The Emergence of Cooperation among Egoists

ROBERT AXELROD
University of Michigan

This article investigates the conditions under which cooperation will emerge in a world of egoists without central authority. This problem plays an important role in such diverse fields as political philosophy, international politics, and economic and social exchange. The problem is formalized as an iterated Prisoner's Dilemma with pairwise interaction among a population of individuals.

Results from three approaches are reported: the tournament approach, the ecological approach, and the evolutionary approach. The evolutionary approach is the most general since all possible strategies can be taken into account. A series of theorems is presented which show: (1) the conditions under which no strategy can do any better than the population average if the others are using the reciprocal cooperation strategy of TIT FOR TAT, *(2) the necessary and sufficient conditions for a strategy to be collectively stable, and (3) how cooperation can emerge from a small cluster of discriminating individuals even when everyone else is using a strategy of unconditional defection.*

Under what conditions will cooperation emerge in a world of egoists without central authority? This question has played an important role in a variety of domains including political philosophy, international politics, and economic and social exchange. This article provides new results which show more completely than was previously possible the conditions under which cooperation will emerge. The results are more complete in two ways. First, *all* possible strategies are taken into account, not simply some arbitrarily selected subset. Second, not only are equilibrium conditions established, but also a mechanism is specified which can move a population from noncooperative to cooperative equilibrium.

The situation to be analyzed is the one in which narrow self-maximization behavior by each person leads to a poor outcome for all. This is the famous Prisoner's Dilemma game. Two individuals can each either cooperate or defect. No matter what the other does, defection yields a higher payoff than cooperation. But if both defect, both do worse than if both cooperated. Figure 1 shows the payoff matrix with sample utility numbers attached to the payoffs. If the other player cooperates, there is a choice between cooperation which yields R (the reward for mutual cooperation) or defection which yields T (the temptation to defect). By assumption, $T > R$, so it pays to defect if the other player cooperates. On the other hand, if the other player defects, there is a choice between cooperation which yields S (the sucker's payoff), or defection which yields P (the punishment for mutual defection). By assumption, $P > S$, so it pays to defect if the other player defects. Thus no matter what the other player does, it pays to defect. But if both defect, both get P rather than the R they could both have got if both had cooperated. But R is assumed to be greater than P. Hence the dilemma. Individual rationality leads to a worse outcome for both than is possible.

To insure that an even chance of exploitation or being exploited is not as good an outcome as mutual cooperation, a final inequality is added in the standard definition of the Prisoner's Dilemma. This is just $R > (T+S)/2$.

Thus two egoists playing the game once will both choose their dominant choice, defection, and get a payoff, P, which is worse for both than the R they could have got if they had both cooperated. If the game is played a known finite number of times, the players still have no incentive to

	Cooperate	Defect
Cooperate	R=3, R=3	S=0, T=5
Defect	T=5, S=0	P=1, P=1

$$T > R > P > S$$
$$R > (S+T)/2$$

Source: Robert Axelrod, "Effective Choice in the Prisoner's Dilemma," *Journal of Conflict Resolution* 24 (1980): 3-25; "More Effective Choice in the Prisoner's Dilemma," *Journal of Conflict Resolution* 24 (1980): 379-403.

Note: The payoffs to the row chooser are listed first.

Figure 1. A Prisoner's Dilemma

I would like to thank John Chamberlin, Michael Cohen, Bernard Grofman, William Hamilton, John Kingdon, Larry Mohr, John Padgett and Reinhard Selten for their help, and the Institute of Public Policy Studies for its financial support.

cooperate. This is certainly true on the last move since there is no future to influence. On the next-to-last move they will also have no incentive to cooperate since they can anticipate mutual defection on the last move. This line of reasoning implies that the game will unravel all the way back to mutual defection on the first move of any sequence of plays which is of known finite length (Luce and Raiffa, 1957, pp. 94-102). This reasoning does not apply if the players will interact an indefinite number of times. With an indefinite number of interactions, cooperation can emerge. This article will explore the precise conditions necessary for this to happen.

The importance of this problem is indicated by a brief explanation of the role it has played in a variety of fields.

1. *Political Philosophy.* Hobbes regarded the state of nature as equivalent to what we now call a two-person Prisoner's Dilemma, and he built his justification for the state upon the purported impossibility of sustained cooperation in such a situation (Taylor, 1976, pp. 98-116). A demonstration that mutual cooperation could emerge among rational egoists playing the iterated Prisoner's Dilemma would provide a powerful argument that the role of the state should not be as universal as some have argued.

