Explaining Social Order: Internalization, External Enforcement, or Equilibrium?

Randall Calvert

Working Paper No. 4

March 1995

Explaining Social Order: Internalization, External Enforcement, or Equilibrium?

Randall L. Calvert
University of Rochester

September 1994

Abstract. In the context of a discussion on institutions, "social order" connotes the most fundamental parts of the system of social institutions, focusing attention on the self-enforcing nature of institutions and on the importance of implicit and informal rules of behavior as opposed to formal, written rules. This focus, together with the problematic nature of social order (its rules are sometimes violated, and it may decay or break down), suggests that we look at social order through the lens of game theory; we can fruitfully theorize about social order by seeing it as an equilibrium pattern of behavior in some fixed, underlying game presented by nature. I present an example of such a model, demonstrating how to model diverse opportunities for individual action and communication, and deriving multiple forms of social order that could in principle emerge from a given underlying situation. The model offers general hypotheses about social order and suggests further issues for study. Moreover, the illustration demonstrates the significant differences between this game theoretic approach to social order and other approaches, such as sociological, sociobiological, interpretive, and "socio-historical" methods, as well as alternative rational choice methods.

draft 1.0

This paper was prepared for the Conference "What is Institutionalism Now?" at the University of Maryland, October 14-15, 1994. The research reported here was partly funded by the Center for Advanced Study in the Behavioral Sciences, which generously supported the author as a Fellow during 1990-91, through National Science Foundation grant BNS-9700864.

Explaining Social Order: Internalization, External Enforcement, or Equilibrium?

Randall L. Calvert

1. Introduction: Social Order and Institutions

Like "institutions," "social order" is a term of such central importance to social science that few theorists even pause to define it. "Social order" means, among other things, the long-lived patterns according to which a society functions as a society, rather than as a random agglomeration of individuals. These patterns occur in a great variety of forms; they may constrain individual behavior in subtle and powerful ways; but they are nevertheless subject, at times, to drastic change. The variability, stability, and changeability of social order shows clearly that, in order to understand particular instances of social order, we need a firm grasp of its general nature. Accordingly, social scientists have tried in numerous ways to theorize about social order as a general phenomenon. My goal in this paper is to theorize about social order in a manner well suited to exploring its mysteriously simultaneous variability, stability, and changeability. Before describing several notable approaches to such theorizing, I consider what social scientists mean by social order and how that idea is related to the idea of institutions.

What is Social Order?

Max Weber, although he does not directly define the term, is centrally concerned with social order. To him a "legitimate order," means more than just patterns; it comprises a system of prescriptions for individual action in specific situations, regarded by the members of a society as being obligatory. Adherence to such prescriptions can be either "instrumentally rational" or "value-rational;" and the order could take the form either of "convention," generally enforced by the disapproval of the members of a social group, or of "law," enforced by a "staff" (Weber 1968: 31-34). Weber's work overall is concerned with determining the sources and nature of these felt obligations ("legitimacy"), their power over behavior, and the resulting patterns of behavior in society.

More recently Friedrich A. Hayek and Jon Elster have focused more self-consciously on the meaning of social order itself. Hayek defines order as "a state of affairs in which a multiplicity of elements of various kinds are so related to each other that we may learn from our acquaintance with some spatial or temporal part of the whole to form correct expectations concerning the rest" (Hayek 1973: 36) and applies this concept directly to behavior in society, in order to focus on the question of where such order comes from. Jon Elster gives a related definition of social order as that which "glues societies together and prevents them from disintegrating into chaos and war" (Elster 1989: 1), distinguishing two main aspects of social order, namely predictability and cooperation.[1]

---

[1] In *Law, Legislation, and Liberty*, Hayek (1973) notes that, although the terms "social system," "structure", or "pattern" may sometimes serve for the kind of discussion in which he engages, his real focus is only captured by the term "social order." Student of social order sometimes fail to maintain this distinction; Talcott Parsons, for example, lumps all these considerations into the idea of "social system," and proceeds to examine the mechanisms defining the social system in the same terms and at the same time with the forces that make social order possible. The result is confusing and at times apparently confused; the repeated socialization of individuals into consistent role orientations, for

Approaches to Understanding Social Order

Whether their methodology is that of economics, psychology, or classical sociology, most students of social order have in one way or another attributed individual adherence to the rules of a social order to some kind of *internalization* of those rules. That is, the individual absorbs the prescriptions of social order, and at least to some extent puts them into practice automatically. This is true of the classical sociology tradition, from which I have adopted the term internalization itself. Weber's emphasis on the obligation that individuals feel to obey the dictates of a legitimate order is of precisely this nature. Parsons, similarly, sees the learning of "role orientations" through socialization as one of the fundamental forces that keeps the "social system" in its equilibrium (1951: 204 ff.), and says that role orientations as well as the social sanctions that are sometimes necessary to deter deviance embody "value-orientations" that are "common to the actors in an institutionally integrated interactive system" (251), i.e., individuals under a stable social order.

However, internalization is also central to other theories views of institutions and social order. March and Olsen (1989) and Denzau and North (1994) propose concepts of institutions in which the natural, limited cognitive techniques of human beings lead them to adopt certain ways of acting in socially defined situations. These ways of acting are more than just habits; they constitute a "logic of appropriateness" that March and Olsen distinguish from the consequentialist logic of economics, a logic according to which individuals can reason their way to the right action in keeping with the existing patterns of social order. Political scientists and sociologists of the "socio-historical" school of new institutionalism see institutions as forces that alter individuals' preferences, thus determining what individuals *want* to do (e.g. Hall 1992, Steinmo and Thelen 1992). Some economists have adopted a similar approach, portraying institutional and cultural forces simply as features of individual utility functions: agents in such models care not only about their material well-being, but also about obeying institutional prescriptions.

A few analysts, the "economic" school of new institutionalism in Hall's (1992) terminology, rely not on internalization but on some system of external enforcement to explain social order. For them social order serves as a system of constraints on individuals who might prefer not to be so constrained. This is the approach taken in the study of "structurally induced equilibrium," or SIE (Shepsle and Weingast 1987). The SIE approach portrays social institutions as defining a game, which individual rational actors are stuck with playing. The results of that game are different from what would happen if the same actors with the same preferences interacted instead in a less structured environment. Another example of this approach is the theory of political entrepreneurship, which seeks to explain successful achievement of collective action as the result of the actions of leaders who are endowed with the ability to apply selective incentives (Olson 1965; Frohlich et al. 1971). Such approaches have in common that the constraints of social order or of institutional rules are based on fixed conditions, not changeable by the actions of those subject to the constraints. Social order is thus defined *externally* to the social group in question.

Social Order: Fundamental and Self-Enforcing Social Institutions

---

example, is clearly important in decribing the social system; but Parsons' analysis of the elements of socialization hardly tells us anything about the limits of socialization in counteracting the forces of "cathexis" (gratification of wants), or in coordinating among the system's confusing welter of different role orientations and their demands (Parsons 1951: 5-6, 204-206, 251, and *passim*).

Hayek therefore proceeds to use the term "order," despite its "frequent association with authoritarian views," which he does not intend. The same caveat applies to my use of the term in this paper.

As defined by Hayek, by Weber, and implicitly by Weber, "social order" certainly subsumes, and may be equivalent to, the system of social institutions -- which are, according to North, "the rules of the game in a society, or, more formally, . . . the humanly devised constraints that shape human interaction" (North 1990: 3)[2]. What does it mean, then, when we discuss "social order" in the context of a general discussion on social institutions? Often "social order" connotes the most basic institutions of a society or social group, institutions somehow prior to or more fundamental than higher-level institutions, the latter being predicated upon a base of already-existing culture, norms, constitutional rules, and the like. Taken in this sense, the term serves to focus our attention on two aspects of social institutions that too often escape attention in discussions of institutions. First, not all institutions are purposefully designed and defined (Hayek's main point); there are important social institutions (under North's definition) that emerge without design and that exist only in the common expectations of individuals rather than in any set of rules that are written down (or even readily articulated by the participants). We cannot understand institutions simply by understanding formal, designed, written-down institutions. Second, there are limits to the rules that institutions can successfully impose on a group of individuals -- you can't just specify any old institutional structure and expect it to create social order through processes of internalization or through external enforcement. This is because there is, ultimately, no external enforcement of fundamental social institutions (Calvert 1992), and because human motivations include an ineradicable kernel of self-interest. To understand institutions and institutional change, it is imperative to understand the limits of institutions' ability to dictate "rules of the game." The term "social order" emphasizes the character of institutions as both formal and informal, and as being ultimately a self-enforcing system. These are the features of institutions on which I wish to focus.

