A Simple Model of Collective Action

RAJIV SETHI
Barnard College, Columbia University

E. SOMANATHAN
Indian Statistical Institute, Delhi

I. Introduction

This article outlines a theory of collective action in common property resource use. There is now a very large empirical literature on the commons, including numerous detailed case studies as well as a few econometric studies. To the best of our knowledge, however, there is no internally consistent model that broadly conforms to the facts that have emerged from this literature and that presents comparative static results on when collective action is likely to be successful. The theory outlined here is intended to fill this gap. It is based on the idea that at least some individuals involved in governance and extraction decisions are not motivated exclusively by material self-interest. Specifically, we allow for the possibility that a concern for reciprocity may be an important consideration in such environments. In this respect, our modeling approach is akin to that of Falk, Fehr, and Fischbacher (2002) but makes simpler assumptions and is explicit about the effects of different parameters on the prospects for successful collective action.

The empirical literature on the commons agrees on the importance of a number of factors that affect the likelihood of successful collective action. These include small user groups, a high level of dependence on the commons, low monitoring costs, and well-established schemes of punishment. "Beyond this apparently massive consensus there are however a number of important 'shadow zones'" (Baland and Platteau 1996, 289). Referring to the influence of group size and heterogeneity, Poteete and Ostrom (2004, 438) remark that "no consensus has emerged on the exact nature of the relationships or the relative importance of either factor." Agrawal (2003) complains that too many

We are grateful to the editor, two anonymous referees, Kaushik Basu, Kanchan Chopra, and participants of the workshop "Conversations between Economists and Anthropologists II" (Goa, August 1–3, 2003) for helpful comments on earlier versions of this article circulated under a different title. Contact the corresponding author, Rajiv Sethi, at rs328@columbia.edu.

 0013-0079/2006/5403-0010$10.00

factors that influence the success of collective action have been identified in the empirical literature, that the relations between different factors are not well understood, and that, in the absence of a model, it becomes impossible to sort out which factors really matter.

We argue that the disconnect between theory and empirical work on this problem stems from a formal modeling approach that does not capture one of the important stylized facts of commons management: the importance of a system for sanctioning those who violate agreements.[1] The model in this article transforms a social dilemma into a coordination game by allowing for the possibility that those not adhering to agreements for behaving cooperatively can be punished. The cost of carrying out the enforcement activity is borne by players whose preference for reciprocity induces them to punish violators of agreements even when it is costly to do so. In the model, cooperative behavior will prevail if the prospect of punishment is sufficient to deter those seeking gains from cheating, if the punishment cost is sufficiently low, and if the cost of communicating to coordinate on the efficient equilibrium is less than the expected surplus from collective action. An examination of these conditions then shows which parameters affect the prospects for cooperation and how they interact.

In Section II, we outline the most common modeling approach taken in the literature and provide some motivation for adopting the alternative taken in this article. Section III presents the basic symmetric model and the defining inequalities for the existence of an equilibrium with collective action. Section III.A presents the conditions for the existence of other equilibria, while Section III.B outlines our method for selecting among multiple equilibria when they exist. Section IV discusses the conditions for cooperation prominent in the empirical literature in the light of the model. Section V generalizes the basic model to allow for various kinds of heterogeneity among players and increasing returns, and Section VI concludes.

II. Review and Critique

Economic analyses of common property typically proceed under the hypothesis that extractors make independent choices with a view to maximizing their material well-being. In a static single-period model, each individual takes the others' actions as given and neglects the implications of their decisions on the payoffs of other extractors. The negative externality in the commons results

[1] The books by Ostrom (1990) and Baland and Platteau (1996) show that successful community management of forests, pastures, irrigation systems, and inland fisheries is widespread but far from universal and almost always accompanied by mechanisms for imposing sanctions on violators of agreements.

in a Nash equilibrium with suboptimal extraction levels from the perspective of the group as a whole (Gordon 1954; Dasgupta and Heal 1979).

To explain why cooperation is possible, time is incorporated into the model. The orthodox way to do this is to suppose that there is a future of infinitely many periods. In each period, the players play the game outlined in the previous paragraph. However, rather than maximizing current payoffs, they maximize a discounted sum of payoffs from the current and all future periods. The rationale is that the future matters less than the present because of impatience, the possibility that interest can be earned on resources converted into cash, or uncertainty that there will be a resource to exploit in the future. Now players take as given not just each others' actions in the current period but also the plans made by others for the infinite future. Each player's plan tells her what to do in each period in response to the entire history of play up to that point. The introduction of the future allows players to punish the other players for excessive exploitation by increasing their own future exploitation. The awareness that other players have such a contingent plan then deters players from harvesting more than their share of the efficient amount. An equilibrium with efficient extraction levels now exists, provided that players do not discount future payoffs too much. Moreover, such an equilibrium need not be based on "incredible" threats of punishment: threats that individuals would not find in their interest to carry out if called on to do so. In other words, there can exist an efficient equilibrium that satisfies the property of subgame perfection.

One difficulty with this modeling approach is that the subgame-perfect equilibrium described above is only one of infinitely many equilibria that exhibit different degrees of resource exploitation. For example, suppose that everyone adopts the following strategy: they extract an equal share of the Nash level from the static game in every period no matter what anyone else does. It follows immediately that no single player can gain by deviating unilaterally from his plan at any stage. This is also a subgame-perfect equilibrium, one in which the tragedy occurs in full force. Among other possible equilibria, some are quite outlandish. For example, it is an equilibrium for players to extract an equal share of the Nash level from the static game in every third period, while exercising restraint in other periods unless someone deviates from this rule, in which case everyone switches to the noncooperative behavior in every period.

Since different equilibria will change in different ways in response to changes in underlying parameters, the multiplicity of equilibria poses a problem for the exercise of comparative statics. As a result, comparative statics are sometimes performed on the set of equilibria or by focusing on a chosen equilibrium,

usually the best attainable for all players, as, for example, in Bendor and Mookherjee (1987). Unfortunately, the equilibrium set often contains equilibria whose outcomes are very different from each other, while focusing on the best attainable equilibrium requires further justification. We will provide such a justification in the model below, although it is not in the repeated-game framework.

