Discussion Papers in Economics

Collective Action in the Commons: A Theoretical Framework for Empirical Research

Rajiv Sethi

E. Somanathan

Discussion Paper 04-21

June 2004

Indian Statistical Institute, Delhi
Planning Unit
7 S.J.S. Sansanwal Marg, New Delhi 110 016, India

Collective Action in the Commons: A Theoretical Framework for Empirical Research*

Rajiv Sethi† E. Somanathan‡

Revised June 2004

Abstract

A model of collective action in the commons that is intended to provide a framework for empirical research into the question of when cooperation is likely to be successful is presented. It is based on the presence of costly punishment opportunities, some players who have a taste for punishing those who violate agreements to cooperate (an assumption strongly supported by recent experimental research), and bounded rationality. It predicts that cooperation is more likely when communication is cheap, the technology of public good provision is sufficiently productive, effective punishment opportunities are available at sufficiently low cost, and when group size is large (holding constant the other parameters mentioned). Heterogeneity in the ability to inflict punishment or be hurt by it may result in collective action becoming infeasible, especially when there are increasing returns to the public good, but there is a range of parameters in which changes in heterogeneity will have no effect and circumstances in which heterogeneity will actually favor cooperation.

*We are grateful to Kaushik Basu, Kanchan Chopra and participants of the workshop "Conversations between economists and anthropologists-II", Goa, August 1-3, 2003, for helpful comments on an earlier version of this paper.

†Department of Economics, Barnard College, Columbia University (rs328@columbia.edu).

‡Planning Unit, Indian Statistical Institute - Delhi (som@isid.ac.in).

1 Introduction

This paper outlines a theory of collective action in common property resource use that is intended to predict the circumstances under which such action will be successful. There is now a very large empirical literature on common property resource use, mostly case studies as well as some econometric studies. However, to the best of our knowledge there is no internally consistent model that broadly conforms to the facts that have emerged from the case study 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 and is presented in the hope that it will be of use to empirical researchers studying common property resource use who may adapt it to fit their problem. It is based on the idea that at least some individuals involved in 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.

Economic analyses of common property typically proceed under the hypothesis that extractors make independent choices with a view to maximizing their material well-being. Since each individual neglects the implications of their decisions on the payoffs of other extractors, this results in suboptimal extraction levels from the perspective of the group as a whole. This effect is clearly illustrated in the following simple, static model of the commons, based on the work of Gordon (1954) and Dasgupta and Heal (1979). Consider a group of individuals with shared access to a resource which is valuable but costly to appropriate. Each appropriator makes an independent choice regarding his level of resource extraction. The aggregate amount of extraction is simply the sum of all individual extraction levels. The total cost of extraction incurred by the group as a whole rises with aggregate extraction in accordance with the following hypothesis: the higher the level of aggregate extraction, the more it costs to extract an *additional* unit of the resource. Think of a fishery. The more nets there are in the water, the fewer the fish that will be caught in each net. That is, the more nets it will take to catch a given amount of fish. The cost of catching a fish rises as the total effort devoted to fishing rises. The share of the total cost of extraction that is paid by any given appropriator is equal to the share of this appropriator's extraction in the total extraction by the group. In other words, costs are proportional to harvests. These assumptions imply that an increase in extraction by one appropriator raises the cost of extraction for *all* appropriators.

Figure 1 depicts the manner in which aggregate benefits and costs vary with the level of aggregate extraction. The straight line corresponds to the monetary value of aggregate

extraction and the curve to the aggregate costs of extraction. The costs rise gradually at first and then rapidly, so that there is a unique level of aggregate extraction X^* at which net benefits are maximized. If each appropriator were to extract an equal share of this amount, the resulting outcome would be optimal from the perspective of the group. However, if all appropriators were to choose this level of extraction, self-interested individuals would prefer to extract more since this would increase their own private payoffs. Returning to the fishery example, at X^* an additional net in the water would catch enough fish to more than justify its private cost, but it would lower the catch in all the other nets by enough that the resulting change in total profits would be negative. However, since some of the other nets are owned by other individuals, it would still be privately profitable to use the additional net. The fact that this increase in profit would come at the cost of lowering the combined payoff to the group as a whole would not deter a self-interested appropriator.