In this paper, I suggest an approach that relies neither on internalization nor on fixed or external enforcement to explain the power of social order. My approach emphasizes how social order structures the expectations and incentives of rational individuals, enforcing itself without the necessity of any internalization of "preferences," properly speaking.[3] This approach allows me to focus simultaneously on the factors that limit the ability of social order to constrain individual behavior. Theories based on internalization or on external enforcement, in contrast, place no limits on what rules could in principle be internalized or enforced and thus what patterns of behavior could in principle be present in a social order. My focus on expectations and incentives leads immediately to a method of predicting how certain changes in conditions might lead to changes in social order, and some general prescriptions for the creation and maintenance of particular features of social order.

2. An Approach to Understanding Social Order

Social order is a continuing, relatively stable pattern of behavior, and it may entail selfless-seeming acts by the individual members of a society, as in the "cooperation" aspect of social order that Elster focuses on. Yet within any pattern of social order individuals pursue their individual needs, which may overwhelm the prescriptions of social order. Two general facts reveal this undercurrent of self-seeking

---

[2] The latter form of North's definition is almost precisely that of Durkheim (1895).

[3] I hasten to admit that internalization, at least in the form of habituation, may play a role in social behavior. However, I argue that internalization is an *unnecessary* condition for social order, and that internalized contraints are overcome by temptation often enough to make them also *insufficient* to maintain social order.

behavior. First, social order is problematic: it can break down, either partially as when norms of honesty or of "helping" (Elster 1989: 11-12) become attenuated in urban life, or completely, as when a community dissolves into warring factions or looting rioters. Second, individuals sometimes stretch the boundaries of what is permitted under a given social order -- people invest considerable ingenuity in circumventing social constraints, and others, consequently, in tightening such constraints. Thus even though we usually observe social order as a phenomenon of stability, the important question about it is *how* it can be so stable under most circumstances.

The facts that social order may break down and that there is sometimes individual resistance to its constraints indicate that a rational actor approach will provide a useful model of social order. This approach, as I hope to demonstrate, can account elegantly for the tension between social constraint and occasional individual resistance, without the need of ad hoc assumptions about the circumstances under which internalized norms might fail to work. The fact that social order involves interaction among many individuals, and represents an ongoing stable pattern that is not externally enforced, indicates furthermore that we ought to look to equilibrium models of noncooperative game theory as a highly promising tool for understanding social order. Specifically, then, I want to portray social order as an equilibrium strategy profile in some underlying game that nature presents to the members of a society. This underlying game consists of the problems and opportunities that the individuals face as a group, including opportunties for gain through cooperation or coordination, opportunities for redistribution through concerted action by a subset of the individuals, opportunities for production, exploitation of resources, communication, group decision-making, coercion, and anything else that the individuals could do or avoid doing through the combination of actions that they undertake. In studying social order, given its connotations of foundational social institutions, this underlying game is unalterable: it represents opportunties and risks inherently available to people. How and to what extent the members of a would-be society realize those opportunities or deal with those risks is the subject to be investigated through game-theoretic techniques.

This approach to understanding social order presents two main theoretical and technical problems. These are: how to model the underlying game; and how to deal with the problem of multiple equilibria. If using my approach to social order requires that we represent all of nature in the underlying game, it isn't very useful. However, there is much to be learned about social order by looking at more narrowly defined problems, such as how a specific form or instance of collective action is achieved, how communication is used to reach normative consensus about a new issue, or how a particular asymmetrical distribution of wealth is maintained. Examining issues on this scale requires much less of our model. The model must present the underlying problem or opportunity in sufficient detail, leaving sufficient room for a variety of outcomes (since to ask how social order could work one way is also ask why it doesn't work in some particular other way instead). It must provide for a range of actions that interestingly reflects the true capabilities of real individuals: to cooperate or not, to act in one way or another, to coerce or threaten, and to communicate and react to communication from others. The latter requirement seems especially daunting at first blush, since one normally thinks of communication opportunities as a rather constant, enormous, and formless set of possibilities not amenable to representation in an extensive-form game. If our goal is to examine a particular aspect of social order, however, it can be quite informative to design a fairly simple underlying game having just a few well-defined communication opportunities and a quite rudimentary "language." As long as the included communication makes it possible for players to engage in an interesting variety of counterfactual communication activities -- saying unexpected things, lying, making unanticipated threats, and the like -- the model can tell us quite a lot about the role of equilibrium communicative behavior in maintaining and guiding social order.

Even a game-theoretic model of a small piece of nature is likely to feature a large number of possible equilibria, among which game theory gives us no clue how to choose. This is particularly true if the model allows for a very interesting variety of communicative possibilities or allows for repeated

interactions among the players.[4] This feature is often seen as a flaw in the applicability of game theory to the study of social interaction, since it is impossible in most interesting games to predict, from initial conditions, what the behavioral outcome of the game will be (see for example Hechter 1992). I believe instead that the multiplicity of equilibria is a basic and real feature of human interaction, a deep truth about social life, accurately represented by game theoretic models. But what can we do analytically with a game having many equilibria? First, the analyst of social order (of institutions in general) is often concerned with understanding a particular pattern of social order; in that case, the proper theoretical exercise is to provide and interrogate a model that reasonably represents the underlying possibilities, *together with* an equilibrium to that game that reflects the interesting properties of the observed social order, in order to produce an explanation of how the observed patterns are maintained and make predictions about how they might change. Second, the multiplicity of equilibria represents an important coordination problem inherent in the design of institutions or the emergence of social order: many different patterns of expectations and behavior among the members of society would be consistent and self-enforcing, and before one such pattern is established there is considerable room for trial-and-error experimentation (Crawford and Haller 1990) or strategic machination (Schelling 1960) before one of them is realized. This process is of considerable interest in understanding the genesis, and even the maintenance, of social order.

In the next section I present one simple example of the kind of model that is useful in examining the possibilities of social order. It has several features that illustrate general ideas about social order that can be further refined and tested through empirical and theoretical work. It also serves to illustrate what can be learned from equilibrium models of social order that cannot be learned from some other common ways of looking at social order. After presenting the model and several of its equilibria, I turn to a discussion of those features.

3. A Model of Interaction in a Social Group

Any analysis of social order, then, begins with the specification of the underlying situation in which social interactions take place -- the actions that individuals may take, and the payoffs that result from each combination of actions. "Social order" consists of any ongoing pattern that emerges, according to which individuals choose their actions. In order to illustrate the analysis of social order by this method, and to draw some general conclusions about the nature of social order, I specify a simple setting in which all relevant interaction consists of prisoner's dilemma situations played by pairs of individuals who are repeatedly mixed and matched at random, together with opportunities for communication among the players.[5] Let $\{1,2,\ldots,N\}$ be the set of individuals ($N \geq 2$ and even). These players are repeatedly and randomly paired up to play the following PD game, G, with one another:

---

[4] On repeated interaction, see Fudenberg and Maskin (1986); on the multiplicity of communication equilibria, see Myerson (1985).

[5] The model used here is identical to that analyzed in Calvert (1992), where the concern was to illustrate the modeling of institutions and to compare the properties of formal institutions with informal norms or decentralized arrangements.

		Player 2 C	Player 2 D
Player 1 C		1 1	α -β
Player 1 D		-β α	0 0

where $\beta > 0$, $\alpha > 1$, and $\alpha - \beta < 2$. Players discount payoffs in future periods by some discount factor $\delta < 1$. In this game, "C" signifies the action of "cooperating, and "D" that of "defecting." Each player has complete, perfect information about her own past and present interactions, but no information about interactions between pairs of other players.

In order to portray much of the richness of social order, however, it is necessary to introduce additional features to this game. These could include variations in the players' payoffs[6] and opportunities for the players to communicate in the course of play. For illustrative purposes I concentrate here only on the latter. So consider a communication-augmented stage game G* that proceeds in the following steps for each iteration $t = 1, 2, \ldots$:

(1) Players are paired at random; each player i is paired with each other player j with probability 1/(N-1), and the pairings in different iterations are independent events. Each player knows only the pairing that she herself is in.

(2) Players may communicate with one another. Specifically, any player i may send a message to any other player j, and messages may be sent to as many other players as desired. For each player j whom she contacts in this manner, player i bears a communication cost c. All these communications take place simultaneously. The messages must be chosen from some message set or "language" L, to be specified below.

(3) A player who received a message in step 2 may reply costlessly with another message from L.

(4) The paired players play one iteration of G; each player learns only the outcome of play in her own pair..

(5) The players may again communicate as in step 2, simultaneously sending messages chosen from L to as many other players as desired, bearing a cost c for each contact. A player *cannot* verify to her opponent in the current phase what messages she has sent to other players.

Denote by $G^*(\delta)$ the repeated game thus defined. I choose this particular specification of communication opportunities, and will specify below a particular form for the message set L, to facilitate simple illustration of some interesting equilibria; those strategies would still be equilibria in any game that includes all the

[6] As analyzed, for example, in Calvert (1993) for a two-player repeated PD game with communication and private information about contribution cost.

steps of G* plus any additional communication steps or larger message sets.[7] In order to examine some of the possibilities for social order in this group, I enumerate several possible equilibria for G*(δ).