Another problem with the repeated-game approach to explaining cooperation is that it is not robust to noise, for example, in the form of mistakes or experimentation by boundedly rational players. As long as such noise is not negligible, strategies of the kind described above will lead to frequent breakdowns and restarts of cooperation.[2] So far as we are aware, this kind of pattern has not been reported in the empirical literature on common pool resources. In fact, if, owing to setup costs, it is costly to start cooperating following a noncooperative phase, as is likely in many situations, this explains why attempts to cooperate on the basis of such strategies are not observed.

Economists often interpret equilibria in which exploitation is restrained by repeated-game strategies as "social norms." This interpretation appears somewhat strained. Social norms do not usually take the form of each person implicitly telling the others that if any of them does not conform to the norm, then neither will he. However, in our view, the most serious limitation of the model as a tool for analyzing cooperation in actual commons is that it does away with the need for governance. In fact, as noted above, and as the vast empirical literature on the commons has shown, successful commons management often or even usually has some institutions to support it (Ostrom 1990). These involve rules or norms, with fines or other punishments specified, often explicitly, for violations. If the shadow of the future were all that were needed to sustain cooperation, such institutions have no reason to exist. Moreover, the repeated-game approach has not addressed the kinds of questions raised by the reviewers of the empirical literature (Baland and Platteau 1996; Agrawal 2003; Poteete and Ostrom 2004) probably because the model is simply too cumbersome to do so effectively.

One alternative to the standard model is to allow for departures from explicitly optimizing behavior in favor of an evolutionary approach. In Sethi and Somanathan (1996), we postulated that the proportion of players playing different, possibly suboptimal, strategies would evolve over time under pressure of differential payoffs, with more highly rewarded strategies displacing less highly rewarded ones in the population. A critically important assump-

[2] See Bowles and Gintis (2005, chap. 4) for a detailed discussion of why repeated game models cannot explain cooperation in the presence of plausible noise and discount rates.

tion was that social punishments of some sort were available: players, at some cost to themselves, could punish other players who did not exercise restraint in harvesting. Under these circumstances, it was shown that a norm of restraint and punishment can be stable under evolutionary dynamics. Such norms can be destabilized, however, by parameter changes that make harvesting more lucrative, such as increases in the market price of the resource or improvements in harvesting technology. While this model gives a better fit to the facts of cooperation in the commons and allows for some interesting comparative statics, it is not tractable when generalized to asymmetrically situated players and is silent as to how a norm of restraint might evolve in the first place.

In this article, we outline a new model that attempts to address these problems. It seeks to specify fully the circumstances under which cooperation will be observed and departs from orthodox economic modeling in two ways. First, it assumes that players do not look far ahead into the future. This is a simplification made for tractability. Second, it relies on the presence of individuals who do not respond only to material payoffs. Economists have traditionally been reluctant to assume that people behave in ways that are not self-interested. The reason for this is that once such assumptions are allowed in the explanation of behavior, it becomes possible to explain virtually anything, but the explanations will often be vacuous since they end up assuming what they purport to explain. In the past few years, however, a new way of disciplining the behavioral assumptions made in modeling, a combination of evolutionary theory and experimental work, has become available.

The relevant departure from the characterization of people as being motivated solely by self-interest is the idea of reciprocity. Both gratitude and indignation are emotions that are felt in connection with reciprocity, the former being associated with what we may call "positive" reciprocity and the latter with "negative" reciprocity. Experiments with human subjects in the past few years have firmly established that many people display reciprocity that is not motivated by the prospect of future gains. Most relevant to us is the work that has been done with public goods games with punishment opportunities (surveyed in Fehr and Gächter 2000). In these games, subjects play a game in which members of a group each choose how much to contribute to a public good. The experimenter sets the payoffs so that contribution is privately costly but socially beneficial. After each round, players learn how much each of the other players contributed. Usually the others are identified only by numbers, so players never find out what another person actually played. Players then have the opportunity to punish others by lowering their payoffs at some cost

to themselves. It is found that even in the last round of such games, when players know there will be no further interaction, some players punish others and do so at considerable cost to themselves. Moreover, the presence of punishment opportunities increases contributions substantially. There have been many experiments by several researchers with variations on this theme in the past few years, and they all display these features.[3]

A natural question that one may ask is, why do players behave in this way? Why should preferences for reciprocity have evolved, when it may be costly to indulge such preferences? Sethi and Somanathan (2003) survey a number of mathematical models of how such evolution could have occurred. Essentially, these involve some combination of repetition, commitment, assortation, and parochialism. Here we mention only the basic idea behind the models that use parochialism. This is that people with preferences for reciprocity behave reciprocally with each other and selfishly when they meet people with selfish preferences. As long as people with selfish preferences cannot perfectly mimic those with reciprocal preferences, those with reciprocal preferences can get higher payoffs from cooperating with others like them, and this can more than outweigh their losses when they are fooled by selfish people pretending to be reciprocators.[4]

In fact, there is a good deal of evidence that people are heterogeneous. Some behave opportunistically, cooperating with others when it pays to do so and exploiting others when that is the most privately profitable strategy. Others are reciprocal, or sometimes even unconditionally altruistic. This heterogeneity is also predicted by many evolutionary models. In what follows, we take it as given that some people are "reciprocators," while others are opportunists, and we explore the implications for the commons of the interaction between these two preference types.

III. The Basic Model

There are n players, $i = 1, 2, \ldots, n$, each of whom has access to a common pool resource. We suppose that some mechanism to monitor resource extraction

[3] In addition to the considerable body of work surveyed by Fehr and Gächter (2000), subsequent papers include Bowles, Carpenter, and Gintis (2001), Carpenter and Matthews (2002), Fehr and Gächter (2002), Sefton, Shupp, and Walker (2002), Masclet et al. (2003), Page, Putterman, and Unel (2005), and Bochet, Page, and Putterman (forthcoming).

[4] The literature on the evolution of reciprocity is now vast; see, e.g., Panchanathan and Boyd (2003), Boyd et al. (2003), and the references cited in both works therein. Although we make no attempt here to endogenize the presence of reciprocators, our model could, in principle, be extended to do so. This is because collective action will be more likely to emerge in groups in which reciprocators have a significant presence, thus raising the average payoffs to reciprocators across a large population composed of many groups. We thank a referee for this observation.

from the common pool, to make rules if necessary, and to levy fines has been set up at some cost. This has, however, to be financed by ongoing contributions, which are observable and voluntary. A failure to contribute may result in punishment, but punishment is costly to impose, and the decision to punish is itself voluntary. Note that we model the punishment of noncontributions to the provision of the public good (in this case, governance) rather than the direct punishment of overextraction, as in Ostrom, Walker, and Gardner (1992) and Ostrom, Gardner, and Walker (1994).

