July 23, 2003 Comments Welcome

COOPERATION IN A REPEATED PRISONERS' DILEMMA WITH OSTRACISM

Published: *Journal of Economic Behavior and Organization* (August 1989) 12: 87-106

The unique Nash equilibrium of the finitely repeated n-person Prisoners' Dilemma calls for defection in all rounds. One way to enforce cooperation in groups is ostracism: players who defect are expelled. If the group's members prefer not to diminish its size, ostracism hurts the legitimate members of the group as well as the outcast, putting the credibility of the threat in doubt. Nonetheless, we show that ostracism can be effective in promoting cooperation with either finite or infinite rounds of play. The model can be applied to games other than the Prisoners' Dilemma, and ostracism can enforce inefficient as well as efficient outcomes.

David Hirshleifer and Eric Rasmusen
Anderson Graduate School of Management

UCLA
Los Angeles, California 90024
(213) 825-4154

Bitnet: IJJ1RAS@UCLAMVS.

File: /papers/ostrac/ostrac.tex.
Draft: 14.13.
UCLA AGSM Business Economics Working Paper # 86-9.

2003 note: Hirshleifer is now at Ohio State, and Rasmusen is at Indiana University.

We thank Sushil Bikhchandani, Robert Boyd, Michael Brennan, Ivan Png, John Maynard Smith, the UCLA Political Science Theory Workshop, and two anonymous referees for helpful comments, and John Mamer for stimulating our interest in this topic.

2000: Eric Rasmusen, Professor of Business Economics and Public Policy and Sanjay Subhedar Faculty Fellow, Indiana University, Kelley School of Business, BU 456, 1309 E 10th Street, Bloomington, Indiana, 47405-1701. Office: (812) 855-9219. Fax: 812-855-3354. Erasmuse@indiana.edu. Php.indiana.edu/∼erasmuse.

A Parable.

The Ostracos are a primitive tribe whose members hunt collectively for large game. Anyone who does not hunt, fishes by himself from the tribal lake for a bare subsistence. Fishing is subject to constant returns: the catch per capita is independent of the number of tribesmen. Hunting is subject to increasing returns: meat-per-capita increases with the number of hunters. Since the inviolable custom of the tribe is that hunters share meat with non-hunters, fishermen get both fish and meat. Consequently, nobody engages in hunting.

One evening the elders have a pow-wow to discover why the precious meat is not being brought home to the tepees. In a divination the spirit Phonos tells them that those who avoid the hunt are cowards, not tribesmen, and must be driven into the wilderness to die. The elders accordingly so ordain.

The next day, all come to the hunt except Uurguu, a powerfully built man with a large stomach. On being chided by his fellows, he pronounces "We tribesmen are all equally good hunters and fishers, and rational men to boot. So we all know that I have just as much reason to join the hunt tomorrow as any of you. And we also know that if I do help with the hunt, there will be more meat for each of us than if you drive me away. Let bygones be bygones, to the benefit of all."

To his immense astonishment, Uurguu was immediately expelled.

1 Introduction.

A topic of continuing interest in social theory is how cooperation can emerge in the repeated Prisoners' Dilemma and similar games. Cooperation in all rounds can be attained as one of many Nash equilibria in the infinitely repeated Prisoners' Dilemma, a two-player example of which is presented in Figure 1. In finitely repeated games, however, it is well known that this action pair cannot be part of an equilibrium. Cooperation is hard to sustain, because in any proposed equilibrium there ultimately is some round t^* (if not earlier, then certainly the last round) at which Player 1 foresees no future cooperation from Player 2. In round t^*, whatever Player 2 may do, Player 1 will choose to defect, which causes the sequence of defections to begin in round $t^* - 1$. The backwards recursion continues all the way to the very first round, so both players defect from the very beginning of the game.

Insert Figure 1: The Prisoners' Dilemma with Two Players.

Several solutions have been proposed to achieve cooperation. Kreps et al. (1982) suggest that incomplete information is important. If there is the slightest possibility that one's opponent is a type who, independently of any rational calculation, will cooperate with you if you cooperated in the past, then cooperation for every round up to some round close to the end of the game can arise in equilibrium. Another source of cooperation is commitment to a retaliatory strategy, an approach used by Schelling (1960) and Thompson & Faith (1981). If commitment is feasible, the players in a Prisoners' Dilemma are both better off if they each commit to never play Defect first and to retaliate heavily if the other player does defect. J. Hirshleifer (1987) argues that emotions of revenge and gratitude evolve as ways to make retaliation credible. Still another way to achieve cooperation is to assume that players are altruistic towards each other. But none of these solutions address the question of how cooperation between self-interested individuals can arise without binding promises, emotional responses, or deception.

The solution proposed here allows retaliation against noncooperators by a means other than future noncooperation. The new form of retaliation is *ostracism*: expulsion of the defector from the group. Ostracism does not require commitment, and the equilibrium satisfies the rationality criterion of "subgame perfectness" whether the game has a finite or infinite number of rounds.[1]

The word "ostracism" derives from Greek word for the broken shards on which the citizens of ancient Athens recorded their votes expelling individuals regarded as threats to the state. More generally, ostracism is the practice of excluding disapproved individuals from interaction with a social group. In one form or another, this plays an important role in enforcing socially approved behavior in most groups. Parliamentary bodies have means by which they may expell members, and professional societies have means of decertifying them: disbarring lawyers and taking away the licenses of doctors and accountants.

If ostracism were a costless way to make threats and promises credible, the social dilemma would be easily solved. But ostracism is usually costly to the group because expelling a member hurts not just the outcast, but indirectly all the remaining members. The very fact that members are in a group indicates some advantage to grouping which would be reduced by expelling members. We will call the gains from joining together *aggregation economies*, which may arise from scale economies in productive technology, gains from trade and specialization, simple sociability, or network externalities.[2]

Under conditions that we will specify, ostracism is a credible threat, and cooperation can be achieved because society can exploit the end period problem rather than be victimized by it. In the final round, defection is a dominant strategy. But this means that there are no gains from cooperation in the final round, so (in contrast to earlier rounds) the outcasts will not be missed. In the next to last round, the threat to expel defectors is therefore credible, and the threat enforces good behavior in preceding rounds.