Equilibrium 0: Unconditional Defection

One equilibrium is of course for all players always to defect unconditionally in step 4. This would be social order only in the almost vacuous sense that each individual is accurately "predicting" the actions of others, and taking the appropriate action in anticipation. The fact of interaction in a social group, however, goes completely unexploited by this pattern of behavior.

Equilibrium 1: Pairwise Reciprocal Cooperation

There are other equilibria, of course, in which cooperation occurs. For example, suppose that each player uses a subgame-perfect version of the standard Tit-for-Tat strategy[8] against each other player, regarding each series of pairwise interactions with each of the N-1 partners as though they were N-1 separate games. I shall refer to this strategy, in which no player ever communicates and any communication from other players is ignored, as "TFT". From a player's standpoint, each of the N-1 separate interactions has an *effective* discount factor γ that is smaller than the value δ used to discount payoffs one iteration in the future in the overall game G*(δ), due to the rarity with which a given player is re-encountered. The value of γ can be specified in terms of δ as follows:

Lemma. In G*(δ), the effective discount factor for interactions with a particular other opponent is given by $\gamma = \frac{\delta}{N - 1 - \delta(N-2)}$.

(The proof of the Lemma, and of all subsequent technical results, can be found in the Appendix.) According to standard results on the repeated prisoner's dilemma, if γ is sufficiently large then this pairwise Tit-for-Tat strategy, played by both players in any pair, constitutes an equilibrium strategy profile for the pairwise game between those two players. However, note that γ is a strictly (and in fact rather rapidly) increasing function of N, so that increasing group size quickly dilutes the "shadow of the future" (Axelrod 1981) necessary for cooperation.

If γ does meet the necessary condition, then the strategy profile in which all N players use TFT against each opponent constitutes an equilibrium in the overall game, G*(δ). In that case, we have what would commonly be called a "norm of reciprocity" (Axelrod 1981; Calvert 1989): each player expects his or her opponent to reciprocate cooperation and retaliate against defection, making defection not worthwhile. Such a pattern of behavior and expectations is a more substantial example of social order than was Equilibrium 0, being predicated not solely on individual maximizing behavior but also on the fact of

---

[7] This is accomplished trivially, by extending the equilibrium strategies of G*(δ) to prescribe that the additional communication opportunities are unused and that any use of them by other players is ignored. Of course, the addition of further communication opportunities also makes possible new equilibria that were not available in G*(δ).

[8] That is, for robustness we add to the usual Tit-for-Tat strategy the prescription that if for any reason a player *does* depart from the strategy and defect, on the next two encounters with the same partner that player cooperates unconditionally, making "restitution" so to speak, and thereafter resumes conditional cooperative play. This version of Tit-for-Tat was, to my knowledge, first suggested by Sugden (1986).

repeated interaction of the same individuals with one another. The following result specifies the condition for pairwise TFT to be an equilibrium of G*(δ) in terms of δ:

Theorem 1. The strategy profile in which each player plays TFT with each partner individually is a subgame perfect equilibrium of G*(δ) provided that δ is greater than both $\frac{(N-1)(\alpha-1)}{(N-2)(\alpha-1)+\beta+1}$ and $\frac{(N-1)\beta}{(N-1)\beta+1}$.

Equilibrium 2: Decentralized Communication and Cooperation

If N is moderately large, even a healthy δ may be insufficient to support this simple cooperative form of social order due to the rarity of repeated meetings between any two players. In principle a more effective form of punishment could be used, in which defection against one partner would be punished by subsequent partners; this would yield a much more "social" form of social order, in which each member of the group takes an interest in all the behavior by members in group interactions. However, due to the lack of information about one's partner's behavior with other partners, no strategy of simple, "silent," groupwise reciprocation is possible. To achieve cooperation in a large group, members must take advantage of their opportunities to communicate, in order to transmit information about who may have cheated whom.

For present purposes, I assume that the language L includes all possible pairs (n,*a*) where n∈{1,2, . . .,N} and *a*∈{C,D}; for this equilibrium, the message (n,*a*) may be taken to mean, "My partner in this stage was n, and she took action *a*." Now define the strategy of "Tit-for-Tat with multilateral communication", TFT/MC, in which a player behaves as follows in the corresponding steps of each iteration of G*:

(2) Send no messages before play.

(3) Make no replies.

(4) If in *cooperation status* (defined below), play C if partner is reported to be in cooperation status and D if partner is reported in *punishment status* (also defined below); if in punishment status, play C.

(5) If in cooperation status, truthfully report partner's identity n and action *a* to each of the N-2 other players, incurring a cost of (N-2)c; if in punishment status, do not communicate.

Any communication in steps 2 or 3 and any communication other than the prescribed reports in step 5 are ignored. A player begins G*(δ) in cooperation status. A player enters (or remains in) punishment status if she fails to play as prescribed in step 4, or if she fails to report when prescribed in step 5. A player in punishment status returns to cooperation status immediately upon playing C in step 4. Note that a player may deviate from this strategy in step 5 by falsely reporting her partner's action; if that happens, then a player in cooperation status may be treated by other players as though she were in punishment status. As the Note on the Proof of Theorem 2 (see appendix) indicates, such lying does not occur in equilibrium.

Just as in Theorem 1, it is possible to specify a condition on δ so that this strategy is in equilibrium:

Theorem 2. The strategy profile in which all players use the strategy TFT/MC is a subgame perfect equilibrium of G*(δ) for sufficiently large δ provided that c is less than each of

the following values: $\frac{1}{N-2}$, $\frac{2-\alpha+\beta}{N-2}$, and $\frac{\delta}{N-1}(1+\beta)$. The exact lower bound for δ to support this equilibrium is $\frac{(\alpha-1)+(N-2)c}{\beta+1}$.

The lower bound on δ will generally be much lower using TFT/MC than it was under the more anomic social order of TFT. For example, if $\alpha=2$, $\beta=1$, and $N=100$, then the lower bound on δ for pairwise reciprocity to be possible (from Theorem 1) is .99. If $c=.005$, however, the lower bound on δ for cooperation under TFT/MC is only .745. Notice, though, that the cost of communication has to be relatively small to make it worthwhile for each individual to contact everybody on every iteration; otherwise the process of communication bleeds away the gains from cooperation.[9]

The social order described in Equilibrium 2 is certainly more "social" than that in Equilibrium 1; it involves a real norm in the sense of a shared attitude that defecting against a partner in cooperative status is bad, and that failing to report is bad. Still, this form of social order does not involve any sort of formal organization; every player has exactly the same role in communicating and in punishing deviation. In Weber's (1968) terminology, this would be a "conventional" rather than a "legal" order. Nevertheless, notice how the "rules of the game" for these players are much more complex than in Equilibrium 1's pairwise TFT; beyond simply interacting with each partner, a player is expected by the whole group to act in a certain way and to communicate certain information at certain times to certain other people. Any failure to act or communicate as expected incurs punishment.

A key point to notice about this form of analysis of social order is that these "rules of behavior" are not, in my analysis, part of the assumed structure of the underlying game. Rather, they are features of the equilibrium strategy profile. If discounting is too heavy or communication cost too high, such a system of rules is impossible to maintain. If such conditions change change unexpectedly during the play of a game, they may even force a change in the rules in the form of a leap from one equilibrium pattern of behavior to another. This is a useful feature for a model of social order: using no ad hoc assumptions about changes in the social order, it provides a possible explanation of the nature of such change.

Equilibrium 3: Communication through a Centralizing Institution

In the form of social order in $G^*(\delta)$ described by Equilibrium 2, the requirement of multilateral communication requires each individual to bear a communication cost of (N-2)c on each iteration. In a large group with significant cost of communication, the maintenance of a cooperative social order would require a more efficient exchange of information. As we will see, such an efficiency gain can be had, but only by achieving a greater degree of organizational complexity.

---

[9] One might reasonably wonder at this point why the players need be required to report on every iteration -- why not have them report only when their partners defect? On the equilibrium path, this would seem to eliminate the communication costs altogether. However, the reporting itself is in the nature of a public good: in a large group, I am unlikely to meet my cheating partner again for a long time, so my squealing on her has a deterrence value that benefits everybody else but may not be worth the cost to me. Since players cannot observe cooperation, defection, or even membership in interactions to which they are not party, there is no way to monitor whether other players are reporting cheaters unless they are simply required to report at every iteration.

In place of the multilateral communication used in TFT/MC, I consider next a scheme of centralized communication.[10] Arbitrarily designate player 1 as the "director", who will serve as a central clearinghouse of information. Assume now that the language L includes at least the elements (n,a) as before, plus the messages Q_j for each $j \in N$, which for purposes of this equilibrium may be interpreted as the query, "My opponent is j; what is his status?"[11] Each player i in {2,3,...,N} follows a strategy of "Tit-for-Tat with centralized communication" (TFT/CC), described as follows. When paired with player 1, always defect and never send any messages (the idea being that, for simplicity, the director refrains from actual play of the PD game). When paired with any other player j in {2,...,N}, a player $i \neq 1$ observes the following prescription for the respective phases of each stage in the game:

(2) In iteration $t = 1$, do not communicate. In iterations $t > 1$, if in *punishment status* (defined anew, below), do not communicate; if in *cooperation status* (also defined anew), then pay c to send message Q_j to the director, where j is i's current partner.