For the time being, let us suppose that all players are identically situated in all respects (this assumption will be relaxed to allow for heterogeneity later). Player i can choose whether to contribute to the public good ($x_i = 1$) or not ($x_i = 0$). The aggregate contribution is denoted $X \equiv \sum_{j=1}^{n} x_j$. This aggregate contribution results in an aggregate benefit of αX, which is shared equally among all players (regardless of their contribution levels). Hence, the net benefit to player i arising from any vector $(x_1, \ldots, x_n)$ of contributions is simply $\alpha X/n - x_i$. It is assumed that

$$\alpha/n < 1 < \alpha, \tag{1}$$

as is standard in public goods environments. Hence, in the absence of punishment, it is individually rational for opportunists to choose not to contribute, although it is efficient for all to contribute.

After contributions have been observed by everyone, each player i can choose whether or not to participate in the collective punishment of all players j with $x_j = 0$. If i punishes, then $y_i = 1$, and if i does not punish, then $y_i = 0$. The total number of punishers, or enforcers, is therefore $e = \sum_{j=1}^{n} y_j$, and, provided that at least one person punishes, the total number of punished individuals is equal to the number of defectors $d = \sum_{j=1}^{n} (1 - x_j)$. Each player who is punished suffers a fixed penalty p regardless of the number of players participating in punishment. This penalty p may consist of partial or total exclusion from the public good or some other social sanction. Finally, the cost of punishing is proportional to the number of defectors (d) and inversely proportional to the number of enforcers (e), with the parameter γ affecting the size of this cost. The material payoff to player i is therefore given by

$$\pi_i(x, y) = \begin{cases} \alpha X/n - x_i & \text{if } e = 0, \\ \alpha X/n - x_i - (1 - x_i)p - \gamma y_i d/e & \text{if } e > 0, \end{cases} \tag{2}$$

where $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ are the vectors of contributions and punishments, respectively. The first term is i's share of the output αX from the public good. The second term is i's contribution, the third the punishment

p (nonzero only if i did not contribute), and the fourth the cost to i of punishing (nonzero only if $y_i = 1$).

We assume that there are two kinds of players, opportunists and reciprocators. There are $0 \leq k \leq n$ reciprocators. Opportunists maximize their material payoffs, and reciprocators maximize utility:

$$u_i(x, y) = \pi_i(x, y) + bx_i y_i.$$

Reciprocators therefore experience a "benefit" b if they have contributed and punished noncontributors. We may interpret this as the psychological satisfaction they get from relieving their feelings of anger at noncontributors. Note that reciprocators get no psychological satisfaction from punishing if they are themselves noncontributors, or from having contributed if they do not punish. This game is played every period, and it is assumed that players are myopic: they look only at the effect of their actions on current-period payoffs. While this is an extreme assumption, it makes the game simple and tractable and is no more implausible than the standard hypothesis that players can work out all the future consequences of their actions and those of others. Moreover, in a steady state, the equilibria we identify under myopic beliefs remain equilibria under forward-looking beliefs.[5]

Let $O \subset \{1, \ldots, n\}$ denote the set of opportunists (material payoff maximizers) and $R \subset \{1, \ldots, n\}$ the set of reciprocators. Myopia ensures that opportunists will never punish, and so $y_i = 0$ for all $i \in O$. By contrast, if a reciprocator punishes, then he must have contributed. That is, for any $i \in R$, if $y_i = 1$ then $x_i = 1$.

A strategy or plan for player i is of the form $[x_i, y_i(x)]$, where $y_i(x)$ is an indicator function of the vector of contributions x. For opportunists, $y_i(x)$ is the zero function. We examine the pure-strategy, subgame-perfect equilibria of this two-stage game in every period. There are two reasons for this choice. The first is that players playing in a context that is familiar are probably quite good at doing the necessary backward induction. Cosmides and Tooby (1992) present evidence that people are quite good at solving logical tasks in a social context that is familiar but are quite bad at solving logically equivalent problems presented in an unfamiliar context. Second, the best-response dynamics

[5] The future can be introduced by allowing p and b to depend on players' discount rates. Punishment can be harsher if there is a future, since exclusion from the public good can be prolonged for several periods. The expected surplus to each player from collective action $\alpha - 1$ can be assumed to be larger for lower discount rates, which in turn could raise b, the utility bonus from punishing a defector, since a player may be more emotionally involved with the collective effort when the surplus from it is expected to be high.

we use below converge rapidly to the subgame-perfect equilibria. We have the following three types of equilibria:

	$i \in O$	$i \in R$
A	Contribute	Contribute, punish if one person defects
B	Defect	Contribute and punish
C	Defect	Defect

In equilibria of type *C*, there is neither contribution nor punishment. In type *B* equilibria, opportunists do not contribute, while reciprocators contribute and punish. The equilibria of interest are those of type *A*, in which all contribute, and, if one person were to defect, all reciprocators punish the defector.[6] The conditions for such a subgame-perfect equilibrium to exist are that $k \geq 2$,

$$p \geq 1 - \frac{\alpha}{n}, \tag{3}$$

and

$$b \geq \gamma \frac{1}{(k-1)}. \tag{4}$$

The first of these inequalities states that the cost of being punished (weakly) exceeds the gain from defection. This ensures that all opportunists will cooperate, provided that they expect to be punished for defecting. The second inequality ensures that, in the event that a reciprocator were to defect, it will be an equilibrium (in the resulting subgame) for each of the remaining reciprocators to punish him. Note that we must have $k \geq 2$; otherwise, a deviating reciprocator could never be punished. Condition (4) implies that $b \geq \gamma/k$, which ensures that, in the event that an opportunist were to defect, it will be an equilibrium (in the resulting subgame) for each of the reciprocators to punish him. Conditions (3) and (4) together imply that a subgame perfect equilibrium with complete cooperation can be sustained. We discuss their implications and generalizations in Sections IV and V. First, however, we describe the other possible equilibria in Section III.A and our method of obtaining a unique prediction when multiple equilibria exist in Section III.B. Readers interested only in the practical implications of the model may skim over these next two subsections.