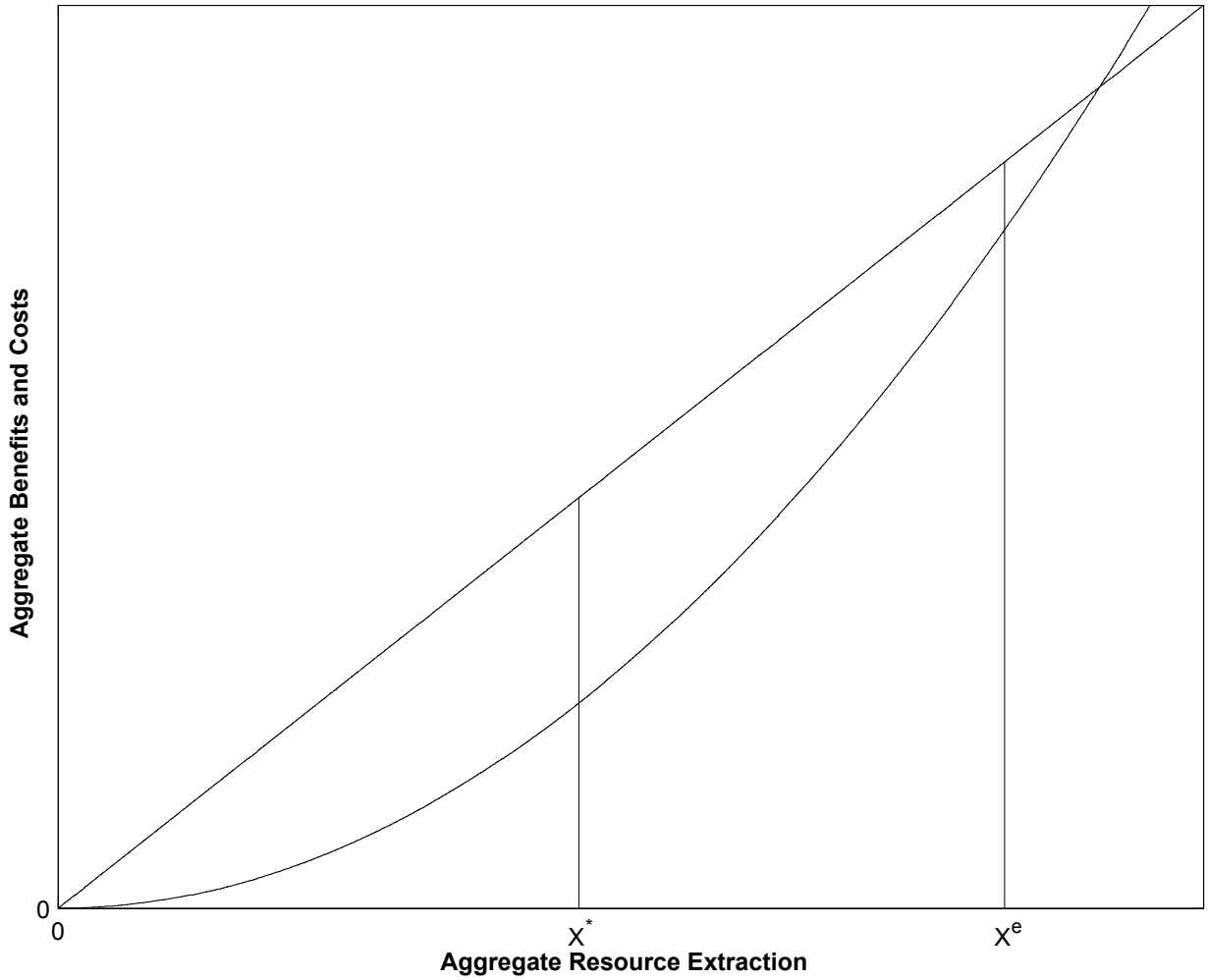

Figure 1. Aggregate Costs and Benefits of Extraction.

If all appropriators were self-interested, and made independent choices regarding their extraction levels, the resulting level of aggregate extraction would *not* be optimal from the perspective of the group. It is possible to show that in a Nash equilibrium of the game played by a group of self-interested appropriators, each one would choose the same extraction level and that the resulting aggregate extraction X^e would exceed X^* (as shown in Figure 1). The level of extraction under such decentralized, self-interested choice is *inefficient*: each member of the group could obtain higher payoffs if all were forced to limit their extraction.