2. *International Politics.* Today nations interact without central control, and therefore the conclusions about the requirements for the emergence of cooperation have empirical relevance to many central issues of international politics. Examples include many varieties of the security dilemma (Jervis, 1978) such as arms competition and its obverse, disarmament (Rapoport, 1960); alliance competition (Snyder, 1971); and communal conflict in Cyprus (Lumsden, 1973). The selection of the American response to the Soviet invasion of Afghanistan in 1979 illustrates the problem of choosing an effective strategy in the context of a continuing relationship. Had the United States been perceived as continuing business as usual, the Soviet Union might have been encouraged to try other forms of noncooperative behavior later. On the other hand, any substantial lessening of U.S. cooperation risked some form of retaliation which could then set off counter-retaliation, setting up a pattern of mutual defection that could be difficult to get out of. Much of the domestic debate over foreign policy is over problems of just this type.

3. *Economic and social exchange.* Our everyday lives contain many exchanges whose terms are not enforced by any central authority. Even in economic exchanges, business ethics are maintained by the knowledge that future interactions are likely to be affected by the outcome of the current exchange.

4. *International political economy.* Multinational corporations can play off host governments to lessen their tax burdens in the absence of coordinated fiscal policies between the affected governments. Thus the commodity exporting country and the commodity importing country are in an iterated Prisoner's Dilemma with each other, whether they fully appreciate it or not (Laver, 1977).

In the literatures of these areas, there has been a convergence on the nature of the problem to be analyzed. All agree that the two-person Prisoner's Dilemma captures an important part of the strategic interaction. All agree that what makes the emergence of cooperation possible is the possibility that interaction will continue. The tools of the analysis have been surprisingly similar, with game theory serving to structure the enterprise.

As a paradigm case of the emergence of cooperation, consider the development of the norms of a legislative body, such as the United States Senate. Each senator has an incentive to appear effective for his or her constituents even at the expense of conflicting with other senators who are trying to appear effective for *their* constituents. But this is hardly a zero-sum game since there are many opportunities for mutually rewarding activities between two senators. One of the consequences is that an elaborate set of norms, or folkways, have emerged in the Senate. Among the most important of these is the norm of reciprocity, a folkway which involves helping out a colleague and getting repaid in kind. It includes vote trading, but it extends to so many types of mutually rewarding behavior that "it is not an exaggeration to say that reciprocity is a way of life in the Senate" (Matthews, 1960, p. 100; see also Mayhew, 1974).

Washington was not always like this. Early observers saw the members of the Washington community as quite unscrupulous, unreliable, and characterized by "falsehood, deceit, treachery" (Smith, 1906, p. 190). But by now the practice of reciprocity is well established. Even the significant changes in the Senate over the last two decades toward more decentralization, more openness, and more equal distribution of power have come without abating the folkway of reciprocity (Ornstein, Peabody and Rhode, 1977). I will show that we do *not* need to assume that senators are more honest, more generous, or more public-spirited than in earlier years to explain how cooperation based on reciprocity has emerged and proven stable. The emergence of cooperation can be explained as a consequence of senators pursuing their own interests.

The approach taken here is to investigate how individuals pursuing their own interests will act, and then see what effects this will have for the system as a whole. Put another way, the approach is to make some assumptions about micro-motives, and then deduce consequences for macro-behavior (Schelling, 1978). Thinking about the paradigm case of a legislature is a convenience, but the same style of reasoning can apply to the emergence of cooperation between individuals in many other political settings, or even to relations between nations. While investigating the conditions which foster the emergence of cooperation, one should bear in mind that cooperation is not always socially desirable. There are times when public policy is best served by the prevention of cooperation—as in the need for regulatory action to prevent collusion between oligopolistic business enterprises.

The basic situation I will analyze involves pairwise interactions.[1] I assume that the player can recognize another player and remember how the two of them have interacted so far. This allows the history of the particular interaction to be taken into account by a player's strategy.

A variety of ways to resolve the dilemma of the Prisoner's Dilemma have been developed. Each involves allowing some additional activity which alters the strategic interaction in such a way as to fundamentally change the nature of the problem. The original problem remains, however, because there are many situations in which these remedies