Chicago Journal of International Law

Volume 5 | Number 1 Article 11

6-1-2004

Two-Tier Market Institutions

Avinash Dixit

Two-Tier Market Institutions*

Avinash Dixit**

ABSTRACT

This paper models a hierarchical system for market governance. A monitoring agency detects any opportunistic behavior in each small sub-market or lower tier, using the superior information available at that level. Trade can occur across sub-markets. A small upper-level group of sub-market monitors arranges communication of the news of any cheating in one sub-market to all other sub-markets. I examine when and how such a system can overcome the diminishing returns to information acquisition and communication that have limited the scope and size of self-governing trading communities in the past. I then offer tentative suggestions for governance of globalized markets.

I. INTRODUCTION

Markets are organized forums for voluntary exchanges among traders who may not know each other personally and may not be engaged in repeated bilateral interactions. In most such situations, one or both of the parties to an exchange can behave opportunistically, or, to use a shorter and more evocative word, cheat.[1] A cheater increases his own gain from trade but leaves the other side with a loss. Therefore each trade is a prisoner's dilemma game. Its bad equilibrium may be, for many or even all traders, worse than abstaining from trade altogether. This prospect will keep these traders away; the market will be thin and may even collapse. Therefore, markets can succeed only if effective

* This paper was presented at the conference on *The Empirical and Theoretical Underpinnings of the Law Merchant*, The University of Chicago (Oct 16–17, 2003). I thank Lisa Bernstein and Avner Greif for valuable discussions, my discussant Eric Talley for perceptive comments, and the National Science Foundation for research support.

** John J.F. Sherrerd '52 University Professor of Economics, Princeton University, Princeton, NJ 08544-1021. Phone: 609-258-4013. Fax: 609-258-6419. E-mail: dixitak@princeton.edu. Webpage: <http://www.princeton.edu/~dixitak/home>.

1 Oliver E. Williamson, *The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting* 47–50 (Free Press 1985); Oliver E. Williamson, *Transaction-Cost Economics: The Governance of Contractual Relations*, 22 J L & Econ 233, 234–45 (1979) (discussing and illustrating the concept of opportunism).

mechanisms to deter cheating are put in place. All these mechanisms must have two components: detection of cheating and punishment of the miscreant. Both are problematic.

Consider punishment first. If traders do not have repeated bilateral relationships, then a third party or institution must punish the cheater either by direct penalties such as fines, or by exclusion from future trading opportunities. Unless the agency that adjudicates the matter enjoys coercive power, its authority usually rests on the threat of exclusion from future trades. But the cheater's future trades will be with third parties other than the victim of his cheating. If they expect positive gain from dealing with the cheater, they do not have the incentive to participate in the punishment. In other words, the punishment is a public good, and its execution is another dilemma game. Theoretical models solve this dilemma by postulating punishments for refusing to participate in punishments of the cheater, ad infinitum, but that is a somewhat unsatisfactory solution. Better solutions exist in the real world. Many experiments have demonstrated that people have an instinct to punish anti-social behavior, even at considerable personal cost.[2] Indeed, one can easily think of evolutionary reasons why such instincts arise and persist in societies. Therefore, the execution of punishments may be a less acute problem than might at first appear. A centralized authority to adjudicate and decide to exclude the cheater from future trades may not even be necessary. So long as information about cheating is accurately transmitted and preserved, a purely voluntary system of social norms can execute punishments and, therefore, can serve to deter cheating. If the group of traders is small and closely knit by social and cultural ties beyond the trading relationships, then such norms and shared expectations of collective punishment can arise and persist.[3] A successful collective punishment system also requires that there are few or no false accusations of cheating, whether caused by misunderstanding or malice, or that any false accusations can be detected and corrected at low cost. Greif describes such an error-correction mechanism among Maghribi traders.[4] False accusations should not be a significant problem if traders do not stand to gain from making accusations. In this respect, barring the cheater's access to future trading

---

2 Colin F. Camerer, *Behavioral Game Theory: Experiments in Strategic Interaction* 46–48, 103–04 (Princeton 2003); Ernst Fehr and Simon Gächter, *Cooperation and Punishment in Public Goods Experiments*, 90 Am Econ Rev 980 (2000).

3 Lisa Bernstein, *Private Commercial Law in the Cotton Industry: Creating Cooperation through Rules, Norms, and Institutions*, 99 Mich L Rev 1724, 1749–50 (2001); Avner Greif, *On the Interrelations and Economic Implications of Economic, Social, Political and Normative Factors: Reflections from Two Late Medieval Societies*, in John N. Drobak and John V.C. Nye, eds, *The Frontiers of the New Institutional Economics* 57, 87–88 (Academic 1997); Lisa Bernstein, *Opting Out of the Legal System: Extralegal Contractual Relations in the Diamond Industry*, 21 J Legal Studies 115, 140–41 (1992).

