Norms, Third-Party Sanctions, and Cooperation

JONATHAN BENDOR
Graduate School of Business
Stanford University

DILIP MOOKHERJEE
Indian Statistical Institute
New Delhi

1. INTRODUCTION

In recent years, a large body of literature has explored the relative roles of bilateral reciprocity and centralized authority in enforcing cooperative behavior. These are, of course, the two most prominent institutional alternatives discussed since the time of Hobbes. More recently, scholars such as Axelrod (1986) and Ellickson (1987) have drawn attention to a third alternative: social norms.

Norms are a pervasive feature of a variety of social and economic groups: examples include traditional village communities (Netting), race and caste ties (Srinivas; Akerlof), political cliques (Schmidt et al.), professional networks of academics, lawyers, or physicians (Freidson), work groups, departments within firms, or firms (Roethlisberger and Dickson; Dalton; Lorenz), bureaucracies (Heclo), bodies of elected politicians such as the U.S. Congress (Matthews; Fenno), relations among whalers (Ellickson, 1989), and trading in the absence of enforceable contracts, such as between merchants in pre-industrial societies (Belshaw) or between nations (Oye). These informal norms provide another powerful influence on individual behavior and are, therefore, important in the analysis of social phenomena such as cooperation, stratification, and discrimination.

Norms—unwritten rules such as "live up to your end of a deal"—may

We would like to thank participants at seminars in the Graduate School of Business and the Department of Political Science at Stanford, as well as two anonymous referees, for their helpful comments on earlier drafts of this article.

Journal of Law, Economics, and Organization vol. 6, no. 1 Spring 1990
 ISSN 8756-6222

enhance cooperation for three reasons. First, decision makers may internalize such codes of conduct (Biddle). Internalization is an inner control of behavior by first parties: if decision makers strongly believe that the norm of honesty is a legitimate rule, they are less likely to lie even if doing so is personally advantageous. Second, those injured by the violation of a norm may punish the deviant. Here, the literature has analyzed bilateral transactions (Axelrod, 1981), as well as collective goods, where a deviant may hurt many partners (Friedman; Taylor; Kurz; Green and Porter; Abreu, 1986; Abreu et al.; Coleman; Radner; Bendor and Mookherjee). Third, norms are typically backed by third-party sanctions: if Smith reneges on a deal with Jones, the latter may spread the word about the former, and other members of the same community may punish Smith in a variety of ways, despite being uninvolved in the original agreement.[1]

This third dimension of norms, third-party sanctions, has been relatively unexplored, particularly compared to the large body of literature on second-party retaliation. Yet the growing interest in the role of social institutions in bolstering cooperation suggests that we should pay more attention to third-party sanctions, for they are a quintessentially social dimension of norms: people who are not parties to an economic transaction become involved because of (1) the social ties connecting them to the transactors and (2) the informal codes of conduct of the community. (In contrast, second-party retaliation is a weak indicator of a norm because it is often indistinguishable from ordinary strategic behavior, e.g., bilateral reciprocity in a repeated game.) It is this social aspect of norms that constitutes the subject of this article.[2]

Specifically, we offer a game-theoretic analysis of the role of third-party sanctions in facilitating cooperation. We are interested in comparing the relative importance of norms, bilateral reciprocity, and more formal alternatives such as contracts or centralized control. Such analyses should help us

1. Empirically, of course, all three aspects of norms are often present simultaneously. A decision maker may feel impelled by reasons of principle to tell the truth, the victims of deception may retaliate, and indignant third parties may become involved as well. (For example, see Fenno for a discussion of both sanctions and internalization in the House Appropriations Committee.) Disentangling the strengths of these different controls is difficult—the actors themselves may be uncertain on this score—hence scholars have often chosen to examine the effects of one type of control in isolation, holding constant the other two by experimental or analytical means.

2. Our models will not directly represent internalization, or more generally the moral dimension of norms. They will, however, readily lend themselves to such interpretations. If i punishes j for cheating k—even though j's reneging did not harm i—it is quite natural to tell a story wherein i, believing in honest dealing as an impersonal code of conduct, is morally outraged by j's behavior, and for this reason punishes j. In contrast, it seems less compelling to tell a story involving internalization and moral outrage if only the victim k retaliates by not cooperating with j. In such circumstances a natural interpretation is that k is using a self-interested strategy of conditional cooperation.

understand patterns in the evolution of institutions facilitating social cooperation, as well as cross-sectional patterns across groups with different technological and structural characteristics.

The study of third-party sanctions also allows better understanding of two distinct roles of formal institutions in facilitating cooperative behavior. One role is to collect and disseminate reliable information about the actions chosen by group members toward one another. The second is to enforce coercive penalties for deviations from a commonly agreed-upon code of behavior. Many institutions, such as governments, combine the two roles. But many others, such as Better Business Bureaus, Amnesty International, medieval trade guilds, election observer groups, and some agencies of the United Nations, confine their role to the first. Their effectiveness is based on how group members use information about deviations between other pairs of members, and subsequently apply third-party sanctions. The factors that enhance the role of third-party sanctions thus also determine the importance of this class of "observer" institutions.[3]

The specific setting that we explore is one of a group with three or more members who exchange bilaterally with one another. We assume every pairwise relationship has the characteristics of a Prisoner's Dilemma. There is no collective good aspect to the group's activities: since by definition pure public goods preclude bystanders (and hence third parties), this emphasis on bilateral dealings is necessary for our analysis. For a similar reason, that is, the public good character of reputations, we assume that the multilateral game is one of complete information, that is, the payoffs of all members are common knowledge. When investigating the role of third-party sanctions, we assume that every member's actions are observed without error by the entire group: such information is presumed to be naturally available, or collected and disseminated by an "observer" institution.[4] Finally, we assume that the game is infinitely repeated, with the same set of players interacting repeatedly. Future payoffs are discounted at a constant rate; this rate represents both impatience and the chances of the game terminating.

We start by posing the following specific question: for what classes of payoff functions and discount rates can third-party sanctions support more cooperative outcomes as subgame-perfect equilibria, compared to purely bilateral sanctions? A natural conjecture is that they *always* facilitate greater cooperation, since they impose stronger punishments for deviations. Section 3 demonstrates this conjecture to be false, by identifying a class of payoff functions (satisfying the twin conditions of separability and symmetry) for

3. For a similar emphasis on the role of institutions as generators of information, see the recent paper by Milgrom, North, and Weingast.

4. See Mookherjee for a model of third-party sanctions arising in an incomplete-information context.

which third-party sanctions do not permit greater symmetric cooperation, no matter what the discount rate.

Subsequent sections demonstrate that the effectiveness of third-party sanctions must be based on the combination of two factors: (1) nonseparabilities or asymmetries[5] of different bilateral relationships and (2) intermediate values of the discount rate. Consider the latter. For sufficiently high rates of discount, neither third-party sanctions nor bilateral sanctions can support any cooperation. In such situations, explicit contracts and/or centralized control is necessary. On the other hand, for sufficiently low discounting, bilateral sanctions suffice to generate Pareto-optimal levels of cooperation, and third-party sanctions have nothing to contribute. This suggests that third-party sanctions are most important in communities with an intermediate degree of turnover and closeness of contact between members.

The roles of nonseparabilities and asymmetries of payoffs are also explored in succeeding sections, via a series of examples. Section 4 considers a game with nonseparable but symmetric payoffs, where the marginal costs of providing help to any particular member is increasing in the number of other members being simultaneously helped. For intermediate values of the discount rate, third-party sanctions are shown to permit Pareto-improving increases in the level of mutual help.

Section 5 then considers an example of a separable but asymmetric game: the group comprises two factions, and interfaction cooperation is more difficult to induce than intrafaction cooperation. In this context, intrafaction third-party sanctions are shown to facilitate interfaction cooperation for intermediate rates of discount. Their effectiveness depends on the relative sizes of the two groups—the more unequal the sizes, the less effective these third-party sanctions turn out to be. Somewhat surprisingly, unequal sizes also enhance the value of asymmetric patterns of cooperation, where majority members give more help to minority members than they receive in return. Thus, inequalities in the size of different factions strain the effectiveness of third-party sanctions, unless majorities demonstrate increasing tolerance for minorities.

One methodological caveat: we do not wish to assert the inevitability of cooperation via informal norms whenever parameters of the game permit the existence of such equilibria. All these games are characterized by multiple equilibria: in particular, perpetual defection between every pair is always an equilibrium.[6] Whether members manage to converge to some equilibrium or not through experimentation and adaptation, and which particular equilibrium they may converge to, is completely beyond the realm of our analy-

5. These terms are defined more precisely in the Appendix.

6. For an empirical argument that nearly universal defection between families is an equilibrium in some *highly stable* peasant societies, see Foster.

sis. These factors may well depend on historical accidents and the precise pattern (as well as costs and benefits) of experimentation and adaptation. We are merely concerned with understanding the potential of third-party sanctions to expand the range of cooperative outcomes that can be supported as equilibria.

Section 2 introduces the basic model and Section 6 concludes.