(3) Make no replies.

(4) In iteration $t = 1$, play C. In iterations $t > 1$, when in cooperation status and told (by the director's reply in step 3) that j is in cooperation status, play C. Otherwise play D. When in punishment status, play C ("make restitution").

(5) If in cooperation status, pay c to report j's action, (j, a), to the director. If in punishment status (i.e., just played D inappropriately in step 4) then do nothing.

Player 1, the director, obeys strategy "A," described as follows. For all t, in step 2, make no statement; in step 4, always defect. In step 5, make no communication. Otherwise:

(3) In iteration $t = 1$, make no replies. In iterations $t > 1$, if message Q_j was received from player i in step 2, reply by truthfully reporting the status of that player's reported opponent -- (j,C) if cooperation, (j,D) if punishment -- and otherwise communicate nothing.

The cooperation and punishment statuses are defined as follows. At iteration $t = 1$, every player is in cooperation status. A player i in cooperation status enters punishment status if any of the following occur: in step 2, she fails to query as required; in step 4, she fails to cooperate even though her partner had been reported by the director to be in cooperation status; or if she fails to report as prescribed in step 5. If a player in punishment status cooperates in step 4, she re-enters cooperation status beginning with step 5.

The following theorem derives the conditions under which the profile TFT/CC can be used by all players, with the director using A, in equilibrium:

---

[10] The equilibrium constructed in this section resembles the "Law Merchant System Strategy" (LMSS) central to Milgrom et al (1990), in its use of the pre-trade (pre-PD) query and the central communicator. In the LMSS, a player whose partner deviates from cooperative play registers a complaint with an outside "judge," who assesses a fine against the deviant and reports the deviant to be a non-cooperator until the fine is paid. The present model dispenses with such fines, instead simply making the "director" the reporter of whether a player is supposed to be punished, TFT-style, in the play of the PD. Further, the present analysis considers the cost of communication, while that of Milgrom et al. does not.

[11] The reason for including the queries is so that the director can learn who is playing whom (and thereby identify anyone who fails to report in step 5) without herself expending any search costs.

Theorem 3. The strategy profile in which player 1 uses A and players 2 through N use TFT/CC is a subgame perfect equilibrium of $G^*(\delta)$ for sufficiently large δ provided that $c < 1/2$. The lower bound on δ is the maximum of $\dfrac{\beta + c}{\beta + c + (1-2c)\dfrac{N-2}{N-1}}$ and $\dfrac{\alpha - 1 + c}{\alpha - 1 + c + (1-2c)\dfrac{N-2}{N-1}}$.

When N and c are large, centralized communication makes cooperation possible under conditions in which the informal social order in Equilibrium 2, using decentralized communication, could not be maintained. To use the same illustrative values as before, when $\alpha=2$, $\beta=1$, $N=100$, and $c=.005$, then the lower bound on δ for Equilibrium 3 is only .506, compared with .745 for Equilibrium 2 and .99 for Equilibrium 1.[12]

In the TFT/CC equilibrium, we have a form of social order that actually involves a bit of formal organization. Despite the fact that the underlying game is the same in all these equilibria, Equilibrium 3 alone defines a special role for one player, forcing the other players to report only to the director and to base their actions on only the director's replies. Moreover, social order under Equilibrium 3 bestows unique prerogatives upon the director, namely the receiving of queries and the sending of messages concerning the status of players. This apparent addition of "new strategies" to the game, however, is akin to the apparent introduction of new rules discussed after Theorem 2: the underlying game really remains constant, and only the equilibrium has new features. A player in {2, 3, . . ., N} could irrationally take it upon herself to make statements about partners' statuses, departing from TFT/CC, but other players will react only to the director. The TFT/CC institution presents the players with rules that they must, out of their own self-interest, follow. Still, a change in the game's parameters, such as a lowering of the discount factor relative to α and β, could render those rules ineffective and unenforceable.[13]

---

[12] Indeed, in Equilibrium 3 larger group size is technically an *advantage*, if only because higher numbers mean that a player is paired with the director (and thus gets a zero payoff) less often. Of course, if larger group size put more strain on the director's ability to cope with her duties, a factor unmodeled here, this advantage might disappear.

[13] An important feature of Equilibrium 3 is that the director has no incentive of any kind; player 1's adherence to strategy A is rational only because she can gain nothing (although she would also lose nothing) by deviating, given the other players' equilibrium strategies. Further elaboration of the model clarifies the instability of this approach, but also shows how the director can be supplied with positive incentives to maintain the cooperative equilibrium.

With a slightly richer strategy space, an alternative equilibrium exists that, as indicated by real-world experience, might be expected to predominate over Equilibrium 3: the director could commit extortion, threatening to falsely report players in punishment status if they do not share their winnings; or players could bribe the director to falsely report them in cooperation status while they cheat their partners.

Calvert (1992) examines directly several such possibilities. By including opportunities for direct ("voluntary") transfers of money in the game, opportunities for extortion are created; these can then be deterred by having players pay the director a small fee when querying about their current partners. The threat of losing this fee (along with the prospect, for other players, of always being extorted in the future once having submitted to extortion) can then, under the right combination of expectations and rational intentions, deter the director from extortion. A similar strategy ought to be available to limit bribery, although I have not examined this possibility in detail.

Comparing Equilibrium 3 with the previous ones demonstrates another desirable feature of understanding social order as an equilibrium in some underlying game. It demonstrates that social order through formal organization is of the same basic nature as social order through informal organization or norms. A more complicated formal organizational structure could be portrayed just as well, in principle, as an equilibrium in an underlying, constant game; and indeed, once we realize that social order must ultimately be self-enforcing, such a view becomes necessary to a full understanding of organizations.[14] A formal organization might be the result of an unwritten but shared understanding, such as the governing system of a preliterate society; or many of its details might be specified in a written document, such as the U.S. Constitution. Although the written rules might influence players' initial expectations about the selection of an equilibrium, however, they take no precedence over unwritten understandings as hallmarks of formal organization (any more than informal understandings are more fundamental to social order than are written rules). In either case, the participants must find it in their interest, given their expectations about the actions of others, to obey the written or unwritten rules, for those rules to have any effect.

The Underlying Game, Social Order as Equilibrium, and the Many Possible Social Orders

The model examined in this section concentrates on one version of Elster's "cooperation" aspect of social order, namely a setting in which repeatedly- and randomly-paired individuals have the opportunity to realize gains through pairwise cooperation. It specifies an "underlying game" reflecting the individuals' (fixed) opportunities for gains from social action, their (fixed) problems in sharing information, and a range of action possibilities (cooperation, communication, retaliation) that gives the model's toy version of social order some interesting internal structure while making the maintenance of any pattern of social order problematic under adverse circumstances. Under favorable conditions, then, a self-enforcing social order is possible. That order can take many forms, only a few of which are derived here. It can use a variety of methods of "social control" ranging from purely pairwise reciprocity to a system relying on general punishment with centrally coordinated monitoring. The social order can be completely decentralized, assigning the same roles of cooperating, reporting, and punishing to all members of the society; or it can be organized, assigning specialized roles to some individuals. Most importantly, even though it is self-enforcing, changing conditions can make new, more productive forms of social order possible, and can sometimes force a change in the pattern of social order. The next section examines some of these dynamic possibilities, after drawing some general lessons about social order, and compares the features of my approach with alternative approaches to social order.

4. Conclusions about Social Order, and about Theorizing on Social Order

The example in Section 3 illustrates important ideas having broader applicability to our general understanding of social order. In the model, as in most instances of social order, repeated interaction plays a central part (Taylor 1976); as a result, discounting, which depends upon participants' uncertainties about the future, is important in setting the limits on what kinds of social order are possible. And since cooperation is an important aspect of social order generally (Elster 1989), the particular conditions for equilibrium in my example illustration have considerable general application: the maintenance of cooperation depends upon a particular relationship holding between the discounting of future payoffs and,

---

[14] This idea fits closely with the message of Kreps (1990), who argues that the nature of "corporate culture" within firms is critical in explaining the nature of the firm, and can be understood in terms of equilibrium among individual rational actors.

to use Rapoport and Chammah's (1965) terms, the reward from mutual cooperation, the temptation to defect unilaterally, the "punishment" of mutual defection, and the "sucker's payoff" of unilateral cooperation. Finally, my example demonstrates how important communication can be in understanding patterns of social order,[15] and how organization and something that looks very much like *authority*, namely the players' adherence to the director's pronouncements, can be critical in achieving mutual gains for the members of a group.