4 Avner Greif, *Contract Enforceability and Economic Institutions in Early Trade: The Maghribi Traders' Coalition*, 83 Am Econ Rev 525, 529–30 (1993).

opportunities, or levying fines that go to third parties, may be better punishments than requiring the cheater to make restitution to the accuser or having the accuser collect any fines.

Information about cheating is the other component of any mechanism to deter cheating, and it is my focus in this article. The information must be acquired, verified, and transmitted to others who will then participate in the punishment of the cheater, and information must be preserved as long as the punishment lasts. Transmission mechanisms may be purely voluntary and decentralized (for example, gossip networks) or may involve various degrees of compulsion and centralization (for example, auditing, and bulletin boards or web sites). All of these have been studied empirically as well as theoretically.

Empirical studies of voluntary information transmission in networks of traders cover a wide variety of locations and times: eleventh century Maghribi traders in northern Africa,[5] cattle owners and herders of the Orma tribe in Kenya in the 1970s,[6] and a community of cattle farmers in northern California in the 1980s.[7] Numerous studies of management of common property resources raise similar issues of communication of information.[8] In that context, cheating is infringing on the group's collective property or on the usufruct rights allocated to other participants. Institutions and organizations run by a central authority to collect and retain information, and also to adjudicate and assess any penalties, show equal variety: law merchants[9] and Genoese traders in medieval Europe,[10] modern trade associations in most countries,[11] and arbitration forums in international trade.[12] These institutions and organizations vary greatly in their scope, methods, and finance, but for my present purpose they have the key common feature that they acquire and retain information about traders' misbehavior. Finally, modern technology has made it possible to run a voluntary

---

5 Avner Greif, *Cultural Beliefs and the Organization of Society: A Historical and Theoretical Reflection on Collectivist and Individualist Societies*, 102 J Pol Econ 912 (1994); Greif, 83 Am Econ Rev at 525 (cited in note 4).

6 Jean Ensminger, *Making a Market: The Institutional Transformation of an African Society* ch 4 (Cambridge 1992).

7 Robert C. Ellickson, *Order without Law: How Neighbors Settle Disputes* ch 1–6 (Harvard 1991).

8 For a general discussion, see Elinor Ostrom, *Governing the Commons: The Evolution of Institutions for Collective Action* (Cambridge 1990).

9 For a general discussion, see Paul R. Milgrom, Douglass C. North, and Barry R. Weingast, *The Role of Institutions in the Revival of Trade: The Law Merchant, Private Judges, and the Champagne Fairs*, 2 Econ & Pol 1 (1990).

10 Greif, 102 J Pol Econ at 917–44 (cited in note 5).

11 For a general discussion, see Bernstein, 99 Mich L Rev at 1724 (cited in note 3); Bernstein, 21 J Legal Studies at 115 (cited in note 3).

12 For a general discussion, see Walter Mattli, *Private Justice in a Global Economy: From Litigation to Arbitration*, 55 Intl Org 919 (2001); Alessandra Casella, *On Market Integration and the Development of Institutions: The Case of International Commercial Arbitration*, 40 Eur Econ Rev 155 (1996).

information transmission system with a central hub, namely a bulletin board or a web site, of which eBay's system of rating buyers and sellers is the best-known example.

Each of these institutions evolved to serve a particular purpose and adapted more or less well to changing circumstances. Therefore, they differ from one another in many respects. But a limitation common to them all is a tendency to diminishing returns, or increasing average and marginal costs, as the scale of the market increases. Most of the case studies emphasize the strains caused by the entry of newcomers and increasing contacts between members of the group and outsiders. Mere increase in the numbers and the heterogeneity of the insiders also causes problems. Information acquisition or monitoring of trade, and transmission of this information, are both subject to increasing costs. Effective monitoring often relies on local information, a point stressed most strongly by Ostrom.[13] Voluntary transmission also becomes more costly as the network becomes larger and less well connected, as Greif's study of the Maghribi traders illustrates.[14] The Maghribis had to disseminate news by writing individual letters to other traders; the disutility costs of this to the writer would increase rapidly with the number of traders. And to be trustworthy, such communication required enough personal relationship between sender and receiver. Therefore, when military and political changes in the Mediterranean created opportunities to expand trade to new areas, the Maghribis had to find other Maghribis to act as their agents. They rarely used Muslims or even non-Maghribi Jews, even though in the absence of opportunism such dealings would have been very profitable. Their governance institution for controlling opportunism required a small well-connected group. Locating other Maghribis in new areas became increasingly difficult; this limited their expansion possibilities.

