitute an interval partition. That is, there exist $0 \leq \theta_1 \leq \theta_2 \leq \cdots \leq \theta_K = 1$ such that $(\theta_{i-1}, \theta_i) \subset W_i \subset [\theta_{i-1}, \theta_i]$ and $L(a_i, \theta_i) = L(a_{i+1}, \theta_i)$ for $i = 1, \ldots, K-1$. ■

Proof of Proposition 3. Any language $\{W_1, \ldots, W_K\}$ induce actions $\{a_1, \ldots, a_K\}$, where a_i minimizes loses given W_i. The sets

$$W_i^* \equiv \{\theta : L(a_i, \theta) \leq L(a_j, \theta) \text{ for all } j\} \cap \{\theta : L(a_i, \theta) < L(a_j, \theta) \text{ for all } j < i\}$$

are convex by assumption and using these words cannot reduce losses. Hence one can always replace an arbitrary language with a language that partitions the state space into convex sets and does not increase cost.

I now show that the assumption that $N(\cdot)$ is monotonic on rays is sufficient for the conclusion. Let $\mathring{X}_i = \{\theta : L(a_i, \theta) < \min_{j<i} L(a_j, \theta)\}$ and $X_i = \{\theta : L(a_i, \theta) = \min_j L(a_j, \theta)\}$. The convexity of X_i and $\mathring{X}_i$ follow from the assumption that $N(\cdot)$ is monotonic on rays. Suppose $\theta', \theta'' \in X_i$ and $\theta^* = \lambda\theta' + (1-\lambda)\theta''$ for $\lambda \in (0, 1)$. It follows that

$$0 \geq L(a_i, \theta) - L(a_j, \theta) \tag{17}$$

for $\theta = \theta'$ and θ''. It follows by assumption that inequality (17) holds for θ^*. This proves that X_i is convex. The same argument (starting with a strict inequality in (17)) establishes that $\mathring{X}_i$ is convex. It follows that W_i^* as defined above is convex and hence defined the language $\mathcal{W}^* = \{W_1^*, \ldots, W_K^*\}$ is a cost-minimizing language consisting of convex words. ■

Proof of Proposition 4. Assume that the density is monotonically decreasing. I will show that the shortest word in the cost minimizing language can be taken to be $[0, l]$ for some l. Iterating the argument demonstrates that the cost minimizing language can be taken to be an interval partition, with the words increasing in length. Let $\mathcal{W} = \{W_1, \ldots W_K\}$ be a cost minimizing language, and let W_1 be the shortest word in $\mathcal{W}$. Denote by l_i the length of W_i. If a subset of W_1 of positive Lebesgue measure is not contained in $[0, l_1]$, then there is a $k > 1$ and a subset $S \subset [0, l_1] \cap W_k$ such that S has positive Lebesgue measure and an S' of identical measure contained in $[l_1, 1] \cup W_1$. Consider an alternative language $\mathcal{W}' = \{W_1', \ldots, W_K'\}$ with word $W_1' = W_1 \cap S \setminus S'$, $W_k' = W_k \cup S' \setminus S$, and $W_i' = W_i$ for $i \neq 1, k$. Denote by $C(\mathcal{W})$ ($C(\mathcal{W}')$) the cost of language $\mathcal{W}$ ($\mathcal{W}'$). $C(\mathcal{W}') = C(\mathcal{W}) - (l_k - l_1)(F(S) - F(S'))$. By assumption, $l_k \geq l_1$. If the density of $F(\cdot)$ is non increasing, then $F(S) \leq F(S')$. Hence $C(\mathcal{W}') \leq C(\mathcal{W})$. ■

Proof of Proposition 5. Assume $f(\cdot)$ is decreasing. Take $t_2 > t_1$ and set $m = (t_1 + t_2)/2$.

$$\begin{aligned}
\int_{t_1}^{t_2} g'(m-\theta)\, dF(\theta) &= \int_{t_1}^{m} g'(m-\theta)\, dF(\theta) + \int_{m}^{t_2} g'(m-\theta)\, dF(\theta) \\
&= \int_0^{m-t_1} g'(s)\, dF(m-s) - \int_0^{m-t_2} g'(r)\, dF(m-r) \\
&= \int_0^{m-t_1} g'(s)\, dF(m-s) + \int_0^{m-t_2} g'(s)\, dF(m+s) \\
&= \int_0^{m-t_1} g'(s)\, \left(dF(m-s) - dF(m+s)\right) \\
&\geq 0. \qquad (18)
\end{aligned}$$

The second line follows from changing variables. The third line uses $g'(s) = g'(-s)$ (and another change of variables). The fourth line follows since $m = (t_1 + t_2)/2$ implies $m - t_1 = t_2 - m$. Finally, the inequality follows because both terms in the integral are nonpositive: $g'(0) = 0$ because the loss function is minimized at 0; $g'(s) > 0$ for $s > 0$ by convexity; and $dF(m-s) \geq dF(m+s)$ by assumption.