Section 2 lays out the model. Section 3 demonstrates that ostracism can enforce cooperation with either a finite or infinite number of rounds. Section

4 discusses the assumptions and compares them with other models. Section 5 adds a touch of morality to eliminate multiple equilibria, and Section 6 shows how even Pareto-inferior outcomes can be supported by ostracism. Section 7 gives examples of ostracism and discusses which applications fit our model.

2 The Model

At the start of the game, players form a group to cooperatively produce a good. If they play Cooperate, more of the good is produced than if they play Defect. Defection may be thought of as shirking. A player who is in the group, having not been ostracized in the immediately preceding round, is called a *member*. We will call the individual member's vote for ostracizing someone his *blackball*. One blackball suffices for ostracism. The term "blackball" refers to the action of a player in voting to expel, while "ostracism" refers to his actual expulsion. Each player gets some base level of satisfaction (normalized to zero below) simply by being in the group, but his satisfaction is greater (a positive payoff) if the other members cooperate. The group can exclude a player from even the base level (i.e., give him a negative payoff) by ostracizing him, so ostracism is a punishment even when no one cooperates, and is more painful to the deviator than merely having other players also Defect. In the parable, fishing by all members leads to the base level of welfare; hunting, which requires cooperation, is the positive level; and expulsion, which prevents a player from either hunting or fishing, is the negative level.

We will examine the equilibrium under the following key assumptions:

1. **Free Rider Problem**. A defector gets a higher payoff than a cooperator in the round in which he defects.

2. **Aggregation Economies**. Payoffs per member are an increasing function of the number of members who cooperate.

3. **No Aggregation Economies without Cooperation.** The payoffs per member when every member defects do not depend on the size of the group.

4. **Excludability of Resources from Non-Members**. A player would rather be in the group than ostracized, even if every member defects.

5. **Costless Enforcement.** Blackballing has no direct cost or benefit to those members who engage in it.

Assumptions less central to the results include

1. Ostracism only lasts one round. A player who is ostracized can re-enter the next round unless ostracized again.

2. To be ostracized, a player need be blackballed by only one member.

3. A player can blackball any number of other players.

Each round t is divided into two phases: an *ostracism phase*, labelled t^{os}, and a *dilemma phase*, labelled t^{pd}. The game starts with $\bar{n}$ players who are all members, and continues either until round T or forever, depending on the particular version of the model. In any round t, let n_t denote the number of members who play in the Prisoner's Dilemma at the end of the round (the dilemma phase) and let n_t^{os} denote the number at the start of the round (the ostracism phase). This will result in $n_t^{os} = n_{t-1}$, as shown in Figure 2.

Figure missing: ostrac1.eps

In the blackballing phase of round 1, each of the $\bar{n}$ members may blackball any of the other members. Any player who is blackballed at 1^{os} is excluded from the next dilemma phase, 1^{pd}, and from the next blackballing phase, 2^{os}. (His exclusion from 2^{os} is not essential for the results.)

The number of members in 1^{pd}, denoted n_1, may be less than $\bar{n}$, since some players may be ostracized at 1^{os}. The game is repeated at round 2 with n_2^{os} members playing in the ostracism phase 2^{os}, which leaves n_2 members to play in the dilemma phase 2^{pd}. This continues through the final dilemma phase T^{pd}, or forever if the game is infinite.

A player ostracized in t^{os} does not play in the multiperson Prisoners' Dilemma in t^{pd}. His total payoff for the round is $-Y$, the cost of being a non-member. In addition, he remains a non-member at $(t+1)^{os}$, which means that he cannot participate in blackballing in that round. Unless he is ostracized again at $(t+1)^{os}$, he is free to rejoin the group at $(t+1)^{pd}$.

Dropping the t subscript to avoid clutter, let us write the payoff functions using n for the number of members in the dilemma phase and n^c for the number of members who cooperate. All output is split equally among the members. The average output per member is denoted by $f(n^c, n)$, and the cost to a single member of cooperating is denoted by X. For the game to be a Prisoners' Dilemma, defecting must be a dominant strategy in the one-period game, so we require that for any number of cooperators $m > 0$ and members $n > 0$,

$$f(m, n) - X < f(m-1, n). \tag{1}$$

We also require that

$$f(m-1, n-1) < f(m, n) \quad , \tag{2}$$

which is the mathematical statement that there exist aggregation economies: the presence of an additional cooperating member raises per capita output. Setting $m = n$, inequality (2) implies that per capita payoffs are larger in a larger cooperating group. Let us normalize the output with zero cooperators to $f(0, n) = 0$, which satisfies "No aggregation economies without cooperation." As a complement to (2), we will assume that the presence of a defecting member does not raise the average output. Therefore,

$$f(m, n-1) \geq f(m, n). \tag{3}$$

Ordinarily, inequality (3) would be strict because the presence of a free-riding member would strictly lower the average output.

When all members cooperate, each receives a payoff of $f(n, n) - X$. If some members defect, the payoff to each defector is $f(n^c, n)$ and the payoff to each cooperator is $f(n^c, n) - X$. When all defect, each member receives a payoff of zero. A player who is ostracized receives $-Y$. The period's payoff to member i is therefore

$$\pi_i = \begin{cases} f(n^c, n) - X & \text{if } i \text{ cooperates} \\ f(n^c, n) & \text{if } i \text{ defects} \\ f(0, n) = 0 & \text{if all players defect} \\ -Y & \text{if } i \text{ is ostracized} \end{cases} \tag{4}$$

Let $\delta \in [0,1]$ be a discount factor common to all players. The total payoff to player i for the entire game is $\sum_{t=1}^{T} \delta^{t-1} \pi_{it}$.