Modern technology has reduced the cost of communication and largely eliminated the diminishing returns with respect to numbers. However, there remain problems of accurate communication in large groups. Even the much-praised voluntary system of eBay has found it difficult to cope with the increase in the size of that market and has had to resort to more formal governance methods. Its CEO, Meg Whitman, has said: "We all had an intuition that as eBay's community of users became more like the size of New York City than the size of Los Gatos, we would have to deal with issues like fraud."[15]

Greif contrasts the contract enforcement system of the Maghribi traders, which was based on private communication in social networks and norms for multilateral sanctions, with that of the Genoese traders, which had a centralized

---

13 Ostrom, *Governing the Commons* at 14, 17, 177 (cited in note 8).

14 See Greif, *On the Interrelations and Economic Implications* at 65, 67 (cited in note 3).

15 John McMillan, *Reinventing the Bazaar: A Natural History of Markets* 78 (Norton 2002), quoting Meg Whitman, CEO of eBay, as quoted in the San Jose Mercury News 1G (Apr 8, 2001).

system of adjudication of formal contracts on a bilateral basis.[16] He finds that the Genoese system was better able to cope with the expansion of trading opportunities that occurred in the late medieval period around the Mediterranean. Greif models these alternative bilateral and multilateral punishment systems and compares them.[17] I have constructed a theoretical model that shows how diminishing returns arise in a decentralized system and why a centralized system can provide governance at lower cost for large-scale markets.[18] Bowles and Gintis[19] and Kali[20] propose related models that help us understand the informational limitations on the size of a self-enforcing trading group.

However, a hybrid system can take advantage of the best of both worlds. If the market is split up into sub-markets, then each can be monitored locally, using the better information available there. Instances of cheating can then be publicized across all sub-markets using an upper-tier organization that has a much smaller membership, namely one representative from each sub-market. Of course, the split is only for the purpose of information acquisition and exchange; individuals from one sub-market continue to visit other sub-markets and trade with partners there. Bernstein describes just such a system in the diamond industry.[21] The World Federation of Diamond Bourses ("WFDB") is a group of twenty member markets or bourses, of which the New York Diamond Dealers Club is the most important, with about two thousand individual trader members.[22] Each bourse monitors the transactions that take place under its auspices (not necessarily on its physical premises), typically using an arbitration system. Information about cheating is transmitted from one bourse to all via the WFDB: if a trader refuses to pay the judgment awarded against him by the

---

16 Greif, *On the Interrelations and Economic Implications* at 64–88 (cited in note 3); Greif, 102 J Pol Econ at 936–41 (cited in note 5).

17 See Greif, 102 J Pol Econ at 917–22 (cited in note 5); Greif, 83 Am Econ Rev at 531–42 (cited in note 4). Greif also discusses other reasons why the Genoese traders fared better. For example, they had family firms with essentially infinite lives, whereas each Maghribi trader's business was dissolved at his death and his sons had to start afresh. Greif, *On the Interrelations and Economic Implications* at 82–83 (cited in note 3). And the individualist philosophy underlying their social and legal structure was more conducive to innovation and change. Greif, 102 J Pol Econ at 943 (cited in note 5).

18 See Avinash Dixit, *Trade Expansion and Contract Enforcement*, 111 J Pol Econ 1293 (2003).

19 Samuel Bowles and Herbert Gintis, *Optimal Parochialism: The Dynamics of Trust and Exclusion in Networks* 5–21 (2000) (working paper), available online at <http://www.santafe.edu/sfi/publications/Working-Papers/00-03-017.pdf> (visited Mar 28, 2004).

20 Raja Kali, *Social Embeddedness and Economic Governance: A Small World Approach* 8–22 (2003) (working paper), available online at <http://wcob.uark.edu/rkali/smallworld.pdf> (visited Mar 3, 2004).

21 See generally Bernstein, 21 J Legal Studies at 115–57 (cited in note 3).

22 Id at 119, 121.

arbitration tribunal of one bourse, his name and photograph are sent to, and displayed in, the clubroom of every bourse.[23] Bernstein sums this up: "[t]he diamond industry . . . has succeeded, at least for the time being, in creating a system that is designed to capture the benefits of both monitoring by small social groups (individual bourses) and monitoring achieved through information intermediaries (institutions such as the world federation and brokers)."[24] My purpose in this paper is to construct a mathematical model of such a two-tier system, to improve our understanding of when and how it can realize such benefits.

The two-tier institution of market governance has an analogy in industrial organization, namely firm size and structure. Williamson pioneered this literature;[25] a good recent treatment is Qian's.[26] The question is why the whole economy is not organized as one firm. The argument goes as follows. To be sure, there are information and agency problems. But information problems exist even within each of the smaller firms and in the dealings that must occur among firms. Any proposed solution to these problems—whether it be monitoring, incentive schemes, self-selection menus, and so on—can be mimicked by the top-level management of a single firm, so the one-firm organization should fare no worse. The answer to this dilemma that is offered in industrial organization theory is that the mimicking is not feasible. Top-level management has a limited span of control. Therefore, the average and marginal transaction costs of running an organization increase with its size. The literature just cited translates this intuition into mathematical models, and I will do likewise to address the question of whether, when, and how a two-tier organization of market governance can do better than a single-level monitoring of the whole market.