2. THE MODEL

There are n individuals involved in an infinitely repeated game. In each period, every player (i, say) chooses an action toward every other player (j, say): this may be interpreted as the amount of help or gift i provides to j. Let a_{ij} denote this action, which is assumed to be real-valued.[7] Also, let A denote the set of all possible levels of help that i can choose to give j: this is a subset of the real line and identical for all pairs. The minimum amount of help that i can give j is zero. Individual i therefore has to choose how much help to give to every other player in the group; the space of possible actions for i in any period is $A_i \equiv \{(a_{i1}, a_{i2}, \ldots, a_{i,i-1}, a_{i,i+1}, \ldots, a_{in}) \mid a_{ij} \in A \text{ for all } j\}$. We will use a_i to denote a specific action for i: $a_i = (a_{i1}, \ldots, a_{i,i-1}, a_{i,i+1}, \ldots, a_{in})$, an element of A_i.

Individual i's payoff W_i in a given period depends on actions $(a_1, \ldots, a_n)$ chosen by different members of the group in that period. However, since the interactions are bilateral, individual i's utility will not depend on exchanges between other pairs. Rather, they will depend only on exchanges between i and other group members:

$$W_i = W_i(a_i; a_{1i}, a_{2i}, \ldots, a_{i-1,i}, a_{i+1,i}, \ldots, a_{ni}) \tag{1}$$

Furthermore, we impose a Prisoner's Dilemma structure: i's utility is increasing in the amount of help received from others and decreasing in the amount of help given to others. If the game were to be played only once, each player would have a dominant strategy of not providing help to any other player. We assume that this outcome is Pareto-inefficient. Specifically, if every player were to give some positive level a^* of cooperation to every other player, then all of them would be better off (compared to the all-defect outcome):

$$W_i((a^*, \ldots, a^*); a^*, \ldots, a^*) > W_i((0, \ldots, 0); 0, \ldots, 0) = 0 \tag{2}$$

where the utility of each player has been normalized to zero at the no-cooperation outcome.

7. Most of our analysis extends to the case where a_{ij} is multidimensional, that is, where it encompasses many different commodities or services.

The one-period game described above is repeated infinitely often.[8] Each player applies a constant discount factor δ to utility in the following period. If $W_{i,t}$ denotes player i's payoff at date t, then i's objective is to maximize

$$\sum_{t=0}^{\infty} \delta^t W_{i,t} \qquad \text{where } 1 > \delta > 0 \tag{3}$$

We assume that the payoff functions W_i of each player are common knowledge within the group. Further, in contexts where third-party sanctions are feasible, the actions of every member with respect to all other members are observable without error to the entire group. Such information may be available naturally, for example, in tight-knit work groups or village communities, or made available by an observer institution. Such a setting can be contrasted to one where members have information only about the history of their own exchanges, thus allowing only bilateral sanctions to be feasible. The main question posed in this article concerns the effectiveness of third-party sanctions in enforcing cooperation, over and above what purely bilateral reciprocity strategies can enforce.

In our model, this question translates more formally into the following. Let $\mathbf{a}(t) = (a_1(t), \ldots, a_n(t))$ denote the outcome of the game, that is, a specification of the amount of help exchanged between *every* pair of individuals at a given date t. A *third-party sanction strategy* for player i in the repeated game can then be represented by a planned action a_i^0 at date 0, and $a_i^t(\mathbf{a}(0), \ldots, \mathbf{a}(t-1))$ at any date $t > 1$, as a function of outcomes (between every pair) at all previous dates.[9] In contrast, a *bilateral sanction strategy* for player i is one where i's action $a_{ij}(t)$ toward player j at any date t depends on the history of exchanges $[a_{ij}(0), a_{ji}(0); a_{ij}(1), a_{ji}(1), \ldots; a_{ij}(t-1), a_{ji}(t-1)]$ only between i and j (in addition to planned action a_i at date 0). Any combination of strategies for the players *generates or supports a certain outcome* for the game at each date. This outcome can be calculated by recursively applying the functions a_i^t representing strategies, along with the initial levels of cooperation $\mathbf{a}(0) = (a_1(0), \ldots, a_n(0))$.[10]

Our primary question can now be posed precisely as follows. Under what circumstances can third-party sanction strategies generate outcomes that are

8. The game does not have to be infinitely repeated, but could have an uncertain termination date: conditional on survival up to any given date there can be a constant probability that it ends at that date. This probability will affect the players' discount rate.

9. Note that a strategy thus specifies not only what the player will actually do in the course of play (i.e., "along the equilibrium path"), but also his threats of retaliation against previous defections.

10. Thus the outcome $\mathbf{a}(1)$ at date 1 is obtained as follows: player 1 chooses $a_1^1(\mathbf{a}(0))$, player 2 chooses $a_2^1(\mathbf{a}(0))$, etc. The outcome $\mathbf{a}(2)$ at date 2 is $a_1^2(\mathbf{a}(0), \mathbf{a}(1)), a_2^2(\mathbf{a}(0), \mathbf{a}(1)), \ldots, a_n^2(\mathbf{a}(0), \mathbf{a}(1))$, and so on.

"more cooperative" (in the sense of generating Pareto improvements), compared to those generated by bilateral sanction strategies? However, we confine attention to strategy configurations that are (noncooperative) perfect equilibria of the repeated game.[11] That is, they must be consistent with the self-interest of players, in the sense that no player would have an incentive to deviate unilaterally from his chosen strategy in any contingency (i.e., a given history of outcomes of the game up to that date). This must hold whether or not the contingency actually arises in the course of play; threats of retaliation to defections must be credible in this sense.

3. THE BASELINE CASE: THIRD-PARTY SANCTIONS ARE INEFFECTIVE

In this section, we explore cases where third-party sanctions never carry greater punitive effects than bilateral sanctions. Intuitively, it might seem that third-party sanctions would *always* be stronger. If i defects against j, and a third party k (as well as j) consequently retaliated against i, one would think that i would suffer a stiffer punishment, thereby lessening his initial incentive to cheat j. We show below that this reasoning is flawed: there is an interesting class of games where it does not apply. Consequently, the explanation for the effectiveness of third-party sanctions is more subtle.

Where is the reasoning flawed? One might suspect that the catch is in the requirement that third-party sanctions be credible threats, as required by our equilibrium notion. If i should cheat j, but not k, is it credible for k to threaten i? If k jeopardizes his relationship with i, doesn't k injure himself? While this counterargument may have some merit, it is *not* a consequence of our equilibrium notion. Perfect equilibrium merely requires that no person have an incentive to deviate *unilaterally* from planned actions, given the planned actions of the other players. Consequent on double-crossing j, suppose i anticipates that k will punish him by withholding cooperation. Given this expectation, it is in i's best interest to stop cooperating with k as well. And k, anticipating that i reasons as above and will therefore defect, also finds it in his best interest to stop cooperating with i. So the suspension of cooperation between i and k, following i's cheating j, is consistent with the notion of *unilateral credibility* embodied in the idea of a perfect equilibrium.

One may, of course, argue that such a credibility notion is not altogether persuasive. After i double-crosses j, bystander k would gain if he could

11. See Selten for a rigorous definition of perfect equilibria of noncooperative games. Note that we are ignoring issues of collective credibility (i.e., renegotiation-proofness) concerning these equilibria. For a discussion of renegotiation in repeated-game contexts, see Farrell and Maskin and Pearce.

arrange a (self-enforcing) agreement with i not to make their relationship contingent on trade with other partners. There are several responses to this argument. First, the group could have a norm whereby if any "noncheater" fails to punish a "cheater," he should also be excommunicated. That is, if k fails to punish i for i's defecting against j, then all other members of the group will retaliate against k. This would strengthen k's resolve to punish i. Axelrod (1986) calls such higher-order rules "meta-norms."[12] Second, if players ignore past defections in the interest of preserving future cooperation, this can happen in bilateral relationships as well. If i cheats j, one could as easily argue that instead of going into a phase of mutual punishment, they would both find it advantageous to let "bygones be bygones," and continue to cooperate as before.[13] Finally, third parties may be "morally" outraged at an infringement of a group norm and thus be motivated to retaliate.

We wish to bypass these complex issues by insisting only on the unilateral credibility of retaliatory threats. The Prisoner's Dilemma structure of every bilateral relationship allows threats of reversion to total noncooperation to be credible, at any stage of the game. Thus in the context of the perfect equilibrium approach, there is no problem with the credibility of k's threat to retaliate against i, should i cheat j.

The source of the possible ineffectiveness of third-party sanctions must therefore lie elsewhere. It would perhaps be most fruitful if we were to describe a class of games where third-party sanctions are ineffective, and show how the heuristic argument fails.

The class of games we will describe is characterized by two features: *separability* and *symmetry*. These are properties of the players' payoff functions $W_1, W_2, \ldots$. Roughly speaking, the game is *separable* if i's relationship with any player j can be decomposed (technologically, though not necessarily strategically) from his relationship with any other player. Specifically, i has a payoff function $U_{ij}(a_{ij}, a_{ji})$, representing the utility he derives from his relationship with j, which depends only on their exchanges. Naturally, U_{ij} is decreasing in a_{ij} and increasing in a_{ji}. Separability requires that player i's overall payoff is the sum of the utilities from his relationship with all other players:

$$W_i = \sum_{j \neq i} U_{ij}(a_{ij}, a_{ji}) \tag{4}$$

12. Of course, one could counterrespond by arguing that this pushes the problem one stage backward: why are the meta-norms credible? An interesting task for future research is to understand the credibility and role of meta-norms, especially when renegotiation-proof equilibria are analyzed (to capture the idea of collective rather than unilateral credibility).