Social Order Can Be Both Problematic and Constraining

A puzzling aspect of social order is how its prescriptions can be so utterly constraining in some circumstances, sometimes pervading the life choices of nearly all individuals for decades or centuries, while in other circumstances it can become problematic, violated by individual deviation, and subject to large-scale alteration and even breakdown. A main weakness of classical sociology, in which the existence and effectiveness of institutions and social order is taken as the analytical starting point (as in Durkheim 1895, Weber 1968, or Parsons 1951) is that it offers little systematic guidance about this tension between constraint and fragility of social order or institutions. Even though Weber and Parsons identify, in their typologies of social forces, tendencies toward both conformity and deviance, their theories offer scant guidance about when one tendency will dominate the other. They attribute conformity to two forces: the internalization of social norms or values, so that an individual simply wants to behave in a way conducive to the existing social order; and social control, which is exercised over an actual or potential deviant by other people.

When we look at social order as a rational-choice equilibrium phenomenon (thus concentrating on the "social contol" aspect of conformity), however, its status as simultaneously constraining and problematic follows quite simply and directly. People do face "rules of the game" in social interaction; these rules are defined by the shared expectations that they hold about how people will behave in relevant situations, including situations in which one has deviated from the expectations. Such rules are binding as long as the same people play the same game with the same payoffs, risks, and uncertainties about the future (and as long as there is no concerted group action to realize an alternative equilibrium through some communicative process, a topic to which I return below). If the stakes of the game change, discounting becomes heavier, or communication becomes more difficult, then those expectations may at some point fail to bind. The precise point at which this occurs is derivable in the theoretical model, and such a derivation yields comparative-statics-type predictions about social order that can be tested in the real world.

If social order worked through "internalization," there would remain the problem of explaining why and when social order might fail to constrain individual behavior. Parsons, for example, posits forces of "cathexis" (want gratification) and social control (internalized and external) that are in tension; but nothing in the theory tells how these forces will balance off, and which will predominate in particular circumstances. The social systems theory cannot generate predictions about failures of social order due to constraints failing to constrain. Similarly, why should individuals' cognitive maps, once arranged, as March and Olsen (1989) and Denzau and North (1994) suggest, to produce a given pattern of social order or institutional constraint, ever change so as to cause weakening or breakdown in social order? Nothing

---

[15] In this model, communication really serves two functions that are not clearly separated: it serves to *coordinate* the actions of players so that all can be agreed on what behavior is appropriate and what circumstances call for punishment; and it serves to share *information* that inititially is asymetrically distributed. Banks and Calvert (1992) presents a model of coordination in a battle-of-the-sexes game with incomplete information about payoffs, in which the two roles of communication are much more clearly delineated.

in the theory points toward any prediction of such changes. Likewise, since there is no accepted theory of preference-change, the economic and socio-historical analyses invoking such change have no mechanism for explaining when pro-social preference changes take place and why they might change back to selfish preferences. A similar problem affects "structure-induced equilibrium" models of rational choice within institutions (Shepsle and Weingast 1987; Bates 1983). Such models are good for deriving behavior while institutional constraints hold, and for showing why one institutional structure might be preferred by individuals to another, but useless for explaining how contraints hold.

Other Aspects of Social Order

Since the model in Section 3 was simplified to include only opportunities for cooperation, communication, and, implicitly, punishment, there are naturally many interesting features of social order, even of Elster's (1989) "cooperation" type of social order, that it fails to capture. These fall into two important categories: types of cooperation not portrayed in the model; and interesting real-world behavior that does not occur in the model. Elster treats several types of cooperation separately: externalities, helping, conventions, joint ventures, and private ordering (Elster 1989: 11-15). To treat different types of cooperation generally requires the use of different basic games (taking the place of G or G* in Section 3), but a very similar type of analysis can then be conducted. For example, Taylor's (1976) original analysis of social order without the state examines simple cooperative equilibria in a repeated, many-player collective action problem (rather than a multiplicity of two-player PDs as I have done) in order to draw conclusions about the possibilty of order under anarchy.

Even within the model of Section 3, there are types of behavior that should be of interest in the study of social order that do not occur in the equilibria derived here. Most obviously, when players adhere to the cooperative equilibria in this model, nobody ever fails to cooperate; there is no social deviance, because the equilibrium strategy profile perfectly deters deviance. In social order in the real world, deviance is always present to some degree. But we can study such behavior by using a more complex model in which player preferences vary (either across players or over time), so that interesting equilibria exist in which deviance does occur. For example, Calvert (1993) examines a PD model in which each player has private information about her true preferences, which vary from one iteration to the next.[16] Ideal efficiency in that model requires that each player be allowed to defect with impunity whenever her contribution to group utility is sufficiently outweighed by her private cost from cooperating. Interesting equilibria, then, generally involve some defection by players, which, to counteract the moral hazard generated by the asymmetric information, must sometimes be punished.

Distribution and Social Order

Another form of behavior that does not appear in the analysis of Section 3 is the striving of players for advantage in the *distribution* of gains, which, Knight (1992) emphasizes, is basic to the politics of institutional creation. A simple distributional issue arises in the model of Section 3 if, as suggested, the director is to receive a payment for her efforts. The size of that payment compared to the other players' remaining gains from the cooperation she facilitates would be an issue in any discussion over institutional arrangments; if the payment were large, this discussion might take the form of a contest over who gets to serve as director. One could use the same model to address distribution among the cooperating players by considering equilibria in which some players are required to cooperate less than others. For example,

---

[16] Specifically, in each iteration, each player draws a new value for the private, immediate "cost of cooperating," which in the case of game G would be given by α-1 or β (depending on whether the other player cooperates or defects).

suppose we designate players M, M+1, . . . , N as "privileged." Consider a variant on the strategies in Equilibrium 2 in which, when a privileged player is paired with a non-privileged player on an even-numbered iteration, cooperation is required only of the non-privileged player. That is, only the non-privileged player would enter punishment status for failing to cooperate. This arrangement would yield an asymmetrical distribution of the gains from cooperation.[17] With sufficiently light discounting (high δ), the threat of retaliation, in the form of lost payoffs from privileged players' cooperation on odd-numbered iterations, would still be sufficient to force non-privileged players to acquiesce in this unfavorable arrangement.

An analysis similar to that performed in Section 3 would show how changes in the parameters, including M, would affect the viability of such an equilibrium. Perhaps more interesting, however, is the question of how such an asymmetrical equilibrium might arise in the first place. I turn next to this question in a more general form. I emphasize here, however, simply that nothing in the general approach that I have described necessarily favors analytical attention to efficiency considerations over distributional ones.

The Dynamics of Social Order

Using standard game theory techniques, I have said little about the dynamics of social order. Although equilibrium analysis shows what kinds of change in social can occur, and on occasion indicates conditions under which change must occur, it says nothing directly about the process of that change and sometimes little about the end result. Recent developments in game theory and its applications offer a number of promising approaches to examining dynamics, however.

The basic form of the dynamics problem is this: How is an equilibrium established in a game with multiple equilibria? Hardin (1982: chapters 10-14) examines informally a number of processes for the establishment of a convention, an equilibrium in a coordination game. More recently game theorists have created a new literature on "learning in games" in which myopic adjustment models show how players might arrive together at a Nash equilibrium (e.g. Jordan 1991; Samuelson 1991; Crawford 1991). For the special case of pure coordination games, Crawford and Haller (1990) have even been able to construct a full-rationality, strategic model of players trying to arrive at a coordinated outcome without the benefit of any pre-established "focal point." All these models show means by which a particular pattern of social order could emerge without any sort of centralized design.

The alternative game theoretic approach to the problem of establishing equilibrium is to portray a pre-play stage of the game in which players propose or negotiate various institutional possibilities. By including a communication stage in which proposals can be made, and dictating some sort of determinate end to that stage, it is possible to derive equilibria in the overall game (negotiation plus play of the original game) in which, if an equilibrium proposal was made and "agreed to," the players will then find it in their interest to adhere to the agreement. Noncooperative bargaining models (Rubinstein 1982) provide the basic toolkit for modeling and solving such communication games. This approach is especially well suited to the study of patterns of social order that are constructed by conscious, collective design, such as formal political institutions.

Any method that addresses the original establishment of equilibrium in a game can also be used to address the problem of changing equilibrium, and thus of changes in social order. Changes in objective conditions, such as the infusion of new players, technological change, or increased uncertainty about future dealings, are represented by changes in the basic payoff and discounting parameters of the game. If an

---

[17] If it is permitted to make interpersonal comparisons of utility, there is also a deadweight loss of some of the available total gains. Of course, this is not an efficiency loss, strictly speaking, since some player would be gaining at the expense of others.

unexpected change in a game's parameters should invalidate the equilibrium in use, the players are essentially in the position of establishing a new equilibrium just as at the beginning of play.[18] If a parameter change makes new equilibria possible, no change is forced, but players who can communicate may be able to take advantage of new opportunities for gains by switching to a new equilibrium, again by the same processes used to arrive at an original equilibrium.