II. A SIMPLE MODEL

Two key ideas underlie the model: traders are different, as if separated by a distance from one another, and the costs of information acquisition and transmission increase with distance. The distance need not be geographic—it can be in any social, cultural, or economic space. A simple formal model where the traders are spread out along a circle suffices to capture the general idea. Let C denote the circumference of the circle, and suppose there are $2D$ traders per unit arc length along the circle. Thus the variable D captures the density of

---

23 Id at 128–29. Bernstein discusses this as a mechanism that can credibly convey a trader's reputation; that is just the obverse of acquiring and conveying accurate information about any cheating.

24 Id at 144.

25 Oliver E. Williamson, *Hierarchical Control and Optimum Firm Size*, 75 J Pol Econ 123 (1967).

26 Yingyi Qian, *Incentives and Loss of Control in an Optimal Hierarchy*, 61 Rev Econ Studies 527 (1994).

traders. Each period, half of the traders stay at their locations. The other half are matched pair-wise with the stationary ones and travel to trade at the partner's location. Thus each period there are D trades per unit arc length along the circle, or CD trades in all. I assume that the matches in successive periods are independently randomly arranged, so there are no repeated bilateral relationships.[27] In reality there are persistent bilateral relationships, and traders do try to continue in such relationships to sustain a simple solution to their two-player prisoner's dilemma based on direct reciprocity. But the need to deal with strangers remains frequent, and market institutions to deter cheating are especially important for these. Therefore, my assumption isolates them for attention.

My purpose in this paper is to compare the costs of alternative organizations or mechanisms for monitoring these trades. Specifically, I study the effects of introducing two tiers or levels at which information is recorded and communicated. To keep matters simple, I assume that the stipulated cost in each case secures perfectly accurate monitoring. Together with the assumption that effective norms for punishment of cheaters are in place, this deters all cheating and sustains honesty in all trades. A tradeoff between cost and the degree of effectiveness of monitoring is an important and interesting question in its own right, but it has less direct bearing on the comparisons between single-level and hierarchical monitoring.

A. Single-Level Monitoring

First suppose that just one central organization monitors the whole market. It can be located at any point of the circle; it will then have to monitor trades at distances ranging from 0 to $C/2$ on each side. One expects the cost of monitoring a trade to increase with the distance between the monitoring organization and the traders. In the context of the New York Diamond Dealers Club,[28] each trader's closeness to the club can represent the frequency of his dealings in the club, or the degree to which he has internalized the culture and the norms of the club. Then one can think of several reasons why monitoring is cheaper for closer members. The club has better local information about them. The consequences of being caught when cheating are more severe for them; therefore, they are less likely to behave opportunistically and will spend less effort or skill to devise clever methods of cheating.

Consider a trade in which the "home" trader is located at distance x from the monitor. If the "visiting" trader comes from distance y, let the cost of

---

27 The probability of matching need not be uniform over the circle; there can be substantial local bias.

28 Bernstein, 21 J Legal Studies at 115 (cited in note 3).

monitoring be written as a function $F(x, y)$, increasing in both its arguments. All the arrangements for monitoring must be in place before y is chosen by the random process described above. Therefore, we are concerned with the expected value of $F(x, y)$ taken over the probability distribution of y. This is a function of x alone, and I assume it to be an increasing function.[29] In the simple model of this section, I assume that it is proportional to x, say φx where φ is a given constant or parameter. This is restrictive in two ways: first, the linear functional form is special, and second, the cost per trade does not depend on the number of trades at that location. These choices bring out the central ideas with the least mathematics. In the next section I will consider extensions of the model to examine how far they generalize.

The parameter φ can be interpreted as an inverse measure of the monitoring technology; the smaller φ is, the more efficient this technology is. Actually the term "technology" should be interpreted in a broad sense, including social or cultural aspects of the group of traders that may make cheating more or less likely. Thus a tight homogeneous community with social contacts that go beyond the interaction in trade may need little explicit monitoring, whereas trades between strangers who interact only occasionally and then disappear may need more monitoring at greater cost. If the monitor has to stand ready to investigate every complaint of cheating but need not investigate every trade, and if in equilibrium only a small amount of cheating goes on because the bulk of the population behave honestly and only a few innately bad traders cheat, then φ can be quite small. I do not model any of this explicitly in this paper, but simply let φ stand for the combination of all these phenomena.

Given the cost function, the total cost of monitoring the whole market is then:

$$\begin{aligned} TC_1 &= 2D\int_0^{C/2}\varphi x\,dx \\ &= 2D\varphi\tfrac{1}{2}(C/2)^2 \\ &= \tfrac{1}{4}\varphi DC^2. \end{aligned} \tag{1}$$

We see that the two aspects of the size of the whole market, namely the circumference of the circle and the density of traders at each point of the circle, affect the total cost differently. The total cost is proportional to the density of traders and to the square of the size of the circle. Thus there are constant returns

---

[29] This will be the case unless x and y are sufficiently strongly negatively correlated, so that an increase in x causes a sufficiently strong leftward shift in the conditional distribution of y, which is unlikely.

if the density increases holding the circumference constant, but diminishing returns with respect to the circumference holding the density constant. The diminishing returns in the circumference dimension (or increasing average and marginal costs) are the analog of the loss of control in the industrial organization literature on firm size. The question I address is whether a two-tier system can counter the diminishing returns.