13. This suggests that players will want to develop reputations for retaliating against previous defections. Suspension of cooperation must be incurred in response to defections, as a cost of building this reputation. But this applies equally well to third-party and bilateral sanctions.

The symmetry condition is also fairly intuitive: any permutation of actions chosen by players also permutes their payoffs in the same way. This is defined more precisely in Proposition 0 in the Appendix, where it is shown that the combination of separability and symmetry imply the existence of a function U such that

$$W_i = \sum_{j \neq i} U(a_{ij}, a_{ji}) \tag{5}$$

Symmetry thus imposes two additional features on the utility function $U_{ij}(a_{ij}, a_{ji})$ pertaining to the pair $\{i, j\}$. First, the functions U_{ij} and U_{ji} are identical: the relationship between i and j is symmetric. We later refer to this property as *pairwise-symmetry.* Second, all bilateral relationships are identical. That is, the same utility function $U(a_{ij}, a_{ji})$ applies to all pairs: the relationship between i and j is exactly the same as the relationship between any other pair k and l.

We are now in a position to describe games where third-party sanctions are ineffective in supporting symmetric patterns of cooperation.

Proposition 1. Assume that the game satisfies separability and symmetry, that is, condition (5) holds. Then any symmetric outcome **a**(1), **a**(2), . . . [i.e., where at any date t, every individual gives an identical amount of help $a(t)$ to every other individual] can be generated by a third-party sanction equilibrium, if and only if it can be generated by a bilateral sanction equilibrium. This holds irrespective of the discount rate.

It seems fairly obvious that an outcome can be generated by a TPS equilibrium if it can be generated by a BS equilibrium, since cooperation in the latter must be based on the threat of bilateral sanctions, which are available in a world with third-party sanctions. Consequently, the main issue is to explain the opposite relation, that is, why any outcome can be sustained in a BS equilibrium if it can be sustained in a TPS equilibrium. A heuristic explanation is the following. Suppose i is tempted to cheat j, and the threat of j's retaliating is insufficient to deter i. Then i and j cannot credibly cooperate with one another. But the relationship between i and j is "identical" to the relationship between i and any other individual k, by condition (5). So neither can i and k cooperate with one another. If that is so, k has nothing to credibly threaten i with, should i cheat j: third-party sanctions can add nothing.

To prove Proposition 1 more formally, suppose **a**(0), **a**(1), **a**(2), . . . is a symmetric outcome generated by a TPS equilibrium. Suppose that it cannot be generated by a BS equilibrium. Then in the presence of bilateral sanc-

tions alone, there must exist some date t, where some individual i wishes to cheat at least one other individual (j, say). Given the separability assumption (4), and given bilateral sanctions, we can "separate" i's decision to cheat individual j from his decision to double-cross any other individual, in the following sense. If i cheats j at t, he will do so by cooperating less than the required amount $a(t)$, whence the level of exchanges between i and j (and j alone) from date $(t + 1)$ onward will decline, following j's retaliation. Since the given outcome cannot be generated by a BS equilibrium, it must be the case that even the strongest punishment that j can credibly inflict on i by retaliating from $(t + 1)$ onward cannot deter i from cheating at t: see Abreu (1988).

Now, the strongest punishment that j can inflict on i, consistent with the requirement of unilateral credibility, is total suspension of all exchanges.[14] This gives individual i a utility level of 0 from $(t + 1)$ onward in his relationship with j, rather than the gains from trade promised under cooperation levels $a(t + 1)$, $a(t + 2)$, ... Hence, i would always deviate to zero cooperation, if he were to deviate at all, and the corresponding deviation benefit cannot be outweighed by his utility loss from $(t + 1)$ onward, from the strongest possible bilateral retaliation:

$$U(0, a(t)) - U(a(t), a(t)) > \sum_{t^*=t+1}^{\infty} \delta^{t^*-t}[U(a(t^*), a(t^*)) - 0] \qquad (6)$$

However, by hypothesis, the presence of third-party sanctions would prevent i from double-crossing j at date t. There, not only may j suspend trade with i from $(t + 1)$ onward, but other individuals k, l, etc., also may do the same. This increases i's losses from $(t + 1)$ onward: the right-hand side of (6) can increase, presumably reversing the inequality in some instances.

This reasoning is faulty for the following reason. Given that individuals k, l, etc., other than j threaten to punish i from $(t + 1)$ onward, i's decision to cooperate with j can no longer be "separated" from his decision to cooperate with k, l, etc. If k and l will suspend trade with i from $(t + 1)$ onward, i might as well also cheat them at date t. Thus, i's benefit from defecting at date t, the left-hand side of inequality (6), also increases in the presence of third-party sanctions.

In fact, (6) implies that i will prefer to defect at date t, even in a world with third-party sanctions. To see this, suppose i cheats every other individual at t, by refusing to offer any help. His utility gain at t is then $(n - 1)$ times the

14. Remember that no-trade is an equilibrium of the repeated game, thus allowing it to satisfy the criterion of unilateral credibility. Reverting to this is the worst punishment for i, because he can guarantee himself a utility level of at least zero in each period by refusing to cooperate at all. So, i cannot be forced below this utility level.

left-hand side of (6). The worst punishment that can be inflicted on him by the rest of the group is total and permanent excommunication—no trade at all with anyone from $(t + 1)$ onward. The utility loss of this is exactly $(n - 1)$ times the right-hand side of (6). Hence, (6) implies that third-party sanctions cannot provide i with the incentive to abide by the proposed cooperation level $a(t)$. This contradicts our premise that the given outcome can be generated by a TPS equilibrium, thus proving Proposition 1. (Note also that the result holds regardless of the rate at which future payoffs are discounted.)

We hasten to add that we are *not* claiming that third-party sanctions are empirically unimportant. Far from it. Proposition 1 describes sufficient conditions for third-party retaliations to be irrelevant in supporting symmetric cooperation. These conditions are "tight," that is, relaxing them allows us to generate examples in which third-party sanctions do increase the amount of sustainable symmetric cooperation. Consequently, if the game is nonseparable or asymmetric, individual and group retaliation are not invariably equivalent. Sections 4 and 5 analyze the effects of nonseparabilities and asymmetries, respectively, as well as the range of discount rates which permit third-party sanctions to be effective.

Further, note that even in the games covered by Proposition 1, the restriction to symmetric levels of cooperation is essential. Consider the following example. There are three individuals, denoted 1, 2, and 3. Each individual can give either no help, or a fixed amount $a > 0$ to any other individual, that is, $A = \{0, a\}$. Take the outcome where each individual cooperates with one other player: player 1 cooperates with 2 ($a_{12} = a$), but not with 3 ($a_{13} = 0$); player 2 gives a to player 3 but nothing to 1, and finally player 3 gives a to player 1 but nothing to 2. This outcome cannot be generated by a bilateral sanctions equilibrium: for instance, player 2 is required to be perpetually "suckered" by player 3. Since the latter gives 2 nothing, there is no sanction he can impose on 2 to persuade him to continue giving him a. But if the players are sufficiently patient, that is, δ high enough so that

$$U(0, 0) - U(a, 0) < \frac{\delta}{1 - \delta}[U(0, a) + U(a, 0) - 2U(0, 0)] \tag{7}$$

then this outcome can be supported by a third-party sanction equilibrium. For instance, suppose player 1 gives a to player 2 as long as the latter gives at least a to player 3, and reverts to no-trade otherwise. Similarly, 2 gives a to 3 as long as the latter gives the same to 1, and so on. Then (7) ensures that each player will abide by the agreement.

However, while this specific allocation of cooperation cannot be achieved by bilateral sanctions, there may exist other allocations, achievable via bilateral sanctions, that generate equivalent (or Pareto-dominating) levels of util-

ity to every individual. This is indeed the case in the above example. For instance, suppose that every bilateral relationship alternates on successive dates between the outcome where one individual helps the other, receiving nothing in return, and the outcome where the roles of the two individuals are reversed. Further, the sequencing in different bilateral relationships is coordinated so that the utility of every individual equals $[U(0, a) + U(a, 0)]$ at every date (e.g., if $a_{12} = a$, $a_{21} = 0$ at any date, then $a_{13} = 0$, $a_{31} = a$ at that date). This ensures that each individual attains identical utilities in every period, compared to the third-party sanction equilibrium discussed above. Finally, these bilateral relationships are incentive compatible if and only if

$$U(a, 0) + \delta U(0, a) \geq 0 \tag{8}$$

which is actually implied by condition (7).

The preceding result generalizes to the n-individual context, in the following sense. Suppose we arrange the n individuals in a circle, and every individual gives help (at level a) without receiving anything in return to m individuals on his left, while he receives help (also at level a) without giving anything in return to m individuals on his right, where $2m \leq n - 1$. Further, no help is given or received from the remaining $(n - 1 - 2m)$ individuals. If this outcome can be supported as an equilibrium with third-party sanctions, then there exists a bilateral sanction equilibrium (involving alternation between $a_{ij} = a$, $a_{ji} = 0$ and $a_{ij} = 0$, $a_{ji} = a$ at successive dates in every bilateral relationship) that gives every individual equivalent utility at every date.