Suppose for example that increased uncertainty about future re-encounters with members of the group caused an unexpected reduction in the discount factor. In any of the equilibria of the model presented in Section 3, then, at some point cooperation could no longer be motivated by the existing expectations about actions and reactions. The prescribed forms of communication and cooperation would cease, and some out-of-equilibrium process could be expected to lead to an alternative equilibrium, very likely unconditional defection. In case of a failure of one of the more patience-demanding equilibria (say Equilibrium 1), however, it is also possible for players to realize some alternative, more robust equilibrium by making more effective use of communication. Such a change could even, in principle, be arrived at spontaneously: for example, if players began for the first time to communicate multilaterally about their experiences with past partners and to react to those reports, Equilibrium 2 might come to replace Equilibrium 1. With sufficiently light discounting, Equilibrium 1 is Pareto-superior to Equilibrium 2, since the former does not require the bearing of communication costs; but once the discount factor dips below the threshold given in Theorem 1, communication becomes advantageous.

By including opportunities for proposing such changes as additional communication stages in the game, these mechanisms for planned change could be examined using the same equilibrium-analysis tools as were used in Section 3. Alternatively, myopic adjustment or strategic coordination processes, just like those used to model the initial establishment of equilibrium, could be applied to study unplanned or decentralized shifts in patterns of social order. Note that these processes of communication and adjustment need not just work even-handedly to produce equilibria in which all players have symmetrical roles, or just to realize Pareto-improvements. They can just well provide the vehicles for a subset of players to exploit other players through the establishment of a favorable, asymmtrical equilibrium that redistributes gains, even at some cost in terms of efficiency. Thus the tools for studying game-theoretic dynamics are important for understanding the redistributive maneuvers whose importance Knight (1992) emphasizes.

Conclusion

The rational actor approach is often criticized for producing a picture of human behavior as "undersocialized," ignoring the social influences on behavior. Sociological and other approaches avoidn this difficulty by making socialization, the internalization of social constraints, the centerpiece of theories about social order. Those approaches generally suffer because they do not account well for change in patterns of social order. Many means have been suggested for retaining desirable features of the rational

---

[18] If parameter changes are *anticipated*, on the other hand, the most direct game theoretic approach is to solve for equilibria that specify behavior under all contingencies; such an equilibrium would not change at all, although a different part of its description would come into effect. However, the processes described above for spontaneous emergence of equilibrium need not yield such fully-specified, parameter-contingent equilibria. As long as the players agree in their expectations about situations that really happen (namely, play under some original parameter values), they need not come into agreement about appropriate behavior under other situations. When the parameters change and the unresolved situation occurs, the players once again face the problem of establishing equilibrium -- just as if the change had been unanticipated. So the right analytical approach often is to portray players as having no expectations about parameter changes.

choice model while taking direct account of social factors: making moral or institutional factors part of the individual's utility function (Koford and Miller (1991), "metapreferences" (Sen 1978), grafting nonconsequentialist norms onto rational actors (Elster 1989), and portraying social factors as fixed constraints on rational action (such as in the SIE approach of Shepsle and Weingast). These models too, however, treat social constraint as a fixed influence, depending either on internalization or on external determination of the constraints.

My primary argument in this paper attacks the basic premise for criticizing our models' rational actors as undersocialized. Rather, I argue that, if the context of social interaction is seen as a game presented by nature to a group of rational actors, then equilibrium behavior by those actors represents precisely the patterns of mutual expectations and intentions that constitute social order. I add no new assumptions to the rational actor approach; I only suggest a new modeling strategy, a different application of the usual rational-actor building blocks. This social order-as-equilibrium approach draws on ideas that have already seen productive application in the literature, notably in the work of Taylor (1976), Hardin (1982), Milgrom et al. (1990), and Knight (1992). Far from neglecting the influence of social interaction in individual behavior, this approach is in fact nicely, and at present probably uniquely, suited to the development of a theory of social order.

Appendix
Proofs of Formal Results

Proof of Lemma. The proof uses induction to calculate the expected discounted present value (DPV) of interactions with a given partner, ignoring the current period; it shows that this is equal to the sum over $t = 1$ to infinity of $\gamma^t y$, where y is the payoff in each period and γ is as defined in the Lemma. Thus γ is the effective discount factor as required. For notational convenience, let $q = 1/(N-1)$, the probability of meeting a given partner on a given turn.

First, calculate the expected DPV of the payoff from the next single interaction with the given player. This will be the sum from periods 1 through infinity following the present period of the probability of having the next interaction with the partner in that period, times the discounted payoff if that happens:

$$\sum_{t=1}^{\infty} (1-q)^{t-1} q \delta^t y = y \frac{\delta q}{1 - \delta(1-q)} = \gamma y ,$$

where γ is as defined in the statement of the Lemma.

Now, consider the expected DPV of the payoff from the $(T+1)$-th encounter with this partner after the current period, assuming that the payoff from the T-th encounter is $\gamma^T y$. The probability of the first encounter taking place t periods after the present is $(1-q)^{t-1}q$; once that happens in period t, the expected DPV (discounting from the present period, $t = 0$) of the T-th encounter thereafter is $\delta^t \gamma^T y$. Summing over t gives the expected DPV of the $(T+1)$-th encounter from the present:

$$\sum_{t=1}^{\infty} (1-q)^{t-1} q \delta^t \gamma^T y = \gamma^T y \, \delta q \sum_{t=1}^{\infty} (1-q)^{t-1} \delta^{t-1} = \gamma^T y \frac{\delta q}{1-\delta(1-q)} = \gamma^{T+1} y .$$

Thus the expected DPV of all future interaction with the given partner is

$$\sum_{t=1}^{\infty} \gamma^t y = \sum_{t=1}^{\infty} \left[\frac{\delta q}{1 - \delta(1-q)} \right]^t y ,$$

and substituting 1/(N-1) for q gives the desired result. □

Method for proofs of theorems. All the proofs below proceed by demonstrating that there is no profitable one-period deviation from the equilibrium strategy; this includes deviations made *after* leaving the equilibrium path, since the equilibrium strategy specifies what is to be done in those situations as well. To see why this is sufficient, consider the following argument. When examining the incentives of a given player, the strategies assumed for all the other players, along with the structure of the game, specify a dynamic programming problem for the given player. This is an infinite-horizon problem with discounting, so if there were some infinitely long sequence of departures from the specified strategy that made the player better off, there would also be a *finite* sequence of departures that improved payoff (since beyond some distant future point all further gains are minuscule once discounted back to the present). But suppose there is no single-period departure that alone makes the given player better off; then no longer-duration, finite-length sequence of departures will do so either, for in the next-to-last period of the sequence of departures, to depart for one more period cannot be profitable. Thus in order to show that there can be no profitable departure of any duration from the specified strategy by a single player, it suffices to show that there is no profitable one-period departure.[19] Again, however, note that it is important to show that

[19] This is the strategy of proof adopted by Milgrom *et al.*, which they describe as an appeal to the optimality principle of dynamic programming (1990, p. 8).

even if a player has departed from the strategy, it must then be optimal to carry out the appropriate off-equilibrium behavior specified by the strategy.

Proof of Theorem 1. Theorem 1 follows directly from the Lemma along with standard results on the repeated Prisoner's Dilemma;[20] it is proved here for completeness, as well as to give a simple illustration of the general technique of proof to be used subsequently. The standard proof is given in terms of γ, and the final result is obtained by application of the Lemma.

Suppose that both players cooperated on the previous iteration (or that the game is in iteration 1). In such a situation, the DPV of all present and future payoffs to a player from adhering to TFT by cooperating on the current iteration, and then playing according to TFT thereafter (for a payoff of 1 in every period), can be written as $1 + \gamma + \gamma^2/(1-\gamma)$. The payoff from the one-period deviation of defecting now and adhering to TFT in the future (including the consequent "restitution") is $\alpha - \gamma\beta + \gamma^2/(1-\gamma)$. For cooperation to be the equilibrium move, then, requires $1 + \gamma \geq \alpha - \gamma\beta$, or, $\gamma \geq (\alpha-1)/(\beta+1)$. Using the Lemma to substitute for γ gives the first bound stated in Theorem 1.

Suppose that one player conformed to the equilibrium strategy on the previous iteration, but that the other deviated. TFT then calls for the conforming player to play D on the current iteration, which for any γ obviously gives a higher payoff than playing C since that D will not be punished subsequently. So consider the decision of the deviant. To return to the equilibrium path by playing C yields a payoff of $-\beta + \gamma + \gamma^2/(1-\gamma)$. The one-time deviation of playing D now and then returning to TFT (that is, beginning "restitution" on the next iteration instead of the present one) pays $0 - \gamma\beta + \gamma^2/(1-\gamma)$. Thus for TFT to be in equilibrium requires $-\beta + \gamma \geq -\gamma\beta$, or $\gamma \geq \beta/(1+\beta)$. Applying the Lemma gives the second bound stated in Theorem 1.