B. A Two-Tier System

Divide the circle into M distinct sub-markets, each covering a contiguous arc of length L, so that $ML = C$. Figure 1 illustrates this with $M = 6$.

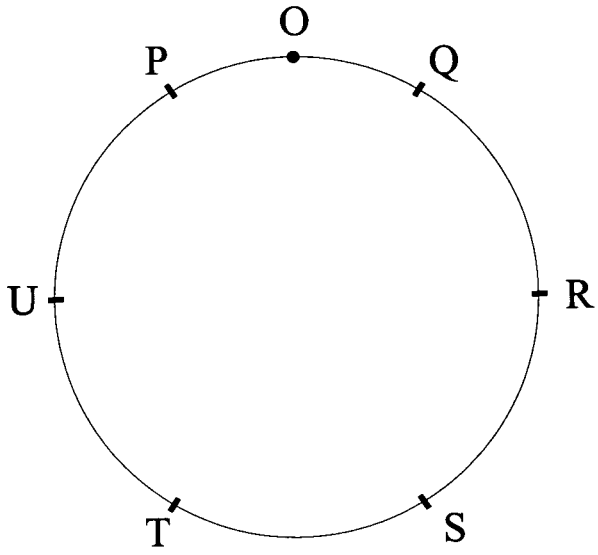

Figure 1: *An Illustration with Six Markets*

For ease of writing and reading, I will henceforth refer to the full circle as "the trading world," and each arc or sub-market simply as "a market." The markets cover the arcs labeled *PQ*, *QR*, *RS*, *ST*, *TU*, and *UP*. This does not affect the matching process. Traveling traders may have to go to markets other than the ones where they are located. Each market has a supervisor who monitors the trades in his own market. The figure shows one of these, namely for the arc *PQ*, located at the midpoint *O* of that arc. Since monitoring costs increase with distance, it is obviously optimal to make the segment covered by

each market a contiguous arc, and to locate the market monitor at the center of that arc.[30] The cost of monitoring each market is then:

$$2D \int_0^{L/2} \varphi x \, dx = 2D \varphi \tfrac{1}{2}(L/2)^2$$

$$= \tfrac{1}{4} \varphi D L^2.$$

In later periods the same traders may be matched with others in different markets. Therefore, information about any cheating must be shared among the supervisors. This happens at the second or top tier and entails its own cost. Each lower-tier monitor sends to the trading-world-level or top-tier supervisor a message containing the names and photographs of any cheaters in his market; the top-tier supervisor collects all these messages and sends back to all market monitors the names and photographs of all miscreants. Even if there is no cheating (as will indeed be the case in equilibrium), a simple "all is well" message must still pass in both directions; otherwise, the recipient cannot be sure whether all is indeed well or there has been a communication failure.

In the illustrative metaphor of the circle, one can think of the top-tier supervisor as located at the center of the circle; his distance from each market monitor is C/π, which is proportional to C. I assume that the costs of sending messages from each lower-tier monitor to the central supervisor and back again are proportional to the distance. This is intended to capture the idea that a message that is being sent across a greater geographic or socio-cultural distance is liable to get more garbled; therefore a higher cost must be incurred in the form of verification or redundancy to ensure accurate transmission.

Each market organization has to monitor LD trades. However, the messages it sends are considerably simpler. In equilibrium, no cheating is going on, so only the "all is well" messages are sent. Of course, to ensure that no cheating is indeed a Nash equilibrium, we have to consider the incentive of any one trader to deviate. In doing so, he will calculate the consequences of his single act of cheating. In particular, he takes into account the fact that his cheating will affect the communication between the two tiers. But the change consists of replacing one simple message ("all is well") by another (his name and his photograph). Then the cost of communication between the markets and the top tier will not depend on the number of trades in each market. Therefore in this section I assume that the communication cost is indeed independent of the number of trades in each market. This assumption is supported by the

---

30 The assumption that each market covers the same arc length L is also optimal in the context of the model where all parts of the circle have the same density of traders and the same information technology. When interpreting the model in the real-world context, one must make an allowance for non-uniformities in different countries or regions.

observation that in many industries that use mechanisms of this kind to police their markets, strategic opportunism is indeed rare. Thus Greif found "only a handful of documents contain[ing] allegations of misconduct."[31] Bernstein reports that in the numerous transactions that occur every year among the 2,000 members of the New York Diamond Dealers Club and the numerous non-members who trade there, only 30 to 40 trades result in a judgment from the arbitration system of the club.[32] While no figures are available for the total number of transactions and the number of cases where the defendant refuses to pay the judgment, a safe guess is that the former is in the hundreds of thousands and the latter in single digits. Therefore, my assumption about the nature of the messages seems justified.[33] However, in the next section I will consider more general cost specifications for completeness.