However, we have been unable to establish a general result to the effect that under the conditions of Proposition 1, corresponding to *any* outcome (symmetric or otherwise) achievable via third-party sanctions, there exists a bilateral sanction equilibrium that (weakly) Pareto dominates it.[15] For instance, consider the case $n = 4$, where $A = \{0, a\}$, and, of course, $U(0, a) > U(a, a) > U(0, 0) = 0 > U(a, 0)$. Suppose also that the following conditions hold:

$$U(a, a) + U(a, 0) + U(0, a) \geq 2(1 - \delta)\, U(0, a) \tag{9a}$$

$$U(a, a) < (1 - \delta)\, U(0, a) \tag{9b}$$

$$U(a, a) \geq \frac{U(a, 0) + U(0, a)}{2} \tag{9c}$$

(9a) ensures that the outcome where every individual participates in one reciprocal relationship ($a_{ij} = a = a_{ji}$), and two unequal relationships (receiv-

15. Note that separability and symmetry of payoffs does not imply that asymmetric outcomes can be ignored: they may be intrinsically desirable [in the previous example if $U(0, a) + U(a, 0) > 2U(a, a)$], or they may be supportable as equilibria whereas symmetric outcomes cannot (see the following example).

ing help in one and giving help in the other) is a third-party sanction equilibrium. Condition (9b) ensures that constant reciprocal cooperation cannot be supported as a bilateral sanction equilibrium: consequently, any viable cooperation must necessarily be asymmetric. Conditions (9a) and (9b) together imply that two players alternating between $(0, a)$ and $(a, 0)$ on successive dates can be a bilateral sanction equilibrium. However, it can be checked that condition (9c) implies that such forms of alternation will leave at least one individual with lower utility than at the third-party sanction equilibrium. This leaves open the question of whether there exist more complicated forms of nonstationary bilateral cooperation, or patterns of randomized cooperation levels, achievable via bilateral sanctions, such that each individual is at least as well off as in the given third-party sanction equilibrium. Thus, the effectiveness of third-party sanctions in supporting asymmetric cooperation outcomes in separable, symmetric games remains unresolved in general.[16]

4. NONSEPARABLE GAMES

In this section, we examine the role of the separability assumption in precipitating the irrelevance of third-party sanctions, as expressed in Proposition 1. We show that when the interactions between different pairs are linked (e.g., if i's marginal cost of cooperating with j depends on levels of cooperation with other players), then third-party sanctions can indeed be effective in generating higher levels of (symmetric) cooperation.

Consider the following example. To isolate the role of nonseparabilities, the game is taken to satisfy the symmetry assumption of Proposition 1. Player i has the following per-period payoff:

$$W_i = b\sum_{j\neq i} a_{ji} - C\left(\sum_{j\neq i} a_{ij}\right) \quad \text{where } b > 0, \text{ and } C \text{ is an increasing, strictly convex function} \tag{10}$$

Payoffs depend separately on benefits of help received and costs of helping others. While benefits are linear, the marginal cost to i of providing an additional unit of help to j depends on the amount of help currently given by i to other players. This reflects limited capacity to provide help to others: for example, giving help may require devoting time to other people's problems, and the marginal cost of spending more time on j's problem may depend on

16. We mention one restriction that should probably suffice to ensure the ineffectiveness of third-party sanctions in general. If the payoffs U are additively separable in help given and received: $U(a_{ij}, a_{ji}) = U_1(a_{ji}) - U_2(a_{ij})$, where U_1 and U_2 are both increasing functions satisfying $U_1(0) = 0 = U_2(0)$, then it is the case that for any $a > 0$, $U(a, a) = U(a, 0) + U(0, a)$. This, of course, rules out conditions like (9a) and (9b). Symmetric cooperation both Pareto dominates alternation between $(a, 0)$ and $(0, a)$, and can be supported as an equilibrium just as easily.

the total time already devoted to helping members other than j. As i devotes more and more time to helping others, it also cuts into the time available for increasingly valuable personal ends.

Proposition 2. Assume payoffs are given by (10). Then third-party sanctions can help support higher levels of stationary symmetric cooperation (where each player gives every other player the same level of cooperation, constant across all dates). Further, the increased levels of cooperation may permit every player to be better off.

The reasoning is as follows. If a stationary, symmetric level $a^\dagger$ of cooperation can be sustained by a bilateral sanction equilibrium, it must satisfy the condition that no player has an incentive to deviate to zero help with respect to any single partner at any date. The benefit is the saving in cooperation cost $[C((n-1)a^\dagger) - C((n-2)a^\dagger)]$. The worst possible punishment that the victim can inflict, in retaliation, is to suspend trade with the cheater at all subsequent dates. Hence

$$C((n-1)a^\dagger) - C((n-2)a^\dagger) \geq \frac{\delta}{1-\delta}\Big\{(n-1)ba^\dagger - C((n-1)a^\dagger) - [(n-2)ba^\dagger - C((n-2)a^\dagger)]\Big\} \quad (11)$$

which implies

$$C((n-1)a^\dagger) - C((n-2)a^\dagger) \leq \delta ba^\dagger \quad (12)$$

Equation (12) is clearly a *necessary* condition for cooperation level $a\dagger$ to be sustained by bilateral sanctions. The convexity of costs implies that it is also sufficient. Providing help to one partner is more costly at the margin than the cost of helping a whole group, whereas the benefits are proportional.[17]

Hence the *maximum* level a_B of cooperation that can be sustained by bilateral sanctions alone is given by the solution to

$$\frac{1}{a_B}[C((n-1)a_B) - C((n-2)a_B)] = \delta b \quad (13)$$

17. Consider the *average* costs saved by cheating, per relationship that is thereby jeopardized. When member i cheats one partner, the (average) cost saved immediately is simply $C((n-1)a^\dagger) - C((n-2)a^\dagger)$. When he cheats in l relationships, the average cost saved is $(1/l)[C((n-1)a^\dagger) - C((n-l-1)a^\dagger)]$, which is smaller. Since players compare the average cost saved immediately to the average benefit lost per relationship in the next period (which is $\delta ba^\dagger$), it follows that each member is most tempted to double-cross exactly one partner.

where each player is just indifferent between cheating and not cheating exactly one partner.

What can third-party sanctions accomplish? If any individual i double-crosses one partner, the entire group can suspend trade with him thereafter. Anticipating this, it would be most profitable for i to cheat *all* his partners. This will not be worthwhile if

$$C((n-1)a^{\dagger}) - C(0) < \frac{\delta}{1-\delta}\left\{(n-1)ba^{\dagger} - C((n-1)a^{\dagger}) - [0 - C(0)]\right\} \tag{14}$$

that is, if

$$C((n-1)a^{\dagger}) < \delta(n-1)ba^{\dagger} \tag{15}$$

Clearly, (15) is also sufficient for $a^{\dagger}$ to be supported by a third-party sanction equilibrium. So the maximum level of cooperation a_T that third-party sanctions can sustain satisfies

$$\frac{1}{(n-1)a_T}C((n-1)a_T) = \delta b \tag{16}$$

Comparing this with equation (13), it follows that $a_B < a_T$, that is, third-party sanctions can sustain higher levels of cooperation. A cooperation level between a_B and a_T cannot be sustained by bilateral sanctions: each individual will do better to cheat at least one partner, since the additional cost of fulfilling this obligation outweighs the prospect of foregoing benefits from the single victim in the future. With third-party sanctions, any deviation can be punished by suspension of trade with the entire group, not just the victims. Consequently, each individual has to balance cheating everyone today against the cost of being ostracized by *everyone* from tomorrow. The convex interdependence of the costs of cooperation implies that each member may find it profitable to jeopardize one relationship, but not all of them simultaneously.

To demonstrate that third-party sanctions may enable Pareto-improving increases in cooperation, consider the example where the cost function takes the form $C(a) = a^2$. Then from (13) and (16), we obtain $a_B = \delta b/(2n-3)$, $a_T = \delta b/(n-1)$. The Pareto-optimal level of (stationary, symmetric) cooperation a^*, which maximizes the per-period utility $[b(n-1)a - (n-1)^2a^2]$ of the representative player is given by $b/[2(n-1)]$. Hence, if $\delta < (2n-3)/(2n-2)$, bilateral sanctions will not permit a^* to be achieved. In such cases, third-party sanctions will allow Pareto-improving increases in cooperation. In fact, if $\delta > \frac{1}{2}$, they will permit the Pareto-optimal level a^* to be achieved.

Third-party sanctions are effective for intermediate values of the discount

rate. If $\delta > (2n - 3)/(2n - 2)$, bilateral sanctions suffice to achieve the symmetric Pareto-optimal level of cooperation, and third-party sanctions add nothing. On the other hand, for values of δ approaching 0, both bilateral and third-party sanctions permit only vanishing levels of cooperation: here explicit contracts or centralized monitoring-cum-incentive schemes are necessary.

Note also that the disparity between a_B and a_T, as well as between a_B and a^*, grows with group size n. In this sense, then, third-party sanctions are more effective in larger groups. The ratio a_T/a^* is independent of n (this result generalizes to any constant elasticity cost function); *in this sense third-party sanctions prevent deterioration of cooperation as group size increases.*

5. ASYMMETRIC GAMES: THE TWO-FACTION MODEL

This section explores the effectiveness of third-party sanctions in asymmetric games. To isolate the role of asymmetries, we consider games where different bilateral relationships are separable, unlike the previous section. Imagine two departments in a government bureaucracy. Because of common professional training and socialization, cooperation between two bureaucrats within the same department is often easier than cooperation between members of different departments.[18] Or consider two ethnic groups that are part of a common political system: a group member may prefer helping those in his own group to helping members from the other group. However, despite the greater difficulty of cooperating across groups, it may be Pareto-inefficient for these two groups not to cooperate, that is, across-group relationships may also constitute a Prisoner's Dilemma.