These cases represent all the possible situations in which a player could consider a one-period deviation from TFT; in each case, the one-period deviation would be unprofitable. Therefore the strategy profile in which both players use TFT is an equilibrium. □

Proof of Theorem 2. Steps 2 and 3 require only that the players not engage in costly communication that will in any case be ignored by other players using the assigned strategy. It remains only to show that there are no profitable one-period deviations beginning in steps 4 or 5.

In step 4, if player i is in punishment status, playing C as prescribed by TFT/MC yields an expected payoff (in discounted present value from that point on) of

$$[-\beta - (N-2)c] + \delta[1-(N-2)c] + \frac{\delta^2}{1-\delta}[1-(N-2)c]$$

while playing D and then returning to TFT/MC would yield a payoff of

$$0 + \delta[-\beta-(N-2)c] + \frac{\delta^2}{1-\delta}[1-(N-2)c].$$

Thus in order for TFT/MC to be in equilibrium we must have the latter no larger than the former, or

$$-\beta - (N-2)c + \delta[1-(N-2)c] \geq \delta[-\beta-(N-2)c].$$

Since $c < 1/(N-2)$ and $\delta \leq 1$, this is always satisfied.

Consider next a player in cooperation status in step 4. If the player's partner is reported to be in punishment status, then obviously it is optimal for the player to defect as prescribed. If the partner is

---

[20] Such as Taylor 1976 or Axelrod 1981. To see the standard result in the present notation, see Milgrom et al. 1990.

reported in cooperation status, playing C gives a payoff of [1-(N-2)c]/(1-δ), which can be written as [1-(N-2)c] + δ [1-(N-2)c] + δ^2/(1-δ) [1-(N-2)c], while deviating to D gives

$$\alpha - \delta[\beta + (N-2)c] + \frac{\delta^2}{1-\delta}[1-(N-2)c].$$

The resulting necessary condition for equilibrium is then

$$[1 - (N-2)c](1-\delta) \geq \alpha - \delta[\beta + (N-2)c]$$

which reduces to

$$\delta \geq \frac{\alpha - 1 + (N-2)c}{\beta + 1}.$$

Since we assumed $c < (2-\alpha+\beta)/(N-2)$, this can be satisfied by sufficiently large $\delta < 1$.

Finally, turn to step 5. Obviously a player in punishment status will be content not to report, as prescribed. For a player in cooperation status, reporting yields a payoff of

$$-(N-2)c + \frac{\delta}{1-\delta}[1-(N-2)c],$$

while failing to report at all puts the player into punishment status and yields a payoff of

$$0 + \frac{-\delta\beta - \delta(N-2)c}{1-\delta}.$$

The resulting equilibrium condition is

$$-\delta\beta - \delta(N-2)c \leq -(1-\delta)(N-2)c + \delta[1-(N-2)c],$$

which reduces to

$$\delta \geq \frac{(n-2)c}{(N-2)c + \beta + 1}.$$

Thus again, sufficiently large $\delta < 1$ satisfies the condition. Notice that this right-hand side is smaller than that derived for cooperation status in step 4, so the previous condition subsumes this one.

It remains to show that it is optimal in step 5 to report to all of the remaining N-2 players, rather than to just some of them. Let K be the number of players to whom player i reports in step 5, $0 \leq K \leq N-2$. Let $V_{ab}(K)$ represent the value of optimal continuation when player i's status is $a \in \{C,D\}$, i's current partner's status is $b \in \{C,D\}$, i reports to K other players in step 5 of the current iteration and to N-2 others thereafter, and all other players always report to all N-2 other players in step 5. Finally, let U(a,b) represent the payoff in step 4 of the current iteration to a player i of status a whose partner's status is b. Then in the next iteration, the probability that player i meets one of the players to whom she reported is (K+1)/(N-1) (the K to whom she reported plus her current partner), while her probability of meeting a player to whom she did not report, thus having to make restitution, is (N-K-2)/(N-1). We can then write player i's total expected payoff beginning in the current iteration as

$$V_{ab}(K) = U(a,b) - Kc - \delta\left[\frac{K+1}{N-1}V_{CC}(N-2) + \frac{N-K-2}{N-1}V_{DC}(N-2)\right].$$

The derivative of $V_{ab}(K)$ with respect to K is

$$-c + \frac{\delta}{N-1}[V_{CC}(N-2) - V_{DC}(N-2)].$$

Clearly $V_{CC}(K) \geq V_{DC}(K)$ for all K since in the former i's partner begins by cooperating, while in the latter the partner begins by defecting. Hence as long as c is sufficiently small, the optimal K is as large as possible, that is, K = N-2. The bracketed term in the derivative reduces to $1+\beta$, giving the third bound on c in the statement of the Theorem. □

Note on the Proof of Theorem 2. There is no temptation for player i to report falsely that i's partner has played D in step 5 since, because the partner will not know that i lied, the partner will assume that she is in cooperation status in the future and will not make restitution. If a player's report could be made known to that player's partner, a temptation to lie would be present -- subgame perfection would require a player falsely reported to be in punishment status to behave as though really in punishment status, so lying would pay off if the liar met the same partner on the very next iteration. Thus if the report were assumed to be known to the player's partner, the equilibrium behavior of Theorem 2 could not be maintained without a modification of the equilibrium strategy to deter lying. This modification could be accomplished by the addition of a kind of Tit-for-Tat in truthful reporting, that is, a separate punishment scheme in which a player whose partner lies retaliates by lying on the next meeting between the same two players in which both are in cooperation status. This would add a second condition on the discount factor, but the new condition would be less stringent than that presently given in Theorem 1 since, although the retaliation is heavily discounted, the reward from lying is itself discounted, accruing as it does only with probability 1/(N-1). Details of this proof are available from the author on request. A similar modification would apply to Theorem 3.

Proof of Theorem 3. The director has no incentive to violate any of the strategy's prescriptions, so consider the incentives of players 2 through N. In step 2, obviously no such player in punishment status will wish to query. For a player in cooperation status, querying gives a payoff of

$$1 - 2c + \frac{\delta}{1-\delta}(1-2c)\frac{N-2}{N-1},$$

while failing to query yields

$$-\beta - c - \frac{\delta}{1-\delta}(1-2c)\frac{N-2}{N-1}.$$

The resulting condition for equilibrium is simply $1-2c \geq -\beta-c$, which is true since $0 < c < 1/2$ and $\beta > 0$.

In step 4 when the player is in punishment status, cooperating as prescribed gives a payoff of

$$-\beta - c + \frac{\delta}{1-\delta}(1-2c)\frac{N-2}{N-1}$$

while defecting yields 0 in the present period and a present value of X beginning with the next period, where

$$X = \frac{N-2}{N-1}\left[-\beta - c + \frac{\delta}{1-\delta}(1-2c)\frac{N-2}{N-1}\right] + \frac{1}{N-1}\delta X,$$

for a total payoff of δX for defecting. (The complicated structure here is due to the fact that a deviant cannot make "restitution" until she is paired with a player other than the director; this occurs in any given iteration with probability (N-2)/(N-1).) Solving for X,

$$X = \frac{N-2}{N-1-\delta}\left[-\beta - c + \frac{\delta}{1-\delta}(1-2c)\frac{N-2}{N-1}\right],$$

so the relevant equilibrium condition is

$$-\beta - c + \frac{\delta}{1-\delta}(1-2c)\frac{N-2}{N-1} \geq \delta\frac{N-2}{N-1-\delta}\left[-\beta - c + \frac{\delta}{1-\delta}(1-2c)\frac{N-2}{N-1}\right],$$

which is true if and only if

$$-\beta - c + \frac{\delta}{1-\delta}(1-2c)\frac{N-2}{N-1} \geq 0,$$

that is,

$$\delta \geq \frac{\beta + c}{\beta + c + (1-2c)\frac{N-2}{N-1}}.$$

The latter is always possible for sufficiently large $\delta < 1$ since $c < 1/2$.

If a player is in cooperation status in step 4 and his partner is in punishment status, obviously there is no reason not to play D as prescribed. If both the player and the partner are in cooperation status, the payoff to playing C is

$$1 - c + \frac{\delta}{1-\delta}(1-2c)\frac{N-2}{N-1},$$

while the payoff from defecting is $\alpha + \delta X$, where X is as defined above. Moving α to the left-hand side, the condition for equilibrium in cooperation status in step 4 becomes

$$\alpha + c - 1 + \frac{\delta}{1-\delta}(1-2c)\frac{N-2}{N-1} \geq \delta\frac{N-2}{N-1-\delta}\left[-\beta - c + \frac{\delta}{1-\delta}(1-2c)\frac{N-2}{N-1}\right].$$

This is identical to the condition derived for punishment status in step 4 above, except that α-1 replaces β on the left-hand side. The resulting condition on δ is thus

$$\delta \geq \frac{\alpha - 1 + c}{\alpha - 1 + c + (1-2c)\frac{N-2}{N-1}}.$$

This is true for sufficiently large $\delta < 1$ provided that the right-hand side is less than 1, which it is since $c < 1/2$ and $\alpha > 1$.