For now, take the cost of information transmission between the two tiers to be ψC for each market, where ψ is a given constant. This can be interpreted as an inverse measure of the effectiveness of the technology of communication, just as φ was an inverse measure of the technology of monitoring, and I intend it to be interpreted in a similar broad sense.

The total cost of the two-tier system can now be computed:

$$\begin{aligned} TC_2 &= M(\tfrac{1}{4}\varphi DL^2 + \psi C) \\ &= \tfrac{1}{4}\varphi DML^2 + \psi MC \\ &= C(\tfrac{1}{4}\varphi DL + \psi M), \end{aligned}$$

where in getting the last line I have used the fact that $ML = C$.

It remains to choose L and M optimally, namely, to minimize the cost subject to the constraint $LM = C$. I will solve it by treating L and M as continuous variables without restricting M to be an integer. When M is around 20 as in the diamond industry,[34] this is a harmless simplification. Then we have a simple Lagrange problem, which yields the solution:

$$L = 2\varphi^{-\frac{1}{2}}\psi^{\frac{1}{2}}C^{\frac{1}{2}}D^{-\frac{1}{2}}, \qquad (2)$$

---

31 Greif, 83 Am Econ Rev at 528 (cited in note 4) (citation omitted).

32 Bernstein, 21 J Legal Studies at 127 (cited in note 3).

33 The general idea is that if the governance mechanism works well, there is very little strategic opportunism; the rare breaches such as non-payment arise only for reasons of unavoidable financial distress. In private communications, Bernstein confirms this for most industries she has studied. It also conforms with Schelling's theory that threats are "not concerned with the efficient *application* of force but with the *exploitation of potential force* . . ."; an effective deterrent threat never has to be carried out. Thomas C. Schelling, *The Strategy of Conflict* 5 (Harvard 1963).

34 Bernstein, 21 J Legal Studies at 121 (cited in note 3).

$$M = \tfrac{1}{2}\varphi^{1/2}\psi^{-1/2}C^{1/2}D^{1/2}. \quad (3)$$

The resulting minimized cost of the two-tier system is

$$TC_2^{\min} = \varphi^{1/2}\psi^{1/2}D^{1/2}C^{3/2}. \quad (4)$$

We can now compare the cost of running the trading world as one market, given by Equation (1), and that of the optimal two-tier system, given by Equation (4). The most interesting question is how the two systems fare with regard to the two variables that measure two aspects of the size of the trading world—the circumference C and the density D. The cost of the single market is proportional to D and to C^2. The cost of the two-tier system increases less rapidly in both dimensions, being proportional to smaller powers of each, namely the square root of D and the 3/2 power of C. Thus the two-tier system has better returns to scale than the one-tier system—constant returns to the density D change to increasing returns, and diminishing returns to the circumference C are less severe. Thus the two-tier system is likely to be advantageous when either dimension of the size of the trading world is large. A more explicit calculation confirms this: the two-tier system is less costly than the one-tier or unified world market if:

$$4\psi < \varphi DC. \quad (5)$$

The effect of the size variables is as was explained above. The technology parameters enter in obvious ways. If the communication technology is good (low ψ), this is conducive to superiority of the two-tier system; if the monitoring technology is good (low φ), this is conducive to superiority of the unified world market.

Expressions (2) and (3) show how the optimal division of the trading world into distinct markets responds to the two dimensions of size. The job of coping with the circumference C is split equally between the number of markets and the length of arc covered by each market: L and M each vary as the square root of C. The response to density is apparently asymmetric: if D increases, the number of markets goes up as the square root of D, but each market covers an offsettingly smaller distance. However, each market handles LD transactions, so the number of transactions handled in each market goes up as the square root of D, symmetrically with the increase in the number of markets. These specific formulas are again dependent on the specific forms of the cost functions for monitoring and communication, but the qualitative idea of optimally splitting up the responsibility for coping with the volume of trades by the two methods at the two tiers is general.

III. Some Extensions

In this section, I consider two generalizations of the linear and quadratic cost specification of the simple model above. This analysis shows which of the intuitions discussed above are more generally valid, and how some others need to be changed.