To simplify the analysis, assume that each individual can, in every pairwise relationship, choose any level of help between 0 and some upper limit $\bar{a}$. Payoffs are separable, that is, each individual's per-period payoff W_i is the sum of his payoffs U_{ij} from different bilateral relationships, as represented in equation (4). Players belong to either of two factions, a majority faction containing M members and a minority faction containing m ($\leq M$) members. The benefit from receiving help is the same for across- and within-faction relationships, but giving help is less costly in an intrafaction relationship. We assume that i's payoff from trade with j is

$$U_{ij}(a_{ij}, a_{ji}) = ba_{ji} - c_k a_{ij} \tag{17}$$

where k takes two possible values H or L, depending on whether i and j belong to the same faction or not. c_H, the cost of helping in an across-faction

18. Consider, for example, cooperation between the Army and the Air Force versus cooperation within the Army.

relationship, exceeds c_L, the corresponding cost in a within-faction relationship. Both kinds of relationships are Prisoner's Dilemmas:

$$b > c_H > c_L > 0 \tag{18}$$

so increases in mutual cooperation make both parties better off. Consequently, the Pareto-optimal level of symmetric cooperation is the maximum possible amount $\bar{a}$ of help that one can give another.

This game is asymmetric because within-faction and across-faction relationships are different: (5) does not hold, despite separability. Note however that the game is *pairwise symmetric:* each bilateral relationship is symmetric (so $U_{ij} = U_{ji}$: if it is costly for i to help j, it is also costly for j to help i).

Consider first the range of discount factors for which third-party sanctions have any potential to be effective. We focus attention on stationary (i.e., time-independent) levels of cooperation.

Lemma A. If $\delta > c_H/b$, then bilateral sanctions can help generate maximal cooperation (i.e., $a_{ij} = \bar{a}$) in every relationship, across-faction as well as within-faction.

The proof of Lemma A is straightforward.[19] The next result (which is proven in the Appendix) describes another range of discount factors where third-party sanctions are ineffective, although for the opposite reason.

Lemma B. If $\delta < c_L/b$, then no positive level of (stationary) cooperation is possible in any equilibrium, for either bilateral or third-party sanctions.

Lemmas A and B together imply that the interesting range of discount factor values is

$$\frac{c_H}{b} > \delta \geq \frac{c_L}{b} \tag{19}$$

For this range of discount rates, bilateral sanctions sustain maximal cooperation within factions, but cannot sustain any cooperation at all across fac-

19. Since without loss of generality, every defection is punished by reversion to zero cooperation, it is most profitable for i to cheat j by not cooperating at all in any period, if he is to cheat at all. For an across-faction relationship, the immediate gain from this is the saving on cooperation cost $c_H\bar{a}$, and the loss from tomorrow onward is $[\delta/(1 - \delta)]\ (b\bar{a} - c_H\bar{a})$ in present-value terms; the latter outweighs the former. Since $c_H > c_L$, the same is true for a within-faction relationship as well.

tions. Can third-party sanctions facilitate cooperation between factions? If so, what kinds of norms would work?

An obvious candidate is the following between-faction norm: individual i_A from faction A warns individual j_B from faction B that if j_B cheats i_A, then not only will i_A never trade again with j_B, but also none of i_A's colleagues from faction A will ever cooperate with j_B. However, this norm cannot work, for reasons similar to those in Section 3. If *all* members of faction A threaten to punish j_B for deviating against one of them, it makes sense for j_B to cheat everyone in A, rather than just one of them. So, while the punishment is multiplied by some factor, the immediate benefits from deviating are also multiplied by the same factor. Individual j_B will find it profitable to deviate against the entire rival faction simultaneously, rather than each of the faction members separately.[20] The net result is the same: mutual defection between factions.

Suppose, however, there is the following, more universalistic norm: if member j_B cheats anyone—inside or outside his faction—then all members of both factions stop trading with him.[21] In particular, j_B's *own* colleagues in faction B threaten to discipline him *for double-crossing people from the other faction.*[22]

The possibility of within-faction sanctions for across-faction transgressions can help support cooperation across factions. Consider initially the case where cooperation levels are also pairwise-symmetric, so across-faction relationships are associated with mutual cooperation level of a_A, and within-faction relationships with a level of a_W. Given both within- and across-group sanctions for any defection, if j_B has to cheat, he may as well double-cross every partner simultaneously. If j_B is a minority member, he will not cheat as long as

$$c_L(m - 1)a_W + c_H M a_A \leq \delta b[(m - 1)a_W + M a_A] \tag{20}$$

while if he is a majority member, the corresponding constraint is

20. Put differently, if i_A cannot cooperate with j_B, neither can any other member k_A in faction A. Then k_A has nothing to threaten j_B with, in order to retaliate against j_B's double cross of his colleague i_A.

21. One can tell different stories about this norm. One interpretation is that if j_B cheats someone, everyone else anticipates that j_B will double-cross them thereafter. Hence in self-defense everyone stops cooperating with j_B. Alternatively, the norm may be internalized, so everyone punishes j_B for violating it even if they were unaffected by the transgression.

22. One could tell the following story about this part of the norm: if any member of one faction (B, say) double-crosses anyone in the other faction, everyone in A will stop cooperating with everyone in B. Anticipating this collective punishment, everyone in B warns each other that if any one of them cheats anyone in the other faction, then every member of B will also punish the cheater. Meta-norms could strengthen this: members of A may threaten to stop cooperating with everyone in B if they fail to punish their deviant member.

$$c_L(M-1)a_W + c_H m a_A \leq \delta b[(M-1)a_W + m a_A] \tag{21}$$

These two constraints reduce to the following condition on the discount factor:

$$\delta \geq \frac{[(m-1)a_W]c_L/b + (Ma_A)(c_H/b)}{(m-1)a_W + Ma_A} \tag{22}$$

which says that δ should exceed a weighted average of c_H/b and c_L/b. Hence, if δ lies between c_H/b and c_L/b, as we assumed in (19), and if we want within-faction relationships to be characterized by maximum cooperation ($a_W = \bar{a}$), then a certain amount of a cross-faction cooperation can be supported. In fact the *maximum* across-faction cooperation that can be supported is [obtained by converting (22) to an equality, and imposing the upper bound $\bar{a}$]:

$$a_A^* = \min\left[\frac{m-1}{M}\left(\frac{\delta - c_L/b}{c_H/b - \delta}\right), 1\right]\bar{a} \tag{23}$$

Equation (23) shows that the upper limit on across-faction cooperation depends on two principal factors: (1) the size of the discount factor δ: specifically, how close δ is to c_H/b rather than to c_L/b and (2) the *relative sizes of the two factions:* the more lopsided these two sizes are, the smaller the amount of across-faction cooperation.

This second effect seems especially interesting. Its importance can be explained as follows. Across-faction cooperation is induced by the prospect of jeopardizing within-faction ties. The strength of this incentive depends on the ratio between the number of within-faction relationships that may be risked following a defection, and the number of across-faction relationships that the member would ordinarily prefer not to sustain. This ratio is more unfavorable for minority members who have to maintain more across-faction ties, with fewer within-faction relationships as inducements—and depends on the relative number $[(m-1)/M]$ of within- to across-faction relationships. *Absolute as well as relative increases in the size of the minority group can thus increase cooperation across groups.*

So far we have confined attention to pairwise-symmetric levels of cooperation, where every across-faction relationship is characterized by equality between the amount of cooperation given and received. We now argue that allowing pairwise asymmetries in across-faction cooperation levels will often permit increases in the amount of cooperation flowing in either direction, making *everybody* better off. This is particularly so when the two groups are of extremely unequal sizes.

More concretely, the idea is the following. We have seen that minority

members pose the binding constraint in supporting across-faction cooperation, partly because of weaker internal discipline and partly because they must support more across-faction ties per capita than do majority members. So, one way of inducing minority members to give more to majority members is to make these across-faction ties more attractive to them. This may be the case if in each across-faction relationship, *the majority member gives more to the minority member than he receives.*

Let g denote the amount of cooperation given by a majority member to a minority member, and r the help given by a minority member to a majority member. As before, suppose that maximum cooperation $\bar{a}$ prevails within factions. This outcome can be made consistent with individual incentives by third-party sanctions of the kind described above, if the following two constraints are satisfied. For a minority member we require

$$(m - 1)c_L\bar{a} + Mc_Hr \leq \delta b[(m - 1)\bar{a} + Mg] \tag{24}$$

and for a majority member

$$(M - 1)c_L\bar{a} + mc_Hg \leq \delta b[(M - 1)\bar{a} + mr] \tag{25}$$

Using K to denote $[(\delta b - c_L)/c_H]\bar{a}$, these two contraints reduce to the following linear inequalities in terms of r and g:

$$r \leq \left(\frac{\delta b}{c_H}\right)g + \left(\frac{m - 1}{M}\right)K \tag{26}$$

$$g \leq \left(\frac{\delta b}{c_H}\right)r + \left(\frac{M - 1}{m}\right)K \tag{27}$$

In addition, we must impose the physical feasibility constraints:

$$0 \leq r \leq \bar{a} \qquad 0 \leq g \leq \bar{a} \tag{28}$$

For a specific parametric circumstance, the set of feasible levels of across-faction trade is sketched in Figure 1. The maximal level of pairwise-symmetric across-faction trade (a_A^*, a_A^*), as given by equation (23), is the point where the line representing the minority member's incentive constraint intersects the 45° line. It is apparent from Figure 1 that if we allow majority members to give more than they receive from minority members, then increases in the amount of cooperation, both given and received, may become possible. In fact, there is a unique maximal amount of cooperation that can be sustained by third-party sanctions. The following proposition provides the precise formulae for the maximal amount of across-faction cooperation.