In step 5, a player in punishment status obviously will not wish to report. A player in cooperation status gains $-c + (1-2c)\delta(N-2)/[(1-\delta)(N-1)]$ by reporting, and $0 + \delta X$ by failing to report, where X is as defined above. The resulting equilibrium condition is the same as that derived above for punishment status in step 4, except that $-\beta$ is removed from the left-hand side. Thus the condition above is sufficient to make the cooperation-status player report in step 5 as well. Finally, a argument similar to that explained in the Note on the Proof of Theorem 2 shows
that a player has no incentive to report falsely in step 5. □

References

Axelrod, Robert. 1981. The emergence of cooperation among egoists. *American Political Science Review* 75: 306-18.

Banks, Jeffrey S., and Randall L. Calvert. 1992. "A Battle-of-the-Sexes Game with Incomplete Information." *Games and Economic Behavior* 4, pp. 347-72.

Bates, Robert H. 1983. "The Preservation of Order in Stateless Societies: A Reinterpretation of Evans-Pritchard's *The Nuer*." In Bates, *Essays on the Political Economy of Rural Africa*. Cambridge UK: Cambridge University Press.

Calvert, Randall L. 1992. "Rational Actors, Equilibrium, and Social Institutions." Forthcoming in J. Knight and I. Sened, eds., 1994, *Explaining Social Institutions*. Ann Arbor: University of Michigan Press.

Calvert, Randall L. 1993. "Communication in Institutions: Efficiency in a Repeated Prisoner's Dilemma with Hidden Information". In W. Barnett and N. Schofield, eds., *Political Economy*. Cambridge: Cambridge University Press.

Calvert, Randall L. 1989. "Reciprocity among Self-Interested Actors: Uncertainty, Asymmetry, and Distribution." In Peter C. Ordeshook, ed., *Models of Strategic Choice in Politics*, University of Michigan Press.

Crawford, Vincent P., and Hans Haller. 1990. "Learning how to cooperate: Optimal play in repeated coordination games." *Econometrica* 58: 571-95.

Crawford, Vincent. 1991. "An 'Evolutionary' Interpretation of Van Huyck, Battalio, and Beil's Experimental Results on Coordination," *Games and Economic Behavior* 3: 25-59.

Denzau, Arthur T., and Douglass C. North. 1994. "Shared Mental Models: Ideologies and Institutions." *Kyklos* 47: 3-31.

Durkheim, Emile. 1938 (1895). *The Rules of the Sociological Method*, 8th ed. Edited by George E.G. Catlin. Translated by Sarah A. Solovay and John H. Mueller. Chicago: University of Chicago Press.

Elster, Jon. 1989. *The Cement of Society: A Study of Social Order*. Cambridge: Cambridge University Press.

Frohlich, Norman, Joe A. Oppenheimer, and Oran R. Young. 1971. *Political Leadership and Collective Goods*. Princeton, N.J.: Princeton University Press.

Fudenberg, Drew, and Eric Maskin. 1986. "The Folk Theorem in Repeated Games with Discounting and with Incomplete Information." *Econometrica* 54: 533-54.

Hall, Peter. 1992. "The Movement from Keynesianism to Monetarism: Institutional Analysis and British Economic Policy in the 1970s." In Sven Steinmo, Kathleen Thelen, and Frank Longstreth, eds., *Structuring Politics: Institutional Analysis in Comparative Politics*. Cambridge UK: Cambridge University Press.

Hardin, Russell. 1982. *Collective Action*. Baltimore: Johns Hopkins University Press.

Hayek, Friedrich A. 1973. *Law, Legislation and Liberty, Volume I: Rules and Order*. Chicago: University of Chicago Press.

Hechter, Michael. 1992. "The Insufficiency of Game Theory for the Resolution of Real-World Collective Action Problems." *Rationality and Society* 4: 33-40.

Jordan, Jerry S. 1991. "Bayesian Learning in Normal Form Games." *Games and Economic Behavior* 3: 60-81.

Knight, Jack. 1992. *Institutions and Social Conflict*. Cambridge, U.K.: Cambridge University Press.

Koford, Kenneth B., and Jeffrey B. Miller, eds. 1991. *Social Norms and Economic Institutions*. Ann Arbor: University of Michigan Press.

Kreps, "Corporate Culture and Economic Theory", in Alt and Shepsle, eds., *Perspectives on Political Economy* (1990).

March, James G., and Johan P. Olsen. 1989. *Rediscovering Institutions: The Organizational Basis of Politics*. New York: Free Press.

Milgrom, Paul R., Douglass C. North, and Barry R. Weingast. 1990. The role of institutions in the revival of trade: the Law Merchant, private directors, and the Champagne fairs. *Economics and Politics* 2: 1-23.

Myerson, R. 1985. "Bayesian Equilibrium and Incentive Compatibility: An Introduction. In L. Hurwicz et. al., eds., *Social Goals and Social Organization: Essays in Memory of Elisha Pazner*. Cambridge: Cambridge University Press.

North, Douglass C. 1990. *Institutions, Institutional Change, and Economic Performance*. Cambridge: Cambridge University Press.

Olson, Mancur S. 1965. *The Logic of Collective Action*. Cambridge, Mass.: Harvard University Press.

Parsons, Talcott. 1951. *The Social System*. New York: The Free Press.

Rapoport, Anatol, and Albert M. Chammah. 1965. *Prisoner's Dilemma*. Ann Arbor: University of Michigan Press.

Rubinstein, Ariel. 1982. "Perfect Equilibrium in a Bargaining Model." *Econometrica* 50, pp. 97-109.

Samuelson, Larry. 1991. "Limit Evolutionarily Stable Strategies in Two-Player, Normal Form Games." *Games and Economic Behavior* 3, pp. 110-28.

Schelling, Thomas C. 1960. *The Strategy of Conflict*. Canbridge, Mass.: Harvard University Press.

Sen, Amartya K. 1978. "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory." In H. Harris, ed., *Scientific Models and Men*. London: Oxford University Press.

Shepsle, Kenneth A., and Barry R. Weingast. 1987. The institutional foundations of committee power. *American Political Science Review* 81: 85-104.

Steinmo, Sven, and Kathleen Thelen. 1992. "Historical Institutionalism in Comparative Analysis." In Sven Steinmo, Kathleen Thelen, and Frank Longstreth, eds., *Structuring Politics: Institutional Analysis in Comparative Politics*. Cambridge UK: Cambridge University Press.

Sugden, Robert. 1986. *The Economics of Rights, Co-operation, and Welfare*. Oxford, U.K.: Basil Blackwell.

Taylor, Michael. 1976. *Anarchy and Cooperation*. London: John Wiley.

Weber, Max. 1968. *Economy and Society: An outline of Interpretive Sociology*. Ed. by Guenther Roth and Claus Wittich. New York: Bedminster Press.

W. ALLEN WALLIS INSTITUTE OF POLITICAL ECONOMY
1994-95 Working Paper Series

University of Rochester
107 Harkness Hall
Rochester, NY 14627

WP# 1 "What Constitutions Promote Capital Accumulation? A Political-Economy Approach,"
Krusell, Per and Ríos-Rull, José-Victor, December 1994.

WP# 2 "Rationalizing School Spending: Efficiency, Externalities, and Equity, and their Connection to Rising Costs,"
Hanushek, Eric A., January 1995.

WP# 3 "Interpreting Recent Research on Schooling in Developing Countries,"
Hanushek, Eric A., January 1995.

WP# 4 "Explaining Social Order: Internalization, External Enforcement, or Equilibrium?"
Calvert, Randall, March 1995.

W. Allen Wallis Institute for Political Economy

Working Paper Series

To order copies of the working papers, complete the attached invoice and return to:

Mrs. Terry Fisher
W. Allen Wallis Institute of Political Economy
107 Harkness Hall
University of Rochester
Rochester, NY 14627.

Three (3) papers per year will be provided free of charge as requested below. Each additional paper will require a $5.00 service fee which must be enclosed with your order.

An invoice is provided below in order that you may request payment from your institution as necessary. Please make your check payable to the W. Allen Wallis Institute of Political Economy.

OFFICIAL INVOICE

Requestor's Name:

_____

Requestor's Address:

_____

_____

_____

_____

_____

Please send me the following papers free of charge:
(Limit: 3 free per year)

WP# _____ WP# _____ WP# _____

I understand there is a $5.00 fee for each additional paper.
Enclosed is my check or money order in the amount of $ _____.
Please send me the following papers.

WP# _____ WP# _____ WP# _____

WP# _____ WP# _____ WP# _____

WP# _____ WP# _____ WP# _____

WP# _____ WP# _____ WP# _____