A. Power Functions

Replace the gathering information about the D trades at distance x in one market at the lower tier by:

$$\varphi D^{\alpha} x^{\beta}.$$

The simple model had $\alpha = 1$ and $\beta = 1$. The extension allows a parameterization of the returns to scale with respect to both density and circumference in the basic monitoring technology. With respect to density, we have increasing returns to scale if $\alpha < 1$, and diminishing returns to scale if $\alpha > 1$. With respect to the circumference, we have diminishing returns (costs increase proportionately more rapidly than the circumference) if $\beta > 1$, and increasing returns (costs increase less rapidly than the circumference) if $\beta < 1$. Then the total cost of monitoring in any one market is:

$$\begin{aligned} 2D^{\alpha} \int_0^{L/2} \varphi x^{\beta} dx &= 2D^{\alpha} \varphi \frac{1}{1+\beta} \left(\frac{L}{2}\right)^{1+\beta} \\ &= \frac{2^{-\beta}}{1+\beta} \varphi D^{\alpha} L^{1+\beta}. \end{aligned}$$

If the whole trading world is organized into one market, so $L = C$, the cost is:

$$TC_1 = \frac{2^{-\beta}}{1+\beta} \varphi D^{\alpha} C^{1+\beta}. \tag{6}$$

Now consider the two-tier system. Replace the cost of information transmission between the M markets and the top tier by:

$$\psi M^{\gamma} (LD)^{\delta} C^{\theta}.$$

The simple model had $\gamma = 1$, $\delta = 0$, and $\theta = 1$. Thus I now allow parameterized returns to scale with respect to the number of markets and the distance from each market to the top-tier center, and also allow the costs to depend on the number of transactions in each market.

The total cost of the two-tier system is then:

$$TC_2 = M\frac{2^{-\beta}}{1+\beta}\varphi D^{\alpha}L^{1+\beta} + \psi M^{\gamma}(LD)^{\delta}C^{\theta}$$

$$= \frac{2^{-\beta}}{1+\beta}\varphi D^{\alpha}C^{1+\beta}M^{-\beta} + \psi D^{\delta}C^{\theta+\delta}M^{\gamma-\delta}. \quad (7)$$

It remains to choose M and L to minimize this expression subject to $LM = C$. The second line is expressed as a function of M alone, so the problem is even simpler.

If $\gamma < \delta$, both terms on the right hand side of Expression (7) are decreasing functions of M, so the minimum of cost is approached by making M as large as possible. This is a situation where communication between each market and the center becomes rapidly more costly as the volume of trade in each market increases, while the cost of communication increases less slowly with respect to the number of markets. Then it makes sense to have numerous small markets, each handling few transactions to be monitored, and relying almost entirely on communication to the top level.

However, the discussion in the previous section about the rarity of breaches suggests that the cost of communication should not increase very rapidly with an increase in the total number of transactions in each market; that is, δ should be small. Therefore, the rest of the analysis assumes that $\gamma > \delta$. In this case there is an interior optimum. To make the algebra somewhat less messy and focus on the size variables that are of primary interest here, I shall leave out other multiplicative factors. Then we find:

$$L \propto D^{\frac{\delta-\alpha}{\beta+\gamma-\delta}}C^{\frac{\theta+\gamma-1}{\beta+\gamma-\delta}}, \quad (8)$$

$$M \propto D^{\frac{-(\delta-\alpha)}{\beta+\gamma-\delta}}C^{\frac{1+\beta-\delta-\theta}{\beta+\gamma-\delta}}, \quad (9)$$

and the resulting minimized total cost of the optimal two-tier system is:

$$TC_2^{\min} \propto D^{\frac{\alpha(\gamma-\delta)+\beta\delta}{\beta+\gamma-\delta}}C^{\frac{\beta(\theta+\gamma)+(\gamma-\delta)}{\beta+\gamma-\delta}}. \quad (10)$$

Comparing Expressions (6) and (10) gives us the conditions for the two-tier system to have better returns to scale with respect to density and distance. For each, the condition is that the power of that size measure should be smaller in (10) than in (6). For density, the condition is:

$$\frac{\alpha(\gamma-\delta)+\beta\delta}{\beta+\gamma-\delta} < \alpha,$$

which simplifies to $\delta < \alpha$. This formula simply says that the communication cost at the upper tier should have better returns to scale with regard to density than does monitoring cost at the lower tier. I have argued that this is the plausible case. Indeed, in the simple model, I assumed $\alpha = 1$ and $\delta = 0$. Now we see that the same qualitative result remains true for a much broader range of these parameters. Turning to the circumference, the condition is:

$$\frac{\beta(\theta+\gamma)+(\gamma-\delta)}{\beta+\gamma-\delta} < 1+\beta.$$

This simplifies to:

$$\delta+\theta < 1+\beta.$$

To interpret this, note that $\delta + \theta$ is the power of C in the final expression for the upper-tier communication cost in Expression (7). The condition requires this to be less than the power of C in the expression for the cost of running the unified world market in Expression (6). Thus the condition is intuitive. However, it is not so clear whether we should expect this to be likely in practice; that depends on the comparison of the behavior of monitoring and communication costs as distance increases. If modern technology makes communication more accurate, but the detection of cheating remains dependent on local information, then the condition will be more likely to be met. This outcome will tilt the balance of advantage toward the two-tier system.

Incidentally, from Expressions (8) and (9) we see that the same two conditions, $\delta < \alpha$ and $\delta + \theta < 1 + \beta$, together ensure that the length covered by each market and the number of markets respond to changes in the two size variables in the same directions as in the simple model. Of course the powers are no longer ½ each, but more general functions of the underlying parameters.