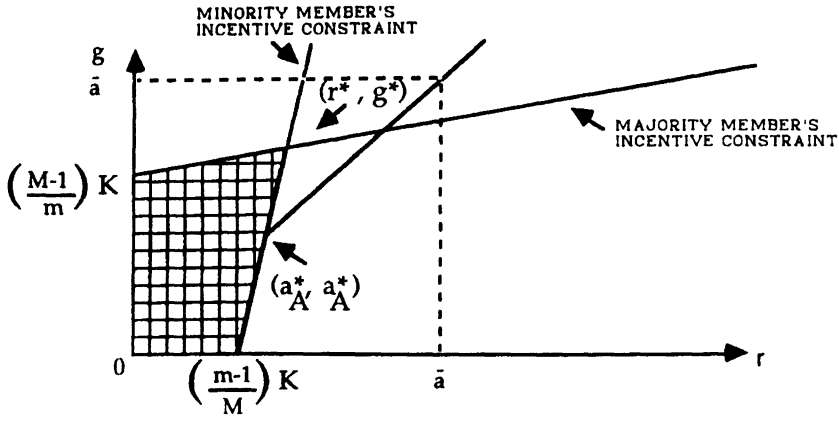

Figure 1.

Proposition 3. There exist unique maximal levels (g^*, r^*) of across-faction cooperation, characterized as follows. There are critical discount factors δ^* and δ^{**} satisfying

$$\frac{c_H}{b} > \delta^{**} > \delta^* > \frac{c_L}{b} \tag{29}$$

such that if:
(i) δ lies between δ^* and c_L/b,

$$r^* = \left[1 - \left(\frac{\delta b}{c_H}\right)^2\right]^{-1}\left(\frac{\delta b - c_L}{c_H}\right)\bar{a}\left[\left(\frac{\delta b}{c_H}\right)\left(\frac{M-1}{m}\right) + \left(\frac{m-1}{M}\right)\right] \tag{30}$$

$$g^* = \left[1 - \left(\frac{\delta b}{c_H}\right)^2\right]^{-1}\left(\frac{\delta b - c_L}{c_H}\right)\bar{a}\left[\left(\frac{\delta b}{c_H}\right)\left(\frac{m-1}{M}\right) + \left(\frac{M-1}{m}\right)\right] \tag{31}$$

which are both less than $\bar{a}$ (as in Figure 1).
(ii) δ lies between δ^* and δ^{**},

$$r^* = \left(\frac{\delta b}{c_H}\right)\bar{a} + \left(\frac{m-1}{M}\right)\left(\frac{\delta b - c_L}{c_H}\right)\bar{a} \tag{32}$$

and $g^* = \bar{a}$. In this case, r^* is less than $\bar{a}$.
(iii) δ lies between δ^{**} and c_H/b,

$$r^* = g^* = \bar{a} \tag{33}$$

In cases (i) and (ii), majority members give more than they receive from minority members (that is, $g^* > r^*$).

The proof of this is straightforward, and the details are given in the Appendix. We draw attention to three main implications:

1. To ensure that minority members give more help than in the symmetric solution (a_A^*, a_A^*), *majority members have to tolerate some inequality in their individual relationships with minority members.* Except when third-party sanctions can sustain maximum across-faction cooperation levels of $\bar{a}$ [as in case (iii) above], maximal cooperation requires that majority members give more than they receive.
2. However, *tolerating this inequality may permit everybody to be better off*, that is, the increased cooperation can be Pareto-improving. To see this, we calculate the per-period utility levels of majority members:

$$W_M = (M - 1)(b - c_L)\,\bar{a} + m(br - c_H g) \tag{34}$$

and of minority members:

$$W_m = (m - 1)(b - c_L)\,\bar{a} + M(bg - c_H r) \tag{35}$$

The first terms of (34) and (35) represent the utility from within-faction trades, and the second terms represent the benefit from across-faction trades. It is easy to check, and not surprising, that minority members always benefit from the move from the symmetric outcome (a_A^*, a_A^*) to the maximal asymmetric one (r^*, g^*). Using (34), it is easily checked that majority members benefit if and only if

$$\delta b^2 > c_H{}^2 \tag{36}$$

This condition is consistent with our basic assumption (19) that δ lies between c_L/b and c_H/b. Thus, the increased amount of giving may be worthwhile even for majority members, despite receiving less in return.[23]
3. The more unequal the two groups are in size, the more unequal are the maximal amounts of cooperation given and received by majority members. Further, if condition (36) is satisfied, the greater is the welfare increment for *every* member in moving from the best "equal" outcome (a_A^*, a_A^*) to the best "unequal" outcome (r^*, g^*), as group sizes become less equal.

The first part of (3) is apparent from Figure 2, which shows how Figure 1 is affected by a decrease in the relative size m/M of the minority. The second

23. Note that given any value for δb where either case (i) or (ii) in Proposition 3 apply, if b is sufficiently large relative to δ, then condition (36) will be met. It is also interesting to note that condition (36) is independent of the sizes of the groups.

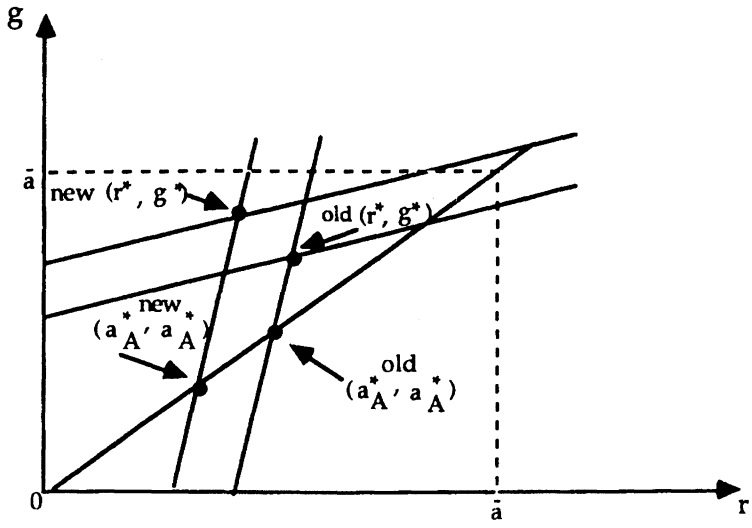

Figure 2.

part follows from the fact that for any unit increase in cooperation received (r) by a majority member, the additional amount of help he has to give in return is $\delta b/c_H$, which is independent of the sizes of the two groups. From Figure 2, it is clear that when the minority shrinks relative to the majority, the total increase in cooperation (undertaken according to the above exchange ratio) sustainable by allowing inequality is itself increased.

These three implications suggest the importance of norms permitting minority members to "exploit" majority members by returning less aid than they receive, particularly when the groups differ significantly in size. To the extent that such asymmetries are perceived as "unfair" by majority members, inequality in group size will increasingly strain interfaction cooperation.

6. CONCLUSION

We start by summarizing some implications of our analysis.

1. Third-party sanctions matter only when future payoffs are discounted at intermediate rates. This suggests that norms are an important form of social control when groups are stable enough to make decentralized third-party sanctions feasible, yet not so stable that bilateral reciprocity suffices on its own. Very unstable groups must turn to more centralized and specialized methods of enforcement, such as legal controls. This hypothesis regarding the intermediate "window of

opportunity" for norms may be of interest to historians of economic institutions.

2. The effectiveness of third-party sanctions also requires nonseparabilities or asymmetries in the relationships between different pairs. In general, a theme runs through all the examples where third-party sanctions matter: the incentive constraints display surpluses and deficits. Surpluses in some relationships are used to cover deficits in others, thereby making it worthwhile to cooperate. This theme suggests the following heuristic for a scholar trying to describe a cooperation-enhancing norm in a particular institution: first, locate bilateral relations with surpluses, and second, hypothesize a rule that in effect transfers some of these surpluses to relations that cannot be sustained by individual retaliation.[24] Conceivably, the evolution of norms that promote cooperation is a trial-and-error process involving such reallocations of incentive surpluses.
3. With nonseparabilities, the precise form of nonseparability matters. For instance, with nonseparable costs third-party sanctions require increasing marginal costs of cooperating with larger numbers of partners. With decreasing marginal costs on the other hand, third-party sanctions will not help enforce greater cooperation.
4. In asymmetric games of the two-faction model type considered, across-group cooperation requires third-party sanctions exercised within groups for across-group infractions. Their effectiveness depends upon inequality of sizes of different groups, and also on the willingness of majority groups to tolerate unequal flows of cooperation in their relationship with minority members.
5. The effects of group size on cooperation based on third-party sanctions are interesting and complex. In some models, such as the one based on nonseparable but symmetric payoffs, larger group size enhances the effectiveness of third-party sanctions.[25] They prevent intensification of free-rider problems as group size increases (again subject to the qualification of footnote 24). This is in sharp contrast to free-rider problems arising in the presence of collective goods, as elaborated in Olson and Bendor and Mookherjee.