In addition to the general assumptions, we must also restrict the magnitudes of the parameters. Let us assume that

$$Y > X, \tag{5}$$

so that in each round, the penalty from being ostracized is larger than the benefit from cheating against cooperators.

Finally, let us assume that "cooperation" is socially valuable, so that the per capita payoff net of costs is higher if all cooperate than if all defect,

$$f(n,n) - X > 0 \quad \text{if} \quad n > 0 \quad . \tag{6}$$

As will be shown by the example in Section 6, Assumption (6) is not always necessary for cooperation to be an equilibrium, but it is used in the proof of the main proposition below.

The payoffs in the Prisoners' Dilemma game of Figure 1 satisfy these assumptions. Let $\bar{n} = 2$, $f(2,2) = 30$, $f(1,2) = 5$, and $X = 15$. According to payoff function (4), if both players cooperate their payoffs are each $f(2,2) - X$, if both defect the payoffs are each 0, and if only one cooperates, his payoff is $f(1,2) - X$ while that of the other player is $f(1,2)$.

3 The Equilibrium.

In this section we will examine a symmetric perfect equilibrium in which all players adopt a strategy called *Banishment*. A player following *Banishment* cooperates along the equilibrium path and blackballs anyone who deviates from the strategy— which includes defectors, players who blackball when unprovoked, players who fail to blackball defectors, players who fail to blackball those who fail to blackball defectors, and so forth. On the equilibrium path, everyone cooperates; if anyone deviates in any way, the others still cooperate, but they blackball him. Banishment is forgiving in the sense that retribution is limited: after a single round of ostracism, the outcast is permitted to return to the group without prejudice.

In the final round, defecting is a dominant action because no punishment can follow. In models without ostracism, this is the fatal first domino that successively overthrows cooperation back to the first round. In using this inductive argument, we are imposing the requirement, now standard in game theory, that the equilibrium be subgame perfect: the relevant portions of an equilibrium strategy are Nash equilibria for every subgame of the original game, whether or not that subgame is reached in equilibrium.[3] Both players always defecting is a perfect equilibrium as well as a Nash equilibrium of the finitely repeated game when ostracism is not used.

If players' strategies incorporate ostracism, however, they may cooperate until the last round even in a subgame perfect equilibrium. In the last round all players defect, just as in the single round Prisoners' Dilemma. But in the next-to-last round there is no cost to expelling defectors, since everyone knows there are no future gains from having a large group. The credible threat of expulsion for the last round enforces cooperation and forces players to blackball cheaters in earlier rounds. Even though in early rounds the group gain from having an additional cooperating teammate is large, ostracism can still be enforced, because the gain from not ostracising a deviator is spread among the group, while the punishment for failing to blackball him falls on individuals.

Banishment Strategy

Dilemma Phase. Before round T, cooperate unless you have violated *Banishment* in the immediately preceding ostracism phase, in which case defect. At round T, defect.

Ostracism Phase. Blackball any player who in the immediately preceding dilemma or blackballing phases deviated from the strategy *Banishment*, and do not blackball anyone else.

Banishment specifies cooperation as the player's equilibrium behavior until round T. The blackballing action rules are iterative. In phase 1^{os}, a player refrains from blackballing. In phase 2^{os}, he blackballs any player who either defected in 1^{pd} or blackballed in 1^{os}. For $t \geq 3$, in phase t^{os} he blackballs any player who (i) defected in $(t-1)^{pd}$; or (ii) blackballed in $(t-1)^{os}$ without provocation; or (iii) failed to blackball in $(t-1)^{os}$ when he should have in response to a deviation in round $t-2$.

We will prove that *Banishment* supports an equilibrium with cooperation in every round but T. Deviators are blackballed, because failing to blackball properly provokes blackballs against oneself.[4] Since all players will defect at T^{pd} anyway, there are no gains from having more members at T^{pd}, so there is no loss to ostracizing someone at T^{os} who deviated in round $T-1$.

Proposition 1: *With sufficiently little discounting, there exists an equilibrium with cooperation in rounds 1 through $T-1$ of the T-round ostracism game.*

Proof: We use backward induction to verify that the strategy combination in which all players follow *Banishment* is a subgame perfect equilibrium. The outcome is then cooperation by all players until the last round. To do so, we must show that no player has an incentive to deviate from this proposed equilibrium in any subgame. Let us start with the subgame consisting of round T alone.

(1) Phase T^{pd}.
In phase T^{pd}, if a member deviated by cooperating he would receive $-X$ instead of 0. (In this phase and in any earlier phase, if no members remain then trivially the equilibrium strategy is not violated at that point in the game.)

(2) Phase T^{os}.
In phase T^{os}, a member is weakly willing to blackball any player who violated *Banishment*'s rules in phases $(T-1)^{os}$ or $(T-1)^{pd}$, because in the final phase, T^{pd}, every member will defect in any case, and the all-defect payoff of zero is independent of the number of members.

Let us next consider any round $t < T$, under the inductive assumption that all players will follow *Banishment* in all subgames starting after round t, *including* those subgames which would not arise if the players follow *Banishment* through round t. We first examine the behavior of a given player, whom we will call player A, in the dilemma phase t^{pd}.