[6] If multiple individuals defect simultaneously, subgame perfection requires that reciprocators will participate in punishment if and only if this is consistent with utility maximization. We need only consider unilateral deviations, however, in establishing that a particular strategy profile is an equilibrium.

A. Equilibria with Less than Complete Cooperation

If $k = 0$, then the unique subgame-perfect equilibrium is of type C. If $1 \leq k \leq n - 1$, by contrast, multiple equilibria may exist. Subgame-perfect equilibria of type C with no contributions and no punishments will exist if no reciprocator can gain by unilaterally deviating from a strategy of not contributing to one of contributing and punishing all noncontributors. The payoff at the equilibrium is 0, while the payoff from a deviation of this type is $\alpha/n - 1 + b - \gamma(n - 1)$. Hence, an equilibrium of type C will exist if

$$1 - \frac{\alpha}{n} \geq b - \gamma(n - 1). \tag{5}$$

Next, consider subgame-perfect equilibria of type B, in which opportunists do not contribute and reciprocators contribute and punish opportunists. A necessary condition for such equilibria to exist is that no opportunist can gain by unilaterally switching to cooperation (and thus escaping punishment). Unilaterally switching to cooperation raises by one the total number of cooperators and, hence, causes the provision of the public good to rise by an amount α, which raises each individual's payoff by α/n. In addition, the deviating individual escapes the punishment cost p and incurs the cost of contribution, which is one. In order for such a deviation to be unprofitable, we must therefore have

$$p \leq 1 - \frac{\alpha}{n}. \tag{6}$$

This ensures that the threat of punishment does not deter opportunists from defecting. In addition, we require that reciprocators have an incentive to cooperate and punish. A sufficient condition for this is the following, which guarantees that a reciprocator would not gain from switching to defection (meaning neither contributing nor punishing noncontributors) even if doing so did not result in punishment from remaining reciprocators:

$$b - \gamma\frac{(n - k)}{k} \geq 1 - \frac{\alpha}{n}. \tag{7}$$

This condition is not, however, necessary. Equilibria of type B can also arise if reciprocators believe that switching to defection will result in punishment, and if this belief is warranted given the strategies of other reciprocators. This requires that the following two conditions hold:

$$b \geq \gamma\left(\frac{n - k + 1}{k - 1}\right), \text{ and } b - \gamma\left(\frac{n - k}{k}\right) \geq 1 - \frac{\alpha}{n} - p. \tag{8}$$

The first inequality ensures that all reciprocators who do not defect have an incentive to punish the one reciprocator who does, provided that they all believe that every nondefecting reciprocator will participate in punishment. The second ensures that a reciprocator will not defect under the belief that he will be punished for doing so. Conditions (1) and (7) together imply that

$$b \geq \gamma\left(\frac{n-k}{k}\right),$$

which ensures that a reciprocator will not free ride on punishment (while continuing to contribute). This is also implied by the first inequality in (8). Hence (6), together with either (7) or (8), is necessary and sufficient for a subgame-perfect equilibrium of type B to exist.[7]

We have so far neglected the case $k = n$. Here equilibria of types A and B are identical and will exist if and only if (3) and (4) hold. Except for the nongeneric case of $p = 1 - \alpha/n$, equilibria of types A and B cannot coexist if $k < n$. For $k < n$, when $p > 1 - \alpha/n$, complete compliance with the norm of contribution is possible, but when $p < 1 - \alpha/n$, only partial compliance is possible.

These inequalities completely describe when each of the three types of equilibria will exist. They exhaust all generic possibilities for subgame-perfect equilibria, since equilibria must be intragroup symmetric. That is to say, in any equilibrium, since the incentives facing a reciprocator are the same as those facing any other reciprocator, they must take the same action at any stage of the game. This is, of course, true for opportunists as well.

B. Prediction

Since the model generally permits multiple equilibria, this raises the question of which equilibrium we might expect to prevail in practice. We need to identify conditions under which equilibria of type A will be chosen when these coexist with those of type C, and to perform a similar analysis for the case when types B and C coexist. We deal with the coexistence of types A and C first.

It may be that (4) holds so that all contributing reciprocators will punish a lone defector, but

$$b < \gamma(n-1), \tag{9}$$

so that a reciprocator will not punish if everyone else defects. The latter

[7] When $k = 1$, (7) is inconsistent with (5) for generic parameter values so equilibria of types B and C cannot coexist.

condition implies (5) so an equilibrium of type C exists in this case. If, in addition, punishment is strong enough to deter would-be defectors, that is, if (3) holds, then equilibria of type A will exist as well. Notice that the equilibrium payoff to all players under type A is α, which exceeds one, the payoff from type C. This raises the possibility that communication among players at the start of each period can allow them to coordinate on the preferred equilibrium.

Suppose that the cost to any player of communicating with the other players is c, and

$$c < \alpha - 1. \tag{10}$$

Therefore, players will soon realize that they are better off agreeing to play type A as long as (10) holds. This alone does not solve the problem of equilibrium selection, since it is also an equilibrium for all commitments to be ignored and for all players to defect. However, such "babbling equilibria" are not observed in the everyday experience of coordination problems with preplay discussion. Experiments on coordination games with two or more players confirm that costless preplay communication enables players to coordinate on the Pareto-dominant equilibrium (Cooper et al. 1992; Charness 2000; Burton, Loomes, and Sefton 2005; Blume and Ortmann, forthcoming) even when failure to coordinate involves a considerable payoff loss for those attempting to coordinate and even though the communication permitted in the experiments is extremely sparse. Babbling equilibria seem especially unlikely if communication is at all costly, since in this case only players intending to honor their commitments will bother to make them.[8] For these reasons, we assume that players will coordinate on the type A equilibrium when it exists.[9]

We can use (7) and similar reasoning to show that if the parameters are such that both type B and C equilibria exist, and if the cost of reciprocators communicating with each other is positive but sufficiently small, then we may expect to see only type B equilibria in the long run.

It is worth remarking that this setup allows for noise in the sense that if players make mistakes or experiment with new actions now and then, this

[8] This method of equilibrium selection amounts to what has been called "forward induction." See Osborne and Rubinstein (1994, 110–15) for a discussion and references to the originators of the concept. This, too, has been confirmed experimentally by Van Huyck, Battalio, and Beil (1993).