It follows that when $f(\cdot)$ is decreasing, the optimal action for an interval is less that the midpoint of the interval.

Suppose that the optimal language consists of words W_i^*, where $W_i^* = [\theta_{i-1}^*, \theta_i^*)$ and $\theta_i^* > \theta_{i-1}^*$. Denote by a_i^* the optimal response to the word W_i^*. Observe that when $L(a, \theta) = g(|a - \theta|)$, a_i^* satisfies $g(\theta_i^* - a_i^*) = g(a_{i+1}^* - \theta_i^*)$ and so

$$\theta_i^* - a_i^* = a_{i+1}^* - \theta_i^*. \qquad (19)$$

It follows that

$$\frac{\theta_i^* - \theta_{i-1}^*}{2} < \theta_i^* - a_i^* = a_{i+1}^* - \theta_i^* < \frac{\theta_{i+1}^* - \theta_i^*}{2} \tag{20}$$

where the inequalities follow because the optimal action for an interval is less than the midpoint of the interval and the equation follows (19). ■

Proof of Proposition 6. Suppose that the optimal language under F includes the word $W_i = [\theta_i, \theta_{i+1}]$ with associated action a_i. Consider the language for $\tilde{F}$ that uses the word $\tilde{W}_i = [\rho\theta_i, \rho\theta_{i+1}]$ and responds to the word with $\tilde{a}_i = \rho a_i$. A change of variables argument implies that

$$\int_{\theta_i}^{\theta_{i+1}} g(a_i - \theta)\, dF(\theta) = \int_{\rho\theta_i}^{\rho\theta_{i+1}} g((\tilde{a}_i - v)/\rho)\, d\tilde{F}(v) > \int_{\rho\theta_i}^{\rho\theta_{i+1}} g(\tilde{a}_i - v)\, d\tilde{F}(v). \tag{21}$$

It follows from (21) that the losses associated with this policy under $\tilde{F}$ are less than the losses under F, the desired result. ■

Proof of Lemma 1.

Note that (12) implies that

$$a_1^C - \theta_1 = \gamma^C(E[\theta_1 \mid m_1] - \theta_1) + (1 - \gamma^C)(E[\theta_2 \mid m_2] - \theta_2) + (1 - \gamma^C)(\theta_2 - \theta_1),$$

$$a_2^C - \theta_2 = (1 - \gamma^C)(E[\theta_1 \mid m_1] - \theta_1) + \gamma^C(E[\theta_2 \mid m_2] - \theta_2) + (1 - \gamma^C)(\theta_1 - \theta_2),$$

and

$$a_1^C - a_2^C = (2\gamma^C - 1)\left((E[\theta_1 \mid m_1] - \theta_1) - (E[\theta_2 \mid m_2] - \theta_2) + (\theta_1 - \theta_2)\right).$$

Routine algebra permits one to write the expected value of any linear combination of losses as a linear combination of six terms: $(E[\theta_1 \mid m_1] - \theta_1)^2$, $(E[\theta_2 \mid m_2] - \theta_2)^2$, $(\theta_1 - \theta_2)^2$, $(E[\theta_1 \mid m_1] - \theta_1)(\theta_2 - \theta_1)$, $(E[\theta_2 \mid m_1] - \theta_2)(\theta_1 - \theta_2)$, and $(E[\theta_1 \mid m_1] - \theta_1)(E[\theta_2 \mid m_2] - \theta_2)$. The lemma follows from the observation that the sixth term is zero (because θ_1 and θ_2 are independent) and elementary manipulations. These manipulations enable one to find coefficients for each of the six terms. The coefficients of the $(E[\theta_1 \mid m_1] - \theta_1)(\theta_2 - \theta_1)$ and $(E[\theta_2 \mid m_1] - \theta_2)(\theta_1 - \theta_2)$ terms are zero. ■