(3) Non-deviation subgame in phase t^{pd}.
Suppose that player A did not deviate in t^{os}, though the other players may have, and either he or other players may have deviated earlier in the game. Under *Banishment*, those other players who deviated in t^{os} will defect in t^{pd}, and those who did not will cooperate in t^{pd}. Let n_t^* be the number of cooperators in phase t^{pd} if A cooperates. Then A's immediate payoff from defecting in t^{pd} is the receipt of $f(n_t^* - 1, n_t)$ instead of $f(n_t^*, n_t) - X$. But if he defects, he will be ostracized in phase $(t+1)^{os}$, which yields him a discounted loss of $-\delta Y$ instead of: (a) $\delta[f(n_{t+1}, n_{t+1}) - X]$, if $t < T-1$; or (b) zero, if $t = T-1$. If $t = T-1$, the game ends at $t+1$; otherwise, when A returns in $t+2$ everyone goes back to cooperating through $T-1$, exactly as if A had not deviated, so A's payoffs in $t+2$ and beyond are the same regardless of whether he deviates in t. Therefore, the condition for A to prefer weakly to cooperate is

$$f(n_t^*-1, n_t) - \delta Y \leq \begin{cases} f(n_t^*, n_t) - X + \delta[f(n_{t+1}, n_{t+1}) - X] & \text{if } t < T-1 \\ f(n_t^*, n_t) - X & \text{if } t = T-1. \end{cases} \tag{7}$$

Inequalities (2) and (3) together imply that $f(m-1,n) < f(m,n)$, (per capita output is raised by adding a cooperator), and by (5), $X < Y$, so for δ sufficiently close to one, condition (7) is satisfied. If A did not deviate in t^{os}, he will not defect in t^{pd}.

(4) Deviation Subgame in phase t^{pd}.
If A has already deviated in t^{os}, then knowing that he will be ostracized in $(t+1)^{os}$ anyway he bears no additional penalty to defecting in t^{pd}. His payoff if he defects is $f(n_t^* - 1, n_t)$, and if does not defect, $f(n^*, n) - X$. The net gain to defecting earned at t^{pd} is positive, by (1), and the amount A earns in all later rounds is unaffected. This verifies that if A had deviated from *Banishment* in the immediately preceding ostracism round, he will defect as required by *Banishment*.

We have therefore verified that *Banishment* is followed in t^{pd}.

(5), (6), (7) The decision in phase t^{os}.

We next calculate A's gain from deviating in t^{os}, given that all the other players will follow *Banishment* for the remainder of the game. We must now distinguish carefully between the number of members present in t^{pd} when A deviates in t^{os} versus when he does not. Let $n_t(eq)$ be the number of members in t^{pd} if A obeys *Banishment* in t^{os}, and let $n_t(dev)$ be the number of members in t^{pd} following a deviation called *deviate* (some pattern of extra or insufficient blackballs that A directs at different players) by A in t^{os} . In equilibrium, the number of members in $(t+1)^{pd}$ is $n_{t+1} = \bar{n}$, the total number of players, because no player deviates at t^{os} or t^{pd}, so no blackballs are cast in $(t+1)^{os}$.

(5) A's decision in phase t^{os} given that he will be blackballed in that phase.
If A foresees being blackballed in t^{os} (perhaps because he had deviated in round $t-1$), then he expects to be ostracized during t^{pd} and his payoff at that phase is unaffected by his decision in t^{os}. If he deviated in t^{os}, he would obtain a payoff of $-\delta Y$ from being expelled in $(t+1)^{os}$ instead of the payoff of $\delta[f(\bar{n},\bar{n}) - X]$ from cooperating in $(t+1)^{pd}$. An exception is if his deviation

at t^{os} is a "clean sweep" that eliminates all the other members, in which case he earns zero at $(t+1)^{pd}$, but this payoff is still lower than $\delta[f(\bar{n},\bar{n})-X]$. By the inductive hypothesis, his payoffs are unaffected for the remainder of the game. So he strictly refers not to deviate.

Suppose, on the other hand, that player A does not expect to be blackballed in t^{os}. We will divide the ways in which he might then deviate in t^{os} into two cases ([6] and [7]).

(6) A's Decision in phase t^{os} given that he will not be blackballed in that phase: clean sweep deviation.
One possible deviation in t^{os} is for A to blackball every player, when that is not called for by *Banishment*. This deviation is special because no other members remain to ostracize A in phase $(t+1)^{os}$. By deviating in this way, A earns zero instead of $f(n_t(eq),n_t(eq))-X$ in t^{pd}. The deviation is immediately unprofitable, and by the inductive hypothesis it creates no change in A's payoffs in $(t+1)^{pd}$ or any later point in the game.

(7) A's Decision in phase t^{os} given that he will not be blackballed in that phase: not a clean sweep deviation.
Consider any deviation by player A in t^{os} that does not expell all the other players. Following such a deviation, A will be blackballed in $t+1$, giving him a payoff of $-\delta Y$ instead of $\delta f(\bar{n},\bar{n})-X$. In addition, we have shown in Part (4) of this proof that after he has deviated at t^{os}, A will defect at t^{pd}, for a payoff of $f(n_t(dev)-1,n_t(dev))$ in that phase instead of $f(n_t(eq),n_t(eq))-X$. A's total gain from deviating is nonpositive if

$$f(n_t(dev)-1,n_t(dev))-\delta Y \leq \tag{8}$$

$$\begin{cases} f(n_t(eq),n_t(eq))-X+\delta[f(\bar{n},\bar{n})-X] & \text{if } t<T-1 \\ f(n_t(eq),n_t(eq))-X & \text{if } t=T-1. \end{cases}$$

If part of the deviation is to cast an unwarranted blackball, reducing n_t, then that part of the deviation is unprofitable for the deviator by (2), because it diminishes the size of the group and the number of cooperators equally in the dilemma phase t^{pd}; since we rule out clean sweep deviations here, unwarranted blackballing never prevents the punishment of ostracism at $(t+1)^{os}$.

So if deviations that do not include unwarranted blackballing can be shown to be unprofitable, it will follow that those that do are also unprofitable.

We therefore consider the more tempting deviation of failing to blackball when *Banishment* calls for it after another player or players deviate. We divide this kind of deviation into two cases.