[9] A similar solution to the equilibrium selection problem could be used in a repeated-game model with self-interested players in which each period has a punishment stage following the contribution stage. This could be robust to noise. However, subgame perfection of efficient equilibria would require strategies that involved an infinite regress of punishments: players who did not punish would need to be punished, and so on. There are also evolutionary approaches to equilibrium selection in repeated games; see especially Fudenberg and Maskin (1990) and Binmore and Samuelson (1992).

will not generally result in a change from type A to type C equilibria, unless there happen to be simultaneous mistakes by several players. This is a consequence of the static nature of the game, together with subgame perfection. Moreover, if there is such a collapse of cooperation, cooperation may be recovered if the communication cost is sufficiently low. Thus, we would expect cooperation, if it comes into being, to be persistent, although perhaps subject to occasional random crashes.

We have assumed throughout that the number of reciprocators k and the strength of reciprocity as measured by b are known. However, even if they are not, and provided that the players' expected values of these parameters (as well as the true values) are sufficiently high, the arguments made above remain valid.

IV. Conditions for Cooperation

The conditions for cooperation to take place in this setup (in addition to $k \geq 2$) are (3) and (4) together with (10). The first of these (3) says that punishment must be a sufficient deterrent against noncontribution, the second (4) says that punishment must be cheap enough to inflict that the psychological benefit of doing so is enough to induce reciprocators to carry it out, and the third says that the surplus to each player from collective action should exceed the per-player cost of communication between players that enables them to coordinate collective action. These are all intuitive. Let us compare them with the conditions for cooperation that have emerged from the commons literature.

Agrawal (2003, 253, table 2) provides a comprehensive list of enabling conditions for successful collective action from three influential books by Wade (1988), Ostrom (1990), and Baland and Platteau (1996), as well as other conditions prominent in the literature. Agrawal classifies them as pertaining to resource characteristics, group characteristics, the relation between group and resource characteristics, institutional arrangements, or the external environment. Those pertaining to resource characteristics are small size (of the resource), well-defined boundaries, low mobility, storage possibilities, and predictability. In our model, these factors would matter because they affect the potential surplus $\alpha - 1$ known to be available from collective action.

Group characteristics favoring collective action listed by Agrawal (2003) are (i) small size; (ii) clearly defined boundaries; (iii) shared norms; (iv) past successful experiences–social capital; (v) appropriate leadership—young, familiar with changing external environments, and connected to local traditional elite; (vi) interdependence among group members; (vii) heterogeneity of endowments, homogeneity of identities and interests; and (viii) low levels of poverty. Of these, we note that shared norms and past successful experiences

emerge as a consequence of successful collective action, not a prerequisite. However, if there has been successful collective action in one domain, then the cost of communicating to reach agreement over collective action in another domain may be lowered because of the presence of norms that are transferable from one sphere to another. Clearly defined boundaries for the group are also endogenous: collective action will wholly succeed only if noncontributors can be punished and the set of contributors defines the group.

In our framework, small group size can influence collective action in several ways. It results in a higher private return to collective action, thus reducing the size of the punishment needed to support it (see [3]). It could also be the case that punishment itself is facilitated by small group size since social sanctions will be an effective deterrent only in small groups that are not anonymous. The cost of communication is also likely to be lower for small groups. By contrast, the cost of punishing someone who violated an agreement could be higher in very small groups (because incurring the enmity of someone in a small community may be very costly), which would make collective action more difficult to sustain (see [4]). Increasing returns (discussed explicitly together with heterogeneity and poverty in Sec. V) could mean that small groups may not generate enough surplus to make collective action worthwhile.[10]

Leaders may be those with low costs of communication. Otherwise, our model suggests that leadership is not important. Interdependence among group members may increase the effectiveness of social sanctions but, by the same token, also make punishment more costly because of more effective retaliation. Poverty could constrain contributions and thus render collective action impossible. The inequality in the power to punish that may accompany poverty has effects that are discussed in Section V.

Regarding the interaction between group and resource characteristics, Agrawal (2003) lists the following enabling conditions for sustainable use: (i) overlap between user group residential location and resource location, (ii) high levels of dependence by group members on the resource system, (iii) fairness in allocation of benefits from common resources, (iv) low levels of user demand, and (v) gradual change in levels of demand. The first two are simply indicators of potential surplus α. The third is a possible outcome of collective action, not a contributor to its success. The fourth and fifth are irrelevant in our model unless they affect α.

Under the heading "institutional arrangements," Agrawal (2003) mentions

[10] Although we have not explicitly modeled errors (in either the actions taken by individuals or in the perceptions of actions taken by others), the greater possibility of multiple simultaneous errors in larger groups can also adversely affect the possibility of sustained cooperation.

simple rules, locally devised access and management rules, ease in enforcement of rules, graduated sanctions, the availability of low-cost adjudication, and the accountability of monitors and other officials to users. Our model indicates that these are all outcomes of collective actions, not prerequisites, with the possible exception of external adjudication, which may be exogenous. Finally, under "external environment," Agrawal mentions a low-cost exclusion technology, central governments that do not undermine local authority, supportive external sanctioning institutions, appropriate levels of external aid to compensate local users for conservation activities, and nested levels of appropriation, provision, enforcement, and governance. It is clear that the last is endogenous, while supportive external sanctioning institutions are likely to enlarge the expected value of p.

To summarize: our review of the factors listed in the literature as important for determining the success of collective action indicates that many of them are, in fact, endogenous; they are consequences, not determinants, of collective action. In fact, our model suggests that the success of collective action in using common property effectively hinges on whether it is feasible to develop a regulatory mechanism for the commons that ensures a sufficiently large surplus relative to unregulated exploitation, whether the cost of communication to establish and maintain this institution is sufficiently low, and whether it is feasible to punish noncontributors enough to deter them at sufficiently low cost. Far from there being "too many variables" (Agrawal 2003), there are, in fact, quite few. We have, of course, to qualify this by saying that, as seen above, many factors may affect the parameters α, c, p, γ, and b. But it is much easier to study them if it is understood that their influence is through a small set of parameters and if it is clear how they may be expected to matter.