This model allows one to examine how variables of interest like total extraction and total profits change as various parameters change. For example, an increase in the sale price of the harvested resource would be represented by an increase in the slope of the line representing benefits in Figure 1. This would increase X^* and X^e. An increase in the number of appropriators would not change X^* but would lead to an increase in X^e and to a decline in total profits. Many more realistic features can be added to the model. However, the static model suffers from one basic problem. It does not explain why the "tragedy of the commons" is avoided in some cases but not in others. In the model, the tragedy is inexorable. It may be better or worse, but it is inescapable. Other static models of the commons, for example, Bardhan, Ghatak and Karaivanov (forthcoming), or those surveyed in Baland and Platteau (forthcoming) derive differing degrees of cooperation depending on the degree of inequality in various dimensions as well as other factors. It remains true, however, that inefficiency is the general rule.

To explain why cooperation is possible and efficiency (approximately) attainable, 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 above. However, they no longer maximize current payoffs. Rather, they maximize a discounted sum of payoffs from the current and all future periods. The rationale is that for a variety of reasons, the future matters less than the current period, because of impatience, because interest can be earned on resources converted into cash, because of uncertainty that there will be a resource to exploit in the future, and so forth. 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 him what to do in each period in response to the entire history of play upto 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 which individuals would not find in their interest to carry out if called upon 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 everyone adopts the following strategy: they extract an nth share of X^e 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 nth share of X^e in every third period, while exercising restraint in other periods unless someone deviates from this rule, in which case everyone switches to the non-cooperative 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 is 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, 'trigger strategies' of the kind described above will lead to frequent breakdowns and restarts of cooperation (Kreps, 1990). 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 it is costly to start cooperating following a non-cooperative 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. In addition, this model does away with

the need for governance. In fact, 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.

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 sub-optimal) 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 assumption 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 the 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, a number of shortcomings remain. Individualistic unrestrained exploitation is always stable, even if the parameters are such that a norm of restraint would also be stable. Hence there is still some indeterminacy, although less than in the standard model. The model exhibits persistence, but perhaps too much compared to what is observed in the field. And it is silent as to how a norm of restraint might evolve in the first place.

In the next section, we outline a new model that attempts to address these problems. It seeks to fully specify the circumstances under which cooperation will be observed, and departs from orthodox economic modeling in two ways. First, it assumes a simple form of bounded rationality: myopia combined with static expectations. It will turn out that these expectations will be consistent with the actual outcomes after convergence to equilibrium, but not during the transition. The assumption of myopia rules out elaborate contingent strategies. 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 last few years, however, a new way of disciplining the behavioral assumptions made in modeling has become available, the combination of evolutionary theory and experimental work.

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 last 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 a group of players 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 other players 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 payoff costs 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 last few years and they all display these features.[1]

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) discuss 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

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

pretending to be reciprocators.

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 explore the implications for the commons of the interaction between these two preference types.

2 The Model

There are n players, $i = 1, 2, ..., n$, each of whom has access to a common pool resource. We suppose that some mechanism to monitor resource extraction from the common pool, make rules if necessary, and levy fines has been set up at some cost. This has, however, to be financed by on-going 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.

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 $(x_1, ..., x_n)$ of contributions is simply $\alpha X/n - x_i$. It is assumed that

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

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). 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

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

where $x = (x_1, ..., x_n)$ and $y = (y_1, ..., y_n)$ 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 (non-zero only if i did not contribute), and the fourth the cost to i of punishing (non-zero only if $y_i = 1$). 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 assumption is somewhat extreme, it makes the game simple and tractable, and is more plausible than the standard hypothesis that players can work out all the future consequences of their actions and those of others.

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 $u_i(x,y) = \pi_i(x,y) + bx_iy_i$. Reciprocators therefore get a utility bonus b if they have contributed and punished non-contributors. We may interpret this as the psychological satisfaction they get from relieving their feelings of anger at non-contributors. Note that reciprocators get no psychological satisfaction from punishing if they are themselves non-contributors, or from having contributed if they do not punish.

Let $O \subset \{1, ..., n\}$ denote the set of opportunists (material payoff maximizers) and $R \subset \{1, ..., n\}$ the set of reciprocators. Myopia ensures that opportunists will never punish, and so $y_i = 0$ for all $i \in O$. On the other hand, 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 the 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 while being quite bad at solving logically equivalent problems presented in an unfamiliar context. Second, the best-response dynamics 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	Defect	Defect
B	Defect	Contribute & punish
C	Contribute	Contribute, punish if one person defects

In equilibria of type A, there is neither contribution nor punishment. In type B equilibria, opportunists do not contribute, while reciprocators contribute and punish. Finally in type C equilibria, all individuals contribute and reciprocators punish any single individual who deviates from the equilibrium by defecting.[2]