If at least two members besides A are in the group in phase t^{os}, any player whom *Banishment* requires to be ostracized will still be ostracized even if A were to refrain from blackballing. Therefore, $n_t(dev) = n_t(eq)$, and it follows from (5) that if δ is near 1, (8) is valid and no deviation involving a failure to blackball is profitable.

If only one other member is in the group in phase t^{os}, then if A fails to blackball him as required by *Banishment*, the number of members in the group for phase t^{pd} increases from 1 to 2. The strategy *Banishment* calls for A to continue by defecting in the dilemma phase, so his payoff there changes from $f(1,1) - X$ to $f(1,2)$. We do not know whether this is an increase or a decrease, since perhaps $f(1,1) > f(1,2)$, but offsetting any possible increase is the fact that after failing to blackball and then defecting, A himself will be ostracized in $(t+1)^{os}$ and his payoff (discounted back to t) will fall by at least δY. The gain to this deviation is negative if

$$f(1,1) - X > f(1,2) - \delta Y.$$

By assumption (3) (adding a non-cooperator does not raise the per capita payoff), $f(1,2) \leq f(1,1)$, and by assumption (5), $Y > X$. Therefore, if δ is close enough to 1, player A loses if he fails to blackball appropriately.

Having verified that *Banishment* is followed in t^{os}, by induction it is a subgame perfect equilibrium in all rounds.

Q.E.D.

Infinite Rounds

It is well known that cooperation can arise in the perfect equilibrium of the infinitely repeated Prisoners' Dilemma, in contrast to the game with a

large but finite number of repetitions. This is implied by the "Folk Theorem", which says that virtually any pattern of actions can be generated by the equilibria of infinitely repeated games, if the discount rate is sufficiently low (see Fudenberg & Maskin [1986] and Rasmusen [1987]). The end round argument from the introduction that ruled out cooperation in the finite game does not apply to the infinite game. Because infinitely repeated games allow so many patterns of behavior, the fact that adding ostracism to the game still allows cooperation is unsurprising, but we will discuss it briefly.

One type of equilibrium strategy for the infinite game is any non-ostracism strategy that enforces cooperation plus the rule "Never blackball." A slight modification of the strategy *Banishment* also enforces cooperation in the infinite game. Define *Banishment* as before, but eliminate the inapplicable part of the definition which refers to defecting in period T. By reasoning similar to Proposition 1's proof, *Banishment* enforces cooperation when discount rates are sufficiently low. If a player deviates, he will be ostracized for a round and then the game proceeds as before along the equilibrium path. This is perhaps a more attractive equilibrium than those which rely on retaliatory defecting, because the cost of ostracism is heaviest for the deviator, which fits our sense of what happens in the world.

Moreover, when the infinitely repeated multiplayer Prisoners' Dilemma is expanded by introducing ostracism, cooperation is possible under a wider range of parameters. If there is heavy discounting (δ near zero), even the infinitely repeated game has all-defect as its unique equilibrium outcome. But if ostracism is possible, and the ostracism penalty of Y is large enough relative to the low discount factor, cooperation can be achieved.

4 Discussion.

Relaxing Assumptions.

The arguments explaining how ostracism can enforce cooperation can be applied not only to the Prisoner's Dilemma, but also to other social dilemmas such as coordination games (see Rasmusen [forthcoming] or Sugden [1986] for examples: e.g., the Battle of the Sexes). Assumption (1), which characterizes the strong temptation to defect in the Prisoners' Dilemma, was not crucial in the proof in the previous section. Its only significance was in making Defection part of the equilibrium strategy on an irrelevant off-equilibrium path; even without this assumption, cooperation is still an equilibrium.

Perhaps the most important assumption to check in deciding whether the ostracism model applies is "No Aggregation Economies Without Cooperation": if all players defect, is there no benefit from a larger group size? Lack of such a benefit is crucial to the argument that the members are willing to ostracize a deviator in the last period.

The particular blackballing rule is not important. We assumed that one blackball sufficed for ostracism, but similar results can be derived if ostracism requires blackballs by a majority of members, or even if it requires unanimity aside from the member in question. Ostracism also works in much the same way if it is irrevocable, i.e., if once a player is ostracized he can never rejoin the group. Executing a player is an example of irrevocable ostracism. In the last period, execution is just the same as temporary ostracism, at least from the ostracizer's point of view: the offender disappears for a round. Earlier period executions have the same effect on the group as irrevocably ostracizing the player, and the qualititative effect on him is also the same: he forever loses the benefit of being in the group. Both modifications, different blackballing rules and irrevocable ostracism, require changes to the details of Proposition 1's proof, but do not change the thrust of the argument.

Execution is not the only punishment that can be modelled as ostracism. Imprisonment is another example. What is required for the model to apply

is that the punishment diminish the deviator's ability to contribute to the group. When society imprisons a criminal, it loses the benefits it would have had from his cooperation, if that cooperation could have been ensured. The main thing that might distinguish ostracism from prison is that in the ostracism model we assumed that the punishment imposesd no direct cost on either the group's output or the blackballer. The tax to pay for the deviator's stay in jail is a direct cost. Such a cost would prevent ostracism at T^{os} and cause the equilibrium to unravel. On the other hand, we also ruled out possible direct benefits such as seizing the property of the deviator. Since our equilibrium in the last round is weak, relaxing these assumptions would lead to either never ostracizing or ostracizing even without provocation, both of which would eliminate the equilibrium with cooperation. In Section 5, we will demonstrate that a small amount of morality (desire to punish past wrongdoers) converts our weak equilibrium into a strong one. If there are direct gains or losses to individuals from ostracizing, or if in the all-Defect outcome payoffs are not perfectly independent of the number of players, then morality can still enforce cooperation if the costs are small.

A Comparison with Other Models.