Finally, it is interesting to note that a cost of communication smaller than the expected surplus (eq. [10]) is not included in Agrawal's (2003) list, which is compiled from field studies of the commons. The possibility of punishment in our model results in a coordination game. The experimental literature clearly shows that communication is critical to whether or not coordination is achieved.[11] We suspect that communication has not received attention despite the experimental literature because it has no role to play in the repeated-game

[11] In fact, even in common-pool resource extraction games in which there is a unique and inefficient equilibrium involving overextraction, communication has been found to be effective in raising rents close to the maximum in experimental situations, even in the absence of punishment (Ostrom et al. 1994) and even with heterogeneous extractors (Hackett, Schlager, and Walker 1994). The almost ubiquitous presence of punishment mechanisms in the field, however, suggests that communication without punishment would not be sufficient to stabilize cooperation in settings in which interaction takes place for much longer periods than in the laboratory.

models that are, almost exclusively, the only formal theories of cooperation in social dilemmas.[12]

We point to some more underlying variables that can be expected to influence some of the model parameters. With regard to α, not only must the return to cooperation be high, but it must be known to be high by all concerned, or by at least some who are in a position credibly to communicate this information to the others at low enough cost. Reciprocators' utility bonus from punishment b will be sufficiently positive only if their emotions are engaged, which in turn is likely if the public good provides them with sufficient surplus.

The punishment p has to be effective. Effective punishment will vary from case to case, but the most likely punishment is exclusion from the commons. Whether it is technologically and socially feasible may be critical. It will be weak if individuals expect to leave the area soon, so short time horizons and a high probability of migration are not conducive to cooperation. A dense network of social interaction may also favor punishment as exclusion can then be used in the domain in which it is cheapest.

From the point of view of empirical testing, it is important to note that the conditions for cooperation are given by inequalities. It follows that cooperation varies discontinuously with the parameters. Changes in the parameters that are not large enough to reverse any of the inequalities will have no effect. This general point applies to the discussion in Section V as well.

V. Heterogeneity and Other Generalizations

Consider as a benchmark the homogeneous player case in which (3) and (4) hold so that the equilibrium of type A prevails. Now suppose that instead of punishment resulting in a uniform loss p, the effects of punishment vary across players. The material payoffs (2) may now be written as

$$\pi_i(x, y) = \begin{cases} s_i\alpha(X) - x_i & \text{if } e = 0 \\ s_i\alpha(X) - x_i - (1 - x_i)p_i - \gamma y_i \dfrac{d}{e} & \text{if } e > 0, \end{cases} \tag{11}$$

where p_i is the cost to player i of being punished, and we are now allowing for a (possibly) nonlinear production function $\alpha(X)$, which describes the output obtained as a function of total contributions. As before, we initially fix the share s_i accruing to player i at $1/n$.

Suppose for simplicity that there are now just two possible values of p,

[12] Explicit punishments are also unnecessary in repeated-game models of cooperation in social dilemmas, but, unlike the ease of communication, they cannot be easily overlooked in the field.

namely, p_l and p_h, and that (3) holds for p_h but not for $p_l < p_h$. Suppose that $p_i = p_l$ for n_l players and k_l reciprocators and that $p_i = p_h$ for the remaining $n - n_l$ players and $k - k_l$ reciprocators. Those players i with $p_i = p_l$ are little affected by punishment and will find it optimal to defect since

$$p_l < 1 - \frac{\alpha(X)}{n}. \tag{12}$$

In the period following this, there may be too few reciprocators who have contributed to punish the players with $p_i = p_h$ at reasonable cost, that is,

$$b < \gamma\left(\frac{1}{k - k_l - 1}\right).$$

Furthermore, the free riding of some players will lower the returns to the others, possibly making it not worthwhile for them to contribute even if they were to be punished, that is,

$$p_h < 1 - \frac{\alpha(X_h)}{n},$$

where $X_h = \sum_{j=1}^{n-n_l} x_j$ denotes aggregate contributions by those players i with $p_i = p_h$. Notice that this inequality is more likely to hold if the production function $\alpha(\cdot)$ displays increasing returns. Thus, heterogeneity in susceptibility to punishment, especially in combination with increasing returns, may lead to collective action becoming infeasible.

The model so far fixed both the shares of the public good accruing to each player and the contributions. However, if side payments are possible or, equivalently, contributions can be varied continuously so that the distribution of the surplus $\alpha(X) - X$ from the public good may be changed (within limits) without affecting the total surplus, then it becomes easier to achieve cooperation. This would be the case, for example, if the production function $\alpha(\cdot)$ were such that there exists a surplus-maximizing total contribution X^*, players are not wealth constrained, and punishment is a sufficient deterrent on average. A precise statement of this fact is in the appendix.

In this case, the players may, after discussion and bargaining, agree on a vector of contributions leading to a total contribution of X^*, and that ensures that the necessary inequalities for successful collective action hold. Heterogeneity in susceptibility to punishment can be taken into account in the division of the surplus by giving players with less susceptibility to punishment larger shares of the surplus, while still leaving all players with a share of the surplus large enough to motivate them to incur the cost of enforcement when

necessary. This may be exactly what is happening when elites take the initiative to organize collective action as, for example, in Wade (1988). Heterogeneity, at least within limits, is not as inimical to collective action as one might think. Poteete and Ostrom's (2004) survey of collective action in forest management provides some empirical support for this view.

Irrigation systems and fisheries have heterogeneous returns to collective action because of heterogeneity of location and landholding in the former case and of skills and capital equipment in the latter. As outlined above and in the appendix, contributions can be tailored to take such heterogeneity into account. Those with low returns from collective action can be asked for correspondingly lower contributions. This suggests that it may not be heterogeneity of returns but, rather, other factors that explain why collective action so often seems absent in these contexts.[13] When irrigators are from different villages and fishermen from different ports, high communication costs and the inability to punish may explain the failure of collective action.

We have discussed heterogeneity of power and the returns to collective action, but the literature on heterogeneity has mostly focused on other dimensions, with wealth being the most important. It will, however, be immediately obvious to the reader that the model suggests that an important reason why wealth and poverty might matter is because they result in heterogeneity of power. Wealthy people, particularly in small communities, are typically less vulnerable to social sanctions than poor people and more able to impose penalties on others. Wealth has probably received more attention than power in the literature simply because it is easier to measure. Of course, as mentioned earlier, it also matters because it can constrain contributions. This is the aspect of wealth that has received theoretical attention (Baland and Platteau 1997, 1998).