B. Fixed Costs

Next, suppose there are fixed costs of setting up each market and also the communication center in the two-tier system. To keep the algebra manageable, revert to the specification of the variable parts of the costs that were used in the simple model. Writing Φ for the fixed cost of each market and Ψ for the fixed cost of the communication center, the total cost of the unified world market is:

$$TC_1 = \Phi + \tfrac{1}{4}\varphi DC^2. \tag{11}$$

That of the two-tier system before optimization is:

$$\begin{aligned} TC_2 &= M(\Phi + \tfrac{1}{4}\varphi DL^2) + (\Psi + \psi MC) \\ &= \Psi + [\tfrac{1}{4}\varphi DL + (\Phi + \psi C)M]. \end{aligned}$$

When L and M are chosen optimally, we find:

$$L = 2\varphi^{-½}D^{-½}(\Phi + \psi C)^{½}, \tag{12}$$

$$M = ½\varphi^{½}D^{½}C(\Phi + \psi C)^{-½}. \tag{13}$$

The resulting minimized cost of the two-tier system is:

$$TC_2^{\min} = \Psi + \varphi^{½}D^{½}C(\Phi + \psi C)^{½}. \tag{14}$$

The returns-to-scale properties of this expression with respect to the size variables are similar to those in the simple model; the only essential difference is that a component of the cost of communication between the center and each market, ψC, is replaced by $(\Phi + \psi C)$, by adding the fixed cost of setting up each market. And of course the fixed cost of setting up the communication center is added to the whole. When C is large, the fixed costs make little proportional difference.

IV. SUGGESTIONS FOR FURTHER RESEARCH

The model can be elaborated and extended in several ways. The most important advance will be to specify the technology of monitoring and communication in greater detail instead of capturing the concepts in the form of the cost function. This increased detail will make the model deeper or more structural, which is the current fashion in economics. Imperfect monitoring and garbled communication, and a tradeoff between cost and accuracy, should also be studied. The dynamics of the punishment mechanism can be modeled explicitly instead of assuming, as I have done, that they are effective enough to make actual breach a rare occurrence. But this may make little practical difference.

The model assumes that opportunism (or cheating, or shirking) is clearly defined, and that the only job of the monitor is to determine whether this has taken place. This is standard practice in economics, where cheating is modeled in the context of a prisoner's dilemma game. In this paper, even that game could be kept in the background, because monitoring was assumed to be fully effective, and the focus was on minimizing the costs of monitoring. In practice, however, there is ambiguity as to what constitutes cheating, and this is the focus of many legal scholars. Bringing together the two modes of thinking is an important task for future research.

Greif has studied a different kind of two-tier institution, which he calls the "Community Responsibility System."[35] This institution prevailed in Europe in

---

35 Avner Greif, *Impersonal Exchange without Impartial Law: The Community Responsibility System*, 5 Chi J Intl L 109 (2004).

the late medieval period. In those days, it was easy to observe what city or community a trader belonged to, by observing his dress, speech, choice of inn of residence when traveling, and so on. And each community was small enough to know all its individual members, and to control their behavior through various social and economic punishments. Therefore, the overall system to deter cheating worked by holding a community responsible for the cheating of any of its members, and leaving it to the community to punish individuals and obtain indemnification for the fine paid by the community. It would be interesting to examine the relative merits of the two systems in different circumstances and with different technologies of detection, identification, and punishment.

V. Implications for Governance Design

One must be very cautious in drawing any firm policy implications from a simple exploratory exercise. However, one concept that emerges from the model seems sufficiently robust to allow us to venture a little along this route. Globalization brings together traders from many different countries, with different legal systems of different degrees of formality and effectiveness, different social and cultural norms, different expectations about the behavior of others, and so on. If the trading world is treated as a single market and subjected to one level of governance, the benefits of local information are likely to be lost, because the cost of monitoring is likely to increase rapidly, or to put it another way, the accuracy of monitoring is likely to decrease rapidly. A two-tier structure, where monitoring takes place at a country or regional level and communication uses a top-tier coordinator or intermediary, is likely to fare better. It also has the advantage that it can emerge naturally, in a "bottom-up" manner, by taking the existing systems of social networks and norms in the individual markets and merely constructing a top tier that brings together the local leaders into a group for communication purposes. Perhaps the slogan for successful integration of the world economy should be "monitor locally, communicate (and trade) globally."

However, local monitoring introduces a new problem that was not handled in the model, but may be especially important in an international context. In practice, monitoring of cheating is not straightforward observation; it requires interpretation and adjudication. This ambiguity creates the possibility that in a trade between a local and a foreigner, a local monitor may be biased in favor of his compatriot. Perhaps the problem can be solved simply by requiring a foreign observer to be present during the proceedings and discussions of each country's arbitration panel, but more complex appeals procedures or remedies may also be necessary.

CJIL