If $k = 0$, then the unique subgame-perfect equilibrium is of type A. If $1 \leq k \leq n-1$, on the other hand, multiple equilibria may exist. Subgame-perfect equilibria of type A with no contributions and no punishments will exist if

$$1 - \frac{\alpha}{n} \geq b - \gamma(n-1), \tag{3}$$

that is, if the net payoff a reciprocator gets from not contributing is greater than or equal to the cost of punishing everyone else plus the utility bonus from punishment. If this condition holds, a reciprocator will be (weakly) worse off switching from $(0,0)$ to $(1,1)$, assuming that all others remain at $(0,0)$. All other requirements for equilibrium are independent of parameter values.

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

$$p \leq 1 - \frac{\alpha}{n}. \tag{4}$$

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 even if doing so did not result in punishment from remaining reciprocators:

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

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

[2] At such equilibria, if more than one person were to defect, then reciprocators will participate in punishment only if this is consistent with utility maximization.

given the strategies of other reciprocators. This requires that the following two conditions hold:

$$b \geq \gamma\left(\frac{n-k+1}{k-1}\right), \text{ and } b-\gamma\left(\frac{n-k}{k}\right) \geq 1-\frac{\alpha}{n}-p \tag{6}$$

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 non-defecting 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 (5) 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 (6). Hence (4), together with either (5) or (6) are necessary and sufficient for a subgame-perfect equilibrium of type B to exist.[3]

Finally consider equilibria of type C, in which all contribute and if one person were to defect, reciprocators punish him. The conditions for such a subgame-perfect equilibrium to exist are that $k \geq 2$,

$$p \geq 1-\frac{\alpha}{n} \tag{7}$$

and

$$b \geq \gamma\left(\frac{1}{k-1}\right). \tag{8}$$

The first of these ensures that no player would gain by switching to defection, while the second ensures that in the event that a reciprocator were to defect, it will be in the interest of the remaining reciprocators to punish him. The latter holds *a fortiori* if an opportunist were to defect.

We have so far neglected the case $k = n$. Here equilibria of types B and C are identical, and will exist if and only if (7) and (8) hold. Except for the non-generic case of $p = 1-\alpha/n$, equilibria of types B and C cannot coexist (except when $k = n$, making them identical). 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

[3]When $k = 1$, (5) is inconsistent with (3) for generic parameter values so these two types of equilibria cannot coexist.

a reciprocator are the same as that 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.

3 Contingent Commitments

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 C will be chosen when these coexist with those of type A, and to perform a similar analysis for the case when B and A coexist. We deal with the coexistence of C and A first.

It may be that (8) 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 condition implies (3) so an equilibrium of type A exists in this case. If, in addition, punishment is strong enough to deter would-be defectors, that is if (7) holds, then equilibria of type C will exist as well. Notice that the equilibrium payoff to *all* players under C is α, which exceeds 1, the payoff from A. This raises the possibility that communication among players at the start of each period can allow them to coordinate on the preferred equilibrium. Specifically, we shall consider commitments by reciprocators of the following kind: I will participate in the punishment of defectors if enough other individuals also participate.

At the second (punishment) stage of the game, given that d people have defected, a reciprocator who has contributed will want to participate in punishment if

$$b \geq \gamma\frac{d}{e}, \tag{10}$$

where $e-1$ is the number of *other* persons the reciprocator expects will punish. Suppose that reciprocators who have contributed at the first stage all believe that each one of them will punish at the second stage, provided that it is optimal for them to do so *conditional on this belief*. This assumes that reciprocators can coordinate their punishments. One may imagine the reciprocators who have contributed gathering in the village square and assessing whether or not there are enough of them to carry out punishment at an acceptable cost. Hence if the inequality holds for some $e \leq n-d$, where d denotes the number of individuals (opportunists and reciprocators) who defected, then punishment will actually be carried out. This follows if we assume that reciprocators can make commitments of the following kind: I will punish

if at least $e-1$ others do so. In that case, choosing the smallest e that satisfies (10) weakly dominates any other commitment. There is now an equilibrium in which all reciprocators make such commitments at the start of each period, expect that the commitments will be carried out by all contributing reciprocators, and all individuals (including opportunists) contribute to the provision of the public good.