Heterogeneity of power may actually favor collective action in some circumstances. To see this, let us allow the punishments p_i to depend on the entire vector y_{-i} (so that the effect of punishment on i depends on the number and identity of the particular individuals who choose to punish i). As a bench-

[13] As pointed out by Baland and Platteau (1998), there may be asymmetric information about the returns to collective action, and this could hinder transfers to compensate low returns. This problem should not be exaggerated, however. In small communities much of the relevant information is easily found out. Even when it is not, good design could reduce the extent of the problem. For example, if equal saleable input quotas are allocated to fishers to increase the catch to effort ratio in a fishery, then low-skill fishermen who may have a lower return from the quota could raise their return by selling to higher-skill fishermen.

mark, suppose, first, that the parameters are such that punishment is not an effective deterrent even when all individuals punish. That is, for all players i,

$$p_i(1, \ldots, 1) < 1 - \frac{\alpha}{n}.$$

Now suppose, instead, that there exists a group of powerful persons I, who can effectively punish the others J (but not each other), if at least one of the others take part in enforcement, say, by monitoring defection. We need both the powerful and the weak for enforcement. Otherwise, if the weak were not needed, the powerful would be able to coerce the weak and leave them worse off. The powerful need to be given shares large enough so that the private returns to contribution for them are high enough to induce them to participate, even though they cannot be punished. Suppose that for all $j \in J$,

$$p_j(y_{-j}) > 1 - s_j\alpha$$

if at least one component y_i of y_{-j} is 1 for some $i \in I$, and at least one component y_l of y_{-j} is one for some $l \in J$. Suppose also that the cost of punishment depends on the identity of the punishers so that if at least one member from each group punishes, then the punishment cost is less than b. Finally, suppose that for all $i \in I$,

$$s_i\alpha > 1.$$

Now all the inequalities necessary for a type A equilibrium are in place, provided that the communication cost is $c < 1 - \alpha$. For this to be a Pareto improvement over a situation with no contributions, it is necessary that $s_j\alpha n > 1$. Clearly, there are many configurations of the parameters such that these inequalities hold. However, if the weak did not have something to offer, for example, by way of help in monitoring, then it is unlikely that the powerful would allocate a share to them that would make them better off than they would have been under the unregulated outcome.

It is often observed that elites take the lead in the management of common property resources and appropriate the lion's share of the benefits. As Baland and Platteau (1998) point out, this is not always a Pareto improvement over an unregulated outcome because the poor may be worse off. Whether or not this actually occurs has to be assessed on a case-by-case basis.

VI. Conclusion

We hope that the model presented here will prove useful as a framework for empirical research into the issue of when collective action in the commons

will be successful. It can be adapted to particular situations by suitable modeling of the production function, punishment technology, and so forth.

What policy implications can we draw from this theory? If outside intervention to help spur collective action in the commons is to be successful, it has to ensure that enforcement of contributions (or other nondefection) is both effective and cheap. Lowering the cost of communication about such issues may be the role that outside agencies can play. They may do so by helping participants see that collective action has been successful in similar circumstances elsewhere, or simply by initiating and facilitating the process of discussion on the issue. They may need to provide information about the benefits of collective action in cases where this is not clear to the participants. Of course, this will only work if the underlying conditions are favorable. This is less likely when players are transient, so that exclusion has little force; when exclusion is not possible for some reason; or when there is a set of powerful players who cannot be punished and whose private returns cannot be made high enough to make it attractive for them to participate. Legal reforms that allow for either community enforcement or the state to lend force to community enforcement may be called for in some cases. Care needs to be taken, of course, to see that this does not result in an expropriation of the poor. Insisting that the process of legal change requires the consultation and consent of all groups would make this less likely.

We have not addressed some potentially important issues. What factors make it likely that bargaining over the division of the surplus will end in agreement? How does history affect the proportion of reciprocators, or does it not? We leave these interesting but challenging questions to future research.

Appendix

In this appendix we show that heterogeneity in the returns to collective action (as captured by the shares s_i of the surplus from the public good) and in susceptibility to punishment do not change the prediction of an efficient outcome if players' communication costs are sufficiently low, transfers unconstrained by wealth are possible, and punishments are effective on average. Note that using the general payoff function (11) and allowing for heterogeneous susceptibility to punishment, the conditions (3) and (10) for an efficient type *A* equilibrium to prevail are replaced by

$$s_i \Delta\alpha + t_i \geq \frac{X^*}{n} - p_i, \text{ for all } i = 1, \ldots, n, \tag{A1}$$

$$\sum_{i=1}^{n} t_i = 0, \tag{A2}$$

and

$$c_i < s_i \alpha(X^*) - \frac{X^*}{n} + t_i \text{ for all } i = 1, \ldots, n, \tag{A3}$$

where t_i is a transfer received by player i out of the total contributions, $\Delta\alpha = \alpha(X^*) - \alpha(X^* - X^*/n)$ is the increase in surplus that is generated when all players rather than all but one player contribute, and c_i is player i's cost of communicating with the other players. Equation (A1) says that the (possibly negative) increase in a player's payoff when he contributes rather than defects plus the (possibly negative) transfer he receives conditional on contributing should be at least as much as his contribution less the damage from being punished. The condition (4) that requires that punishing not be too costly is unchanged. If, on average, punishment is a sufficient deterrent, meaning that

$$\bar{p} \geq \frac{X^*}{n} - \frac{1}{n}\Delta\alpha, \tag{A4}$$

where $\bar{p}$ denotes the average of p_i, then it is straightforward to check that if we set

$$t_i = \left(\frac{1}{n} - s_i\right)\Delta\alpha + \bar{p} - p_i,$$

then (A1) and (A2) are satisfied. Of course, for the efficient outcome to prevail, we also require (A3) to hold. Note that transfers tend to make up for inequality in returns.

References

Agrawal, Arun. 2003. "Sustainable Governance of Common-Pool Resources: Context, Methods, and Politics." *Annual Review of Anthropology* 32:243–62.

Baland, Jean-Marie, and Jean-Philippe Platteau. 1996. *Halting Degradation of Natural Resources: Is There a Role for Rural Communities?* Oxford: Oxford University Press.