24. An analogous idea—that it would be easier for a pair of actors to cooperate if they encounter each other in many arenas rather than just a single arena—has been explored in models of industrial organization (Bernheim and Whinston). And it is the reallocation of surpluses that is the driving force in Bernheim and Whinston's model as well. [Kissinger's notion of "linkage" of policy issues in relations between states has the same flavor as Bernheim and Whinston's model, though it is less fully developed (129). For a formalization of issue linkage, see McGinnis.]

25. Our model, of course, abstracts from the presence of information costs, which might grow with group size.

In the two-faction (asymmetric) model on the other hand, increases in the size of the whole group have no impact at all. What matters is the *relative* size of the two factions. Increasing the size of the majority faction worsens the problem of between-faction cooperation, while increases in the size of the minority eases the problem.

We now proceed to discuss some interesting extensions of our analysis. Our model assumed that actions are perfectly monitored and payoffs are commonly known. In the presence of imperfect monitoring, punishment strategies will have to be carefully chosen (since they may actually be effected "in equilibrium"). There is also the additional complication arising from "gossip" and its strategic implications. For instance, member i may threaten to sully j's record, to gain an advantage in their relationship. To moderate such incentives, the need for an institution to mediate the exchange of information, and verify claims made, is heightened.

Introducing private information about players' payoffs would open interesting research questions about the relation between reputation formation and norms. Of the three basic methods of implementing norms (first-party internalization, second-party retaliation, and third-party sanctions), we suspect that the first and third are particularly germane to analyses involving incomplete information.

Norms are internalized via imperfect socialization processes. This imperfection creates variability in players' "character" or types. For example, honorable players will take seriously the rule of "live up to your end of a deal," whereas egoistic players will not. If internalization is sufficiently strong, an honorable player will not be tempted to defect if her partner cooperates (i.e., defection will not be a dominant strategy). Naturally, both egoistic and honorable players would prefer to deal with honorable players, particularly if encounters are not repeated. If a player's type is not common knowledge, egoistic players will have an incentive to develop reputations for being honorable. Hence, the imperfect and variable internalization of norms may help explain why certain kinds of reputation formation occur.

The relation between incomplete information and third-party sanctions is more subtle. Consider a series of one-shot bilateral encounters between players who are either honorable or egoistic. Initially a player's type is private information. Actions are perfectly monitored by everyone. Suppose payoffs are such that in period one the types separate: honorable players cooperate with whomever they encounter; egoists defect. Thereafter honorable players cooperate with each other but always "punish" egoists. One could interpret this either as third-party sanctions backing a norm of cooperation, or as bilateral responses in a multiperiod game, where the revelation of i's type make it clear to j what her optimal action is. Clearly, disentangling these two processes will require careful theoretical and empirical work.

The effects of group size on cooperation secured by third-party sanctions—implication (5) above—suggests connections between networks of bilateral exchanges and collective activities of groups. Insofar as larger groups face increasingly severe free-rider problems regarding collective activities, and rely more on third-party sanctions to secure cooperation (e.g., in the nonseparable context), we may observe greater importance placed on socialization and overlapping networks of bilateral interactions in large (and successful) collective action groups (e.g., see Hardin). Linkages between private and collective activities may allow selective incentives, such as social ostracism, to be enforced even in large groups via informal norms rather than centralized forms of control.

Our analysis accorded no explicit role to "meta-norms." This derived from confining attention to subgame-perfect equilibria in Prisoner's Dilemma games. It will be interesting to explore their possible role in alternative models of interaction, especially when ideas of collective rather than unilateral credibility are incorporated via a suitable notion of "renegotiation-proofness."

Finally, it would be useful to explore alternative kinds of asymmetries: prominent among these are temporal asymmetries where any one member encounters different partners at different dates. These would be relevant to the account of norms provided by Ellickson (1989) of the whaling industry, as well as to notions of organizations (such as firms) as entities permitting cooperation between different generations of members (Bull; Cremer).

APPENDIX

Let α be any permutation function from N to N, where N denotes the set of players. That is, we associate with any player i the player $\alpha(i)$. Naturally, there exists a unique player j such that $\alpha(j) = i$, for α to be a permutation. Take any vector of actions $(a_1, \ldots, a_n)$, and permute them using α to obtain the new vector of actions $(a'_1, \ldots, a'_n)$, that is, satisfying

$$a_{\alpha(r),\alpha(s)} = a'_{r,s} \qquad \text{for each pair } r, s \tag{A1}$$

Thus, the exchange between any pair $\{r, s\}$ in the new vector is replicated by the exchange between the corresponding pair $\{\alpha(r), \alpha(s)\}$. Given any permutation α, and any pair of actions satisfying (A1), symmetry requires that

$$W_r(a'_1, \ldots, a'_n) = W_{\alpha(r)}(a_1, \ldots, a_n) \qquad \text{for all } r \tag{A2}$$

Given that r's relationship with any other individual s in $(a'_1, \ldots, a'_1)$ is replicated by the corresponding player $\alpha(r)$'s relationship with $\alpha(s)$ in $(a_1, \ldots, a_n)$, symmetry requires that r's payoff in the former case equals

$\alpha(r)$'s payoff in the latter. The game thus "looks identical" from every player's point of view.

The following result provides a clearer understanding of what the combination of separability and symmetry entails, for the structure of relationships between players:

Proposition 0. Separability and symmetry imply that each player i's payoff W_i may be written as follows:

$$W_i = \sum_{j \neq i} U(a_{ij}, a_{ji}) \tag{A3}$$

where U is a real-valued function defined on $A \times A$.

Proof. Separability implies $W_i = \Sigma_{j \neq i} U(a_{ij}, a_{ji})$. Suppose $\{W_i\}$ satisfies symmetry.

We first show that $U_{ij}(e_1, e_2)$ equals $U_{im}(e_1, e_2)$ for all $e_1, e_2 \in A$. Choose permutation α satisfying $\alpha(m) = j$, $\alpha(j) = m$, $l = \alpha(l)$ for all l not equal to j or m. Choose any $e_1, e_2, e_3, e_4 \in A$, and let $a_{im} = e_1$, $a_{mi} = e_2$, $a_{ij} = e_3$, $a_{ji} = e_4$. Choose a' satisfying $a'_{rs} = a_{\alpha(r),\alpha(s)}$, so that $a'_{ij} = a_{\alpha(i),\alpha(j)} = a_{im} = e_1$, $a'_{ji} = a_{mi} = e_2$, $a'_{im} = a_{\alpha(i),\alpha(m)} = a_{ij} = e_3$, $a'_{mi} = a_{ji} = e_4$, and $a'_{il} = a_{i,\alpha(l)} = a_{il}$ for any l not equal to j or m. Then symmetry implies

$$\begin{aligned}
W_i(a) &= U_{ij}(a_{ij}, a_{ji}) + U_{im}(a_{im}, a_{mi}) + \sum_{l \neq j,m,i} U_{il}(a_{il}, a_{li}) \\
&= U_{ij}(e_3, e_4) + U_{im}(e_1, e_2) + \sum_{l \neq j,m,i} U_{il}(a_{il}, a_{li}) \\
&= W_i(a') \\
&= U_{ij}(a'_{ij}, a'_{ji}) + U_{im}(a'_{im}, a'_{mi}) + \sum_{l \neq j,m,i} U_{il}(a'_{il}, a'_{li}) \\
&= U_{ij}(e_1, e_2) + U_{im}(e_3, e_4) + \sum_{l \neq j,m,i} U_{il}(a'_{il}, a'_{li})
\end{aligned}$$

This implies that

$$U_{ij}(e_1, e_2) - U_{ij}(e_3, e_4) = U_{im}(e_1, e_2) - U_{im}(e_3, e_4)$$

Set $e_3 = 0 = e_4$, and we obtain $U_{ij}(e_1, e_2) = U_{im}(e_1, e_2)$. This implies that

$$W_i = \sum_{j \neq i} U_i(a_{ij}, a_{ji})$$

We now claim that $U_i = U_j$, for all i and j. Choose the permutation $\alpha(i) = j$, $\alpha(j) = i$, and $\alpha(l) = l$, for all $l \neq i, j$. Choose, for any $e_1, e_2 \in A$, $a_{ij} = a_{il} = a_{lj} = e_1$, $a_{ji} = a_{li} = a_{lj} = e_2$, for all $l \neq i, j$. Then a' is given by $a'_{ij} = a'_{li} = a'_{lj} = e_2$, $a'_{ji} = a'_{il} = a'_{jl} = e_1$, and

$$W_i(a') = (n-1)U_i(e_1, e_2) \qquad W_j(a') = (n-1)U_j(e_1, e_2)$$

Since $j = \alpha(i)$, symmetry requires $U_i(e_1, e_2) = U_j(e_1, e_2)$. Q.E.D.