The strategy *Banishment* is reminiscent of Thompson and Faith's (1981) model in that cooperation is enforced by sequences of threats that involve not only threats to punish the defector, but threats to punish those who fail to punish, and so on. But there are important differences. In Thompson and Faith there is a hierarchical structure in which players higher in rank punish those lower in rank. Commitment to punishment is allowed, and the decision hierarchy is a series of moves in which the different players commit to punishment strategies in sequence. The outcome is dictatorial, in the sense that the most highly preferred outcome of the first mover is achieved.

In our model there is no prior asymmetry between players, and commitment is ruled out by requiring subgame perfectness. Players move simultaneously and they are identical, so that a "democratic" outcome is achieved, rather than the favored outcome of a special player (which in context of Thompson and Faith would involve the dictator playing Defect and all other

players Cooperate). The sequencing of threats that enforces cooperation is endogenous, arising strategically from the interaction of identical players. Our model is less applicable than Thompson and Faith's to the punishments of hierarchical organizations, such as excommunication by the Roman Catholic Church.

Bendor and Mookherjee (1987) also analyze social outcomes when the group can threaten punishments, of which expulsion is an example. Their setting emphasizes the problem of observing whether defection has occurred, rather than the credibility of the threat of punishment.

In their well-known paper on the finitely repeated Prisoners' Dilemma, Kreps et al. (1982) base cooperation upon incomplete information. In their game, players defect in the last k rounds, where k is determined by the parameters but is independent of T, the total number of rounds in game. If T is large, the fact that there is defection in the last k rounds is unimportant. On the other hand, if T is small, the social inefficiency can be relatively severe, and if T is less than k, the players never cooperate. Ostracism works very differently: it does not depend on incomplete information, and it works as well with a a small number of rounds as with a large number.

5 A Little Morality.

Since in many contexts ostracism is not purely selfish behavior, it seems reasonable to consider the possibility that players are slightly moralistic, so they gain a little bit of pleasure from blackballing a deviator. As described above, the equilibrium with *Banishment* is a weak equilibrium: given that the other players follow *Banishment*, a player is also willing to follow it, but he is indifferent about following certain of its action rules. Moreover, another weak symmetric equilibrium is *Always Defect, Never Ostracize*. The cooperative equilibrium is Pareto-superior, so the players may hope for it to be a focal point, but choosing focal points is always somewhat arbitrary.

Let us define "a little morality" as a small positive payoff from blackballing a deviator according to the strategy *Banishment*. In the last ostracism phase, the moralistic player is not indifferent about ostracizing a deviator; by ostracizing, he unambiguously raises his payoff, if only slightly. Because of this, *Banishment* is a strong equilibrium for the T-period game. Perhaps even more importantly, it becomes the unique equilibrium. The strategy *Always Defect, Never Ostracize*, for example, is no longer an equilibrium, because players would raise their payoffs by ostracizing defectors in the next-to-last round. The proof that the *Banishment* outcome is unique essentially follows that of Proposition 1, except now the inductive hypothesis is that *Banishment* is the *only* subgame perfect equilibrium. *Banishment* behavior at the last round, which is strongly preferred by morality, deters any kind of action except for following *Banishment* at the next to last round, which enforces a cascade of threats back to the first round.

Morality is similar to altruism as an escape from the Prisoners' Dilemma, but it is not the same. Altruism achieves cooperation because some players unconditionally want to improve the welfare of others. Morality achieves cooperation because some players want to reduce the welfare of others, if those others behave wrongfully. In fact, altruism in some players would prevent morality from enforcing cooperation, by making the altruists unwilling to punish evildoers.

If the model were modified to allow for direct costs or benefits of ostracizing, then as discussed in the previous section, *Banishment* would no longer be an equilibrium strategy. But if these costs or benefits are small, morality could still persuade players to ostracize when appropriate even if it is costly, or to refrain from ostracizing even if it yields direct gains. Of course, if morality were sufficiently strong, cooperation could be supported by an ethic that called for unconditional cooperation ("the golden rule"). But such a scheme places a heavy burden on morality, because it must overcome the temptation to defect to seize large gains in the dilemma phases. With ostracism, the temptation is much smaller: a little morality goes a long way.

6 Bad Equilibria Enforced by Ostracism.

Although ostracism is able to enforce cooperation in the repeated Prisoners' Dilemma, it is not necessarily a good thing. We can apply the notion of ostracism to repeated games that would normally have desirable outcomes, but in which ostracism causes the players to engage in undesirable behavior. This is a game theoretic version of the idea that social custom can result in economic inefficiency (Akerlof [1976, 1980], Romer [1985]). Our interpetation of that idea is that many games have multiple perfect equilibria, and society may be stuck at an undesirable one for historical reasons.

We will use a numerical example to illustrate how ostracism can hurt a group. Let there be 6 players, who choose C (*Customary* behavior) or D (*Desirable* behavior) in a two-round game with ostracism and no discounting. The actions C and D correspond to "Cooperate" and "Defect" in the general model, but here the payoffs are such that D is a better outcome than C. The reason for this somewhat counterintuitive reversal is that we want to provide an example in which, absent ostracism, the individual's temptation is to perform the socially desirable act, and yet with ostracism he does not. In the Prisoners' Dilemma without ostracism, $Defect$ is the action that the individual is tempted to take, and this feature of the payoff scheme applies to strategy D here. In a round in which the group has n members who all cooperate, they each get $5(n-1)$, while if all defect they each get 100. If only some players defect, the defectors each get and the cooperators each get 0. Figure 3 summarizes these payoffs. If a player is ostracized, he gets a payoff of -150 in the second round.

Insert Figure 3.

If this game did not have ostracism, the unique equilibrium would be for every player to play D in both rounds. In the last round D is a dominant action, since by playing D the player gets 100 if all the others play D and 40 if they do not; following customary behavior would give him a payoff between 0 and 30 when there are 6 players. Since everyone will play D in

the last round, defecting is also the Nash strategy for the first round. The equilibrium payoff is 200 per player.