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. Such "babbling equilibria" are not observed in everyday experience of coordination problems with pre-play discussion. Experiments on coordination games with two or more players confirm that costless pre-play communication enables players to coordinate on the Pareto-dominant equilibrium (Russell, DeJong, Forsythe and Cooper, 1992; Burton, Loomes, and Sefton, 1999; Blume and Ortmann, 2000; Charness, 2000) 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.[4] For these reasons, we assume that players will coordinate on the C-equilibrium when it exists.[5]

We are now ready to specify how play will evolve from one period to the next. At the second stage of the game, the number of defectors of each type is known, and therefore, the number of punishers, if any, is also determined. At the start of the first stage, therefore, given his expectation of which other players will contribute, a player's choice of whether or not to contribute is clear since he can compute the payoff he will get in either case. As noted above, we assume a simple form of bounded rationality, so that players have *static expectations* of other players' contribution decisions. They expect that the others will contribute as they did in the last period.

These assumptions are sufficient to fully specify how play will evolve from one period to the next. It is immediate that it must converge to one of the subgame-perfect equilibria

[4] This method of equilibrium selection amounts to what has been called 'forward induction'. See Osborne and Rubinstein (1994) pp. 110-115 for a discussion and references to the originators of the concept. This too has been confirmed experimentally by Van Huyck, Battalio and Beil (1993).

[5] 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. We did not take this approach because it seems both less tractable and less empirically plausible than the one adopted here.

specified above. Now suppose it has converged on A although the parameters are such that C is also an equilibrium. This means that A can be expected to prevail in the next period. Therefore, players will soon realize that they are better off agreeing to play C at cost c if

$$c < \alpha - 1. \tag{11}$$

Therefore, if (11) holds then we cannot expect A to prevail in period after period. If we observe a situation with no contributions, this must be either because the initial cost of setting up the contribution mechanism was too high or because it was set up but subsequently collapsed due to adverse parameter changes.

We can use (5) and similar reasoning to show that if the parameters are such that both A and B-type equilibria exist, and if the cost of reciprocators communicating with each other is positive but sufficiently small, then we may expect to see only 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 will not generally result in a change from C to A equilibria, unless there happen to be simultaneous mistakes by several players. 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.

4 Conditions for cooperation

The conditions for cooperation to take place in this set up are (7) and (8) together with (11). What do they imply? First, for punishment to halt defections, the payoff α/n that each player expects to get from cooperating must be sufficiently large that he does not find defection better. This means, of course, that not only must the return to cooperation be high, but it must be *known* to be high by all concerned.

It also means that 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.

For punishment to be cheap to inflict it may be useful to have groups that are not too small as is clear in (8). We do not know what the determinants of the proportion

of reciprocators may be. However, it seems likely that b will depend on the return that each reciprocator expects to get in equilibrium: if reciprocators have too small a stake in the continuance of cooperation, they will not be sufficiently emotionally involved to pursue punishment. Finally, for the public good to be set up at all, the communication cost c has to be sufficiently low.

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 the next section as well.

5 Heterogeneity and other generalizations

Consider as a benchmark the homogeneous player case in which (7) and (8) hold so that C 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

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

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

Suppose for simplicity that there are just now two possible values of p, namely p_l and p_h and that (7) holds for p_h but not for $p_l < p_h$. Suppose $p_i = p_l$ for n_l players and k_l reciprocators and $p_i = p_h$ 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,

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

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 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^* and players are not wealth-constrained, meaning that there exists more than one vector of feasible contributions that add up to X^*. 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. A limited degree of heterogeneity in one dimension, say of 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. We have discussed only heterogeneity of power, but heterogeneity of other kinds, for example in the distribution of returns s_i can be analyzed in a similar way. Heterogeneity, at least within limits, is not as inimical to collective action as one might think.

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 depends on the number and identity of the particular individuals who choose to punish). Suppose the parameters are such that punishment is not an effective deterrent even when all individuals punish. That is, for all players i

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

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 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 1 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 for all $i \in I$,

$$s_i\alpha > 1.$$

Now all the inequalities necessary for a Cequilibrium are in place provided the communication cost c is sufficiently low. For this to be a Pareto improvement over a situation with no contributions it is necessary that $s_j\alpha > 1$. 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.

6 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 modification 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 non-defection) 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, or 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 community enforcement or allows 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 require the consultation and consent of all groups would make this less likely.

We have not addressed several potentially important issues. What factors make it likely that bargaining over the division of the surplus will end in agreement? How does asymmetric information enter the picture? How does the history of cooperation over other issues affect people's expectations about the likelihood of a stable agreement that will be enforced? How does history affect the proportion of reciprocators, or does it not? We leave these interesting but challenging questions to future research.