Proof of Lemma B. Suppose the result is false, and there exists a stationary outcome where a_{ij} is the amount of help given by i to j, and there exists at least one pair (i, j) for whom $a_{ij} > 0$. Denoting the set of minority members by m^*, and the set of majority members by M^*, we have the following incentive constraint for a minority member ($i \in m^*$):

$$C_H \sum_{k \in M^*} a_{ik} + C_L \sum_{\substack{j \neq i \\ j \in m^*}} a_{ij} \leq \delta b \left[\sum_{\substack{j \neq i \\ j \in m^*}} a_{ji} + \sum_{k \in M^*} a_{ki} \right]$$

and for a majority member ($k \in M^*$):

$$C_H \sum_{i \in m^*} a_{ki} + C_L \sum_{\substack{r \neq k \\ r \in m^*}} a_{kr} \leq \delta b \left[\sum_{i \in m^*} a_{ik} + \sum_{\substack{r \neq k \\ r \in M^*}} a_{rk} \right]$$

Adding up these constraints across all members (minority and majority), we obtain

$$C_H \left[\sum_{i \in m^*} \sum_{k \in M^*} a_{ik} + \sum_{k \in M^*} \sum_{i \in m^*} a_{ki} \right] + C_L \left[\sum_{i \in M^*} \sum_{\substack{j \neq i \\ j \in m^*}} a_{ij} + \sum_{k \in M^*} \sum_{\substack{r \neq k \\ r \in M^*}} a_{kr} \right]$$

$$\leq \delta b \left[\sum_{i \in m^*} \sum_{k \in M^*} a_{ki} + \sum_{i \in m^*} \sum_{\substack{j \neq i \\ j \in m^*}} a_{ji} + \sum_{k \in M^*} \sum_{i \in m^*} a_{ik} + \sum_{k \in M^*} \sum_{\substack{r \neq k \\ r \in M^*}} a_{rk} \right] \quad \text{(A4)}$$

But total help given by any group to itself, or the other group, is equal to the total help received by the recipient group. So, for example, total help given by the minority group to itself

$$\sum_{i \in m^*} \sum_{\substack{j \neq i \\ j \in m^*}} \alpha_{ij},$$

must equal total help received by this group from itself,

$$\sum_{i \in m^*} \sum_{\substack{j \neq i \\ j \in m^*}} a_{ji}$$

Since our hypothesis requires that the total volume of help exchanged is positive, equation (A4) generates a contradiction to the assumption that $\delta b < C_L < C_H$. Q.E.D.

Proof of Proposition 3. Equations (30) and (31) represent the solution to the equality versions of (26) and (27). Clearly, expression (31) is bigger than expression (30). Thus, a necessary and sufficient condition for r^* and g^* to be given by (30) and (31), respectively, is that expression (31) does not exceed the upper bound $\bar{a}$. Since (31) is increasing in δb, and goes to zero as δb approaches C_L, while on the other hand it goes to plus infinity as δb approaches C_H, it follows that there exists a cutoff value $(\delta^* b)$ for δb between C_H and C_L, where expression (31) equals $\bar{a}$. This establishes case (i).

Next, suppose δb exceeds $\delta^* b$, but is less than C_H, so g^* equals $\bar{a}$. Then r^* is either less than $\bar{a}$ [the case where the equality version of (27) intersects the line $g = \bar{a}$ at a value of r less than $\bar{a}$], or it attains the upper bound $\bar{a}$ (where this intersection occurs at a value of r not less than $\bar{a}$). In the former case, r^* is given by (32). Since (32) is increasing in δb, and approaches a number greater than $\bar{a}$ as δb approaches C_H, it follows that there exists an intermediate value $\delta^{**} b$ between $\delta^* b$ and C_H where it exactly equals $\bar{a}$. This is the dividing line between cases (ii) and (iii). Q.E.D.

REFERENCES

Abreu, Dilip. 1986. "Extremal Equilibria of Oligopolistic Supergames," 39 *Journal of Economic Theory* 191.

———. 1988. "On the Theory of Infinitely Repeated Games with Discounting," 56 *Econometrica* 383.

Abreu, Dilip, David Pearce, and Ennio Stacchetti. 1986. "Optimal Cartel Equilibria with Imperfect Monitoring," 39 *Journal of Economic Theory* 251.

Akerlof, George. 1976. "The Economics of Caste and the Rat Race and Other Woeful Tales," 90 *Quarterly Journal of Economics* 599.

Axelrod, Robert. 1981. "The Emergence of Cooperation Among Egoists," 75 *American Political Science Review* 306.

———. 1986. "An Evolutionary Approach to Norms," 80 *American Political Science Review* 1095.

Belshaw, Cyril. 1965. *Traditional Exchange and Modern Markets.* Englewood Cliffs, N.J.: Prentice-Hall.

Bendor, Jonathan, and Dilip Mookherjee. 1987. "Institutional Structure and the Logic of Ongoing Collective Action," 81 *American Political Science Review* 129.

Bernheim, B. D., and M. Whinston. 1987. "Multimarket Contact and Collusive Behavior." Harvard Institute of Economic Research Discussion Paper No. 1317, May.

Biddle, B. J. 1986. "Recent Developments in Role Theory," 12 *Annual Review of Sociology* 1.

Bull, C. 1985. "The Existence of Self-Enforcing Implicit Contracts." C. V. Starr Center, New York University.

Coleman, James. 1986. "Social Structure and the Emergence of Norms Among Rational Actors," in A. Dickmann and P. Mitter, eds., *Paradoxical Effects of Social Behavior.* Heidelberg: Physica-Verlag.

Cremer, Jacques. 1986. "Cooperation in Ongoing Organizations," 101 *Quarterly Journal of Economics* 33.

Dalton, Melville. 1959. *Men Who Manage.* New York: Wiley.

Ellickson, Robert C. 1987. "A Critique of Economic and Sociological Theories of Social Control," 16 *Journal of Legal Studies* 67.

———. 1989. "A Hypothesis of Wealth-Maximizing Norms: Evidence from the Whaling Industry," 5 *Journal of Law, Economics, and Organization* 83.

Farrell, Joseph, and Eric Maskin. 1987. "Renegotiation in Repeated Games." Harvard Institute of Economic Research Discussion Paper No. 1335, July.

Fenno, Richard. 1962. "The House Appropriations Committee as a Political System: The Problem of Integration," 56 *American Political Science Review* 310.

Foster, George M. 1972. "A Second Look at Limited Good," 45 *Anthropological Quarterly* 57.

Freidson, Eliot. 1984. "The Changing Nature of Professional Control," 10 *Annual Review of Sociology* 1.

Friedman, James. 1971. "A Non-Cooperative Equilibrium for Supergames," 38 *Review of Economic Studies* 1.

Green, Edward, and Rob Porter. 1984. "Noncooperative Collusion under Imperfect Price Information," 52 *Econometrica* 87.

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

Heclo, Hugh. 1977. *A Government of Strangers.* Washington, D.C.: The Brookings Institution.

Kissinger, Henry. 1979. *White House Years.* Boston: Little-Brown.

Kurz, Mordecai. 1977. "Altruistic Equilibrium," in B. Balassa and R. Nelson, eds., *Economic Progress, Private Values, and Public Policy.* Amsterdam: North Holland.

Lorenz, Edward H. 1988. "Neither Friends nor Strangers: Informal Networks of

Subcontracting in French Industry," in Diego Gambetta, ed., *Trust: Making and Breaking Cooperative Relations.* Oxford: Basil Blackwell Ltd.

Matthews, Donald. 1960. *U.S. Senators and Their World.* Chapel Hill: University of North Carolina Press.

McGinnis, Michael. 1986. "Issue Linkage and the Evolution of International Cooperation," 30 *Journal of Conflict Resolution* 141.

Milgrom, Paul, Douglass North, and Barry Weingast. Forthcoming. "The Role of Institutions in the Revival of Trade: Part I: The Medieval Law Merchant," *Economics and Politics.*

Mookherjee, Dilip. 1981. "Noncooperative Equilibria in Supergames with 'Almost Cooperative' Outcomes." Discussion Paper No. 81/37, ICERD, London School of Economics.

Netting, Robert. 1972. "Of Men and Meadows: Strategies of Alpine Land Use," 45 *Anthropological Quarterly* 132.

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

Oye, Kenneth. 1986. *Cooperation Under Anarchy.* Princeton: Princeton University Press.

Pearce, David. 1987. "Renegotiation-Proof Equilibria: Collective Rationality and Intertemporal Cooperation," mimeo, Department of Economics, Yale University.

Radner, Roy. 1986. "Repeated Partnership Games with Imperfect Monitoring and No Discounting," 53 *Review of Economic Studies* 43.

Roethlisberger, Fritz, and William Dickson. 1939. *Management and the Worker.* Cambridge: Harvard University Press.

Schmidt, Steffen W., James Scott, Carl Landé, and Laura Guasti, eds. 1977. *Friends, Followers, and Factions: A Reader in Political Clientelism.* Berkeley: University of California Press.

Selten, Reinhardt. 1975. "Re-examination of the Perfectness Concept for Equilibrium Points in Extensive Games," 4 *International Journal of Game Theory* 25.

Srinivas, M. N. 1952. *Religion and Society among the Coorgs of South India.* Oxford, Clarendon Press.

Taylor, Michael. 1976. *Anarchy and Cooperation.* New York: Wiley.