Ostracism adds another equilibrium, in which all the players cooperate in the first round and defect in the second:

Pareto-Inferior Equilibrium Strategy.
Choose C in the first round.
Blackball any player who chose D in the first round.
Choose D in the second round.

The players defect in the second round because that remains a dominant action regardless of ostracism. They are willing to blackball a player who defects in the first round because they foresee that in the second round the payoffs will be the constant 100, which does not decline if they ostracize some players. Each player will cooperate in the first round because his total payoff is then 25 $(= 5[6 - 1])$ plus 100, as opposed to the 40 plus -150 he would receive from defecting and being ostracized. The equilibrium payoff is 125 per player.

C can be replaced by one's least-liked custom, as long as the custom meets the assumptions of the model. Suppose, for example, that the group is a set of trading parties. C and D could mean "Customary Wage" and "Market-Clearing Wage," "Do Not Lend" and "Usury," or "Shun Blacks" and "Hire Blacks." The members of the group would obey the bad customs for fear of being excluded from trade with the other members. Other examples might be students who disapprove of cheating on exams, but shun weaseling as even worse, or societies where hyper-sensitivity to slights and willingness to duel is enforced by the fear of public contempt. The bad equilibrium with ostracism is not the only equilibrium of these games— another exists in which the players never ostracize and always defect. Many theorists would predict that the efficient equilibrium would be the actual one, since it is both simpler and better. This is based on the view that simplicity and efficiency are properties of focal points, and that player who can communicate will settle upon an efficient, self-enforcing equilibrium. But if historical accident and psychological factors are important in establishing norms of behavior,

the result can be Pareto-inferior equilibria.

7 Applications.

Various practices that groups use for disciplining their members can replace ostracism in our model, but some practices are very different. Unlike the players in the standard Prisoners' Dilemma, many groups can use punishments such as lump-sum fines that are not costly to the group, or can precommit to punishments. Fines and other forms of expropriation, however, do not fit the technical requirements of our model, because fining a defector does not harm the rest of the group; indeed, it benefits them. If fines are available, it is not at all surprising that the group can enforce cooperation. But sufficiently severe fines are often infeasible. For example, the group may lack the legal authority to expropriate physical property, and only have the ability to withhold its society. Or, punishment severe enough to deter transgression might have to be nonmonetary and might unavoidably impair the wrongdoer's capacity to contribute to the group.

Masters (1984) maintains that "imprisonment, enslavement and death are particularly important forms of ostracism." Already, in Section 4, we have discussed the relation of imprisonment and execution to ostracism. Although this expands the coverage of the term explosively, these forms of punishment fit well within our framework, because jailing or executing someone sacrifices aggregation economies by ending his contribution to the group. (Enslavement is different because the rest of the group may directly benefit from the services or sale of the deviator).

Ostracism can also take forms that are milder than forced exile. It may have the same incentive effects as exile without requiring a change of location: the other players might just be impolite or refuse to converse with the offender. Voluntary exile to avoid other punishments is also equivalent to ostracism. Recall that Socrates could easily have fled Athens to avoid drinking hemlock, and surprised his friends (and no doubt his enemies) by refusing to do so. Embezzlers in Bermuda and U.S. draft evaders in Canada are other examples.

Another application is to the problem of monitoring a group's behavior.

Suppose that the manager of a team of workers has available a costly technology for monitoring and enforcing cooperation (working rather than shirking, in this case). Ostracism is a mechanism that can enforce cooperation more cheaply. The manager need only state the Banishment strategy with "Inform the manager about Mr X" replacing "Blackball Mr X." The direct cost of informing the manager is very low, and if it is credible that the manager himself will carry out the costly punishment, cooperation within the team has been enforced by inexpensive self-monitoring.[5]

The exclusion of the member from the benefits of being in the group is the obvious aspect of ostracism. This paper stresses the obverse, that ostracizing a member sacrifices the benefits that he can provide the group. Indeed, in ancient Athens ostracism was applied to some of the most dynamic leaders, and Amsterdam lost an exceptionally gifted citizen when it banished Spinoza.

The ostracism model is not intended to apply to all repeated social dilemmas. The model applies when the group not only faces a repeated game, but also: (a) Members can be expelled from the group; (b) Players would prefer membership in the group even if everyone defects; and (c) If everyone defects, the per capita payoff does not vary with the number of members.

These three requirements rule out applying the model to many situations. The model does not, for example, fit the application of the Prisoners' Dilemma that may first come to mind: oligopoly. Oligopolists generally cannot expel a price cutter from the industry, except in markets where sellers must be certified by a regulatory agency controlled by the group.[6] Moreover, when a deviator can be expelled, the elimination of a competing seller would generally be beneficial to the remaining sellers. The problem is not to make ostracism credible, but to prevent unprovoked ostracism of players who did not defect.

Many other situations do fit the assumptions of the model. Trade sanctions are examples of ostracism, whether by housewives in Lake Wobegon against a grocery with misdated goods or Common Market countries against

a country such as South Africa for its race policies. Ostracism from world capital or goods markets can perhaps provide a clue as to why nations such as Mexico and Argentina are reluctant to brazenly default on their debts; without ostracism, it is a puzzle as to why they would be denied new loans just because of past misbehavior.

Ostracism is common in social groups, and we will cite only a few examples of this widespread phenomenon. Gruter (1985) describes Meidung (shunning), and excommunication, forms of ostracism practiced among the Old Order Amish. These arose as church commandments in 1632 from the Dordrecht Confession of Faith as a means of disciplining church members. Boehm (1985) discusses several forms of ostracism in Balkan tribal society in 19th century highland Albania. These range from refusal to talk or listen to an individual regarded as a coward to expulsion from the tribe, and finally execution. Because clans were obligated to unconditionally avenge wrongs to members, the society was prone to blood feuds. However, clan ostracism was sometimes performed on members who were so reckless as to be a liability to the group. Customarily, the clan was not held liable for the actions of an *odlicen* (expelled member).