———. 1997. "Wealth Inequality and Efficiency in the Commons, Part I: The Unregulated Case." *Oxford Economic Papers* 49, no. 4:451–82.

———. 1998. "Wealth Inequality and Efficiency in the Commons, Part II: The Regulated Case." *Oxford Economic Papers* 50, no. 1:1–22.

Bendor, Jonathan, and Dilip Mookherjee. 1987. "Institutional Structure and the Logic of Ongoing Collective Action." *American Political Science Review* 81:129–54.

Binmore, Kenneth G., and Larry Samuelson. 1992. "Evolutionary Stability in Repeated Games Played by Finite Automata." *Journal of Economic Theory* 57:278–305.

Blume, Andreas, and Andreas Ortmann. Forthcoming. "The Effects of Costless Preplay Communication: Experimental Evidence from a Game with Pareto-Ranked Equilibria." *Journal of Economic Theory*.

Bochet, Oliver, Talbot Page, and Louis Putterman. Forthcoming. "Communication and Punishment in Voluntary Contribution Experiments." *Journal of Economic Behavior and Organization*.

Bowles, Samuel, Jeffrey Carpenter, and Herbert Gintis. 2001. "Mutual Monitoring in Teams: Theory and Evidence on the Importance of Residual Claimancy and Reciprocity." Unpublished manuscript, Department of Economics, Middlebury College.

Bowles, Samuel, and Herbert Gintis. 2005. *A Cooperative Species: Human Reciprocity and Its Evolution*. Unpublished manuscript, Santa Fe Institute, Santa Fe, NM.

Boyd, Robert, Herbert Gintis, Samuel Bowles, and Peter Richerson. 2003. "The Evolution of Altruistic Punishment." *Proceedings of the National Academy of Science* 20:123–43.

Burton, Anthony, Graham Loomes, and Martin Sefton. 2005. "Communication and Efficiency in Coordination Game Experiments." In *Experimental and Behavioral Economics*, ed. John Morgan, 63–85. Amsterdam: Elsevier.

Carpenter, Jeffrey, and Peter Matthews. 2002. "Reciprocity." Working Paper no. 29, Department of Economics, Middlebury College.

Charness, Gary. 2000. "Self-Serving Cheap Talk: A Test of Aumann's Conjecture." *Games and Economic Behavior* 33:177–94.

Cooper, Russell, Douglas V. DeJong, Robert Forsythe, and Thomas W. Ross. 1992. "Communication in Coordination Games." *Quarterly Journal of Economics* 53: 739–71.

Cosmides, Leda, and John Tooby. 1992. "Cognitive Adaptations for Social Exchange." In *The Adapted Mind*, ed. John H. Barkow, Leda Cosmides, and John Tooby, 163–228. New York: Oxford University Press.

Dasgupta, Partha, and Geoffrey M. Heal. 1979. *Economic Theory and Exhaustible Resources*. Cambridge: Cambridge University Press.

Falk, Armin, Ernst Fehr, and Urs Fischbacher. 2002. "Appropriating the Commons: A Theoretical Explanation." In *The Drama of the Commons*, ed. Elinor Ostrom, Thomas Dietz, Nives Dolsak, Paul C. Stern, Susan Stonich, and Elke U. Weber, 157–91. Washington, DC: National Academy Press.

Fehr, Ernst, and Simon Gächter. 2000. "The Economics of Reciprocity." *Journal of Economic Perspectives* 14:151–69.

———. 2002. "Altruistic Punishment in Humans." *Nature* 415:137–40.

Fudenberg, Drew, and Eric S. Maskin. 1990. "Evolution and Cooperation in Noisy Repeated Games." *American Economic Review Papers and Proceedings* 80:274–79.

Gordon, H. Scott. 1954. "The Economic Theory of a Common Property Resource: The Fishery." *Journal of Political Economy* 62:124–42.

Hackett, Steven, Edella Schlager, and James Walker. 1994. "The Role of Communication in Resolving Commons Dilemmas: Experimental Evidence with Heterogeneous Appropriators." *Journal of Environmental Economics and Management* 27: 99–126.

Masclet, David, Charles Noussair, Steven Tucker, and Marie-Claire Villeval. 2003.

"Monetary and Nonmonetary Punishment in a Voluntary Contributions Mechanism." *American Economic Review* 93, no. 1:366–80.

Osborne, Martin J., and Ariel Rubinstein. 1994. *A Course in Game Theory*. Cambridge, MA: MIT Press.

Ostrom, Elinor. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge: Cambridge University Press.

Ostrom, Elinor, Roy Gardner, and James Walker. 1994. *Rules, Games, and Common Pool Resources*. Ann Arbor: University of Michigan Press.

Ostrom, Elinor, James Walker, and Roy Gardner. 1992. "Covenants with and without a Sword: Self-Governance Is Possible." *American Political Science Review* 86:404–17.

Page, Talbot, Louis Putterman, and Bulent Unel. 2005. "Voluntary Association in Public Goods Experiments: Reciprocity, Mimicry, and Efficiency." *Economic Journal* 115:1032–53.

Panchanathan, Karthik, and Robert Boyd. 2003. "A Tale of Two Defectors: The Importance of Standing for Evolution of Indirect Reciprocity." *Journal of Theoretical Biology* 224:115–26.

Poteete, Amy R., and Elinor Ostrom. 2004. "Heterogeneity, Group Size and Collective Action: The Role of Institutions in Forest Management." *Development and Change* 35, no. 3:435–61.

Sefton, Martin, Robert Shupp, and James Walker. 2002. "The Effects of Rewards and Sanctions in Provision of Public Goods." CeDEx Discussion paper 2002-02, University of Nottingham.

Sethi, Rajiv, and E. Somanathan. 1996. "The Evolution of Social Norms in Common Property Resource Use." *American Economic Review* 86:766–88.

———. 2003. "Understanding Reciprocity." *Journal of Economic Behavior and Organization* 50:1–27.

Van Huyck, John B., Raymond C. Battalio, and Richard O. Beil. 1993. "Asset Markets as an Equilibrium Selection Mechanism: Coordination Failure, Game Form Auctions, and Forward Induction." *Games and Economic Behavior* 5:485–504.

Wade, Robert. 1988. *Village Republics: Economic Conditions for Collective Action in South India*. Cambridge: Cambridge University Press.