References

[1] Baland, J-M., and J-P. Platteau (forthcoming). "Collective action on the commons: the role of inequality." in J-M. Baland, P. Bardhan, and S. Bowles (eds.) *Inequality, Cooperation and Environmental Sustainability*, (Princeton University Press and Russell Sage Foundation).

[2] _____, (1998). " Wealth inequality and efficiency in the Commons, part II: the regulated case." *Oxford Economic Papers*, 50(1): 1-22.

[3] Bardhan, P., M. Ghatak, and A. Karaivanov (forthcoming). "Inequality and Collective Action" in J-M. Baland, P. Bardhan, and S. Bowles (eds.) *Inequality, Cooperation and Environmental Sustainability*, (Princeton University Press and Russell Sage Foundation).

[4] Bendor, J., and D. Mookherjee (1987). "Institutional Structure and the Logic of Ongoing Collective Action." *American Political Science Review*, 81: 129-154.

[5] Blume, A., and A. Ortmann (2000). "The Effects of Costless Pre-play Communication: Experimental Evidence from a Game with Pareto-ranked Equilibria." mimeo, University of Pittsburgh and Charles University.

[6] Bochet, O., T. Page, and L. Putterman (2002). "Communication and Punishment in Voluntary Contribution Experiments." mimeo, Brown University, Dept of Economics.

[7] Bowles, S., J. Carpenter, and H. Gintis (2001). "Mutual Monitoring in Teams: Theory and Evidence on the Importance of Residual Claimancy and Reciprocity." mimeo, Middlebury College, Dept of Economics.

[8] Burton, A., G. Loomes, and M. Sefton (1999). "Communication and Efficiency in Coordination Game Experiments." mimeo, University of Nottingham.

[9] Carpenter, J., and P. Matthews (2002). "Social Reciprocity." Working Paper #29, Middlebury College, Dept of Economics.

[10] Charness, G., (2000). "Self-Serving Cheap Talk: A Test Of Aumann's Conjecture." *Games and Economic Behavior*, 33: 177-194.

[11] Cooper, R., D.V. DeJong, R. Forsythe, and T.W. Ross (1992). "Communication in Coordination Games." *Quarterly Journal of Economics*, 53: 739-771.

[12] Cosmides, L., and J. Tooby (1992). "Cognitive adaptations for social exchange," In J. H. Barkow, L. Cosmides and J. Tooby (Eds.), *The Adapted Mind*, (Oxford University Press: New York), pp. 163–228.

[13] Dasgupta, P., and G. M. Heal (1979). *Economic Theory and Exhaustible Resources* (Cambridge: Cambridge University Press).

[14] Fehr, E., and S. Gächter (2000). "The Economics of Reciprocity." *Journal of Economic Perspectives* 14: 151-169.

[15] ————, (2002). "Altruistic Punishment in Humans." *Nature* 415: 137-140.

[16] Gordon, H. S. (1954). "The Economic Theory of a Common Property Resource: The Fishery." *Journal of Political Economy* 62: 124-142.

[17] Kreps, D.M., (1990). "A theory of corporate culture." In J. Alt, and K.J. Shepsle (Eds.), *Perspectives on Positive Political Economy*, Cambridge University Press, Cambridge, pp. 90-143.

[18] Masclet, D., C. Noussair, S. Tucker and M-C. Villeval, (2003). "Monetary and Non-monetary Punishment in a Voluntary Contributions Mechanism." *American Economic Review*, 93(1): 366-380.

[19] Osborne, M. J., and A. Rubinstein (1994). *A Course in Game Theory* (MIT Press: Cambridge).

[20] Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action.* (Cambridge: Cambridge University Press).

[21] Page, T., L. Putterman, and B. Unel (2003). "Voluntary Association in Public Goods Experiments: Reciprocity, Mimicry, and Efficiency." mimeo, Brown University, Dept of Economics.

[22] Sefton, M., R. Shupp, and J. Walker (2002). "The Effects of Rewards and Sanctions in Provision of Public Goods." Working Paper, University of Nottingham, Ball State University and Indiana University.

[23] Sethi, R., and E. Somanathan (1996). "The Evolution of Social Norms in Common Property Resource Use." *American Economic Review* 86: 766-88.

[24] ————, (2003). "Understanding Reciprocity." *Journal of Economic Behavior and Organization*, 50: 1-27.

[25] Van Huyck, J.B., Battalio, R.C., and R.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.