Often no formal institutions for expulsion exist, but there are ways in which the group can pressure the deviant into leaving, or deny him the benefits of society. The latter has been suggested as a problem in experimental work on the Prisoner's Dilemma. A student subject deciding whether to defect against students living in the same dormitory may decide that exclusion from dorm parties may outweigh the experiment's monetary incentives. Ostracism of this kind is a basic part of our culture. Readers of Dickens' *Hard Times*, for example, will recall that even then, a deviant worker in an English trade union would be "sent to Coventry," meaning that no other worker would speak to him. And the name of Captain Boycott, a 19th century Irish land agent, entered the common vocabulary when his neighbors shunned him for cooperating with the English (Churchill [1958]).

Conclusion and Summary.

This paper has described a possible escape from the repeated multiplayer Prisoners' Dilemma and related games: ostracizing defectors, an escape that can enforce cooperation until the final round. Ostracism can be effective despite a perfectness problem in incurring costs to punish defectors after the defection has taken place. Our model can explain why ostracism, or social norms which call for censure of wrongdoers can be self-fulfilling in the sense that it pays for everyone to conform not only to good behavior, but also to punish wrongdoers, to punish those who fail to punish, and so on. Indeed, even socially dysfunctional norms can be supported by ostracism.

The key is that since all players defect in the final round, in that round no cost of ostracizing wrongdoers is incurred. The threat of punishment therefore deters both defection and failure to blackball properly in the next-to-last round, and the argument can be carried back to the start of the game. There exist equilibrium strategies involving cooperation at all rounds until the last. The equilibrium with a finite number of rounds is weak, since the decision on whether to blackball in the final round is a razor's-edge choice. However, if the slightest amount of morality is added, cooperation until the last round becomes a unique and strong equilibrium outcome.

		Player 2	
		Cooperate	Defect
	Cooperate	$15, 15$	$-10, 5$
Player 1:			
	Defect	$5, -10$	$0, 0$

Payoffs to: Player 1, Player 2

Figure 1: A Prisoners' Dilemma with Two Members. (missing)

		Other Players		
		All Customary	Some Desirable	All Desirable
Player i	Customary	$5(n-1)$	0	0
	Desirable	40	40	100

(*Payoffs to Player i*)

Figure 3: A Bad Equilibrium Enforced by Ostracism. (missing)

Notes

1. Roughly speaking, perfectness requires that an equilibrium strategy not only be a best response to the other players' strategies early in the game, but also remain a best response once the game has been partly played out. This rules out threats that would not be carried out.

2. One need not subscribe to the "social contract" theory to attempt to explain grouping behavior in terms of costs and benefits to individuals from being in a group. An evolutionary outlook, or Aristotle's view of man as a political animal, are both entirely consistent with grouping behavior being influenced by the costs and benefits individuals derive from membership.

3. One such subgame, for example, is the subgame starting after both players have cooperated the first two periods and defected in the third. Such behavior might never occur in equilibrium, but a player's equilibrium strategy must specify what actions he takes if it does occur, and those actions must maximize his payoffs for the remainder of the game.

4. An ostracism strategy simpler than *Banishment* would be to wait and blackball deviators only in phase T^{os}, even if they had deviated much earlier. Such a strategy could support cooperation, without the iterative blackballing rule of *Banishment*. But this wait-and-blackball strategy works under a narrower parameter range than *Banishment*, because the cost of being ostracized during phase T^{pd} must outweigh the total benefit from defecting in all previous rounds.

5. Thomas Schwartz of the UCLA Department of Political Science suggested this idea.

6. Kessel (1958), however, describes suspension of licenses, expulsion from medical societies, and denial of hospital staff privileges to doctors associated with price-cutting and group health plans.

References

Akerlof, George, 1976, The economics of caste and of the rat race and other woeful tales, Quarterly Journal of Economics, 599-617.

Akerlof, George, 1980, A theory of social custom, of which unemployment may be one consequence, Quarterly Journal of Economics, 94, 749-775.

Bendor, Jonathan and Dilip Mookherjee, 1987, Institutional Structure and the Logic of Ongoing Collective Action, American Political Science Review, 81, 129-154.

Boehm, Christopher, 1985, Execution within the clan as an extreme form of ostracism, Social Science Information, 24, 309-321.

Churchill, Winston, 1958, The Great Democracies, New York: Dodd, Mead and Co.

Fudenberg, Drew and Eric Maskin, 1986, The folk theorem in repeated games with discounting and with incomplete information, Econometrica, 54, 533-554.

Gruter, Margaret, 1985, Ostracism on trial: the limits of individual rights, Social Science Information 24, 101-111.

Hirshleifer, Jack, 1987, On the emotions as guarantors of threats and promises, in The Latest on the Best: Essays on Evolution and Optimality, John Dupre, ed. Cambridge, Mass.: MIT Press.

Kessel, Reuben, 1958, "Price discrimination in medicine." Journal of Law and Economics 1:20-53 .

Kreps, David, Paul Milgrom, John Roberts, and Robert Wilson, 1982, Rational cooperation in the finitely repeated prisoner's dilemma, Journal of Economic Theory, 27, 245-252.

Masters, Roger, 1984, Ostracism, voice and exit: the biology of social participation, Social Science Information, 23, 877-893.

Rasmusen, Eric, 1987, A new version of the Folk Theorem, UCLA AGSM Business Economics Working Paper # 87-6.

Rasmusen, Eric, forthcoming, Games and Information, Oxford: Basil Blackwell Ltd.

Romer, David, 1984, The theory of social custom: A modification and some extensions, Quarterly Journal of Economics, 99, 717-727.

Schelling, Thomas, 1960, The Strategy of Conflict, London: Oxford University Press.

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

Thompson, Earl and Roger Faith, 1981, A pure theory of strategic behavior and social institutions, American Economic Review, 71, 366-380, vol. 71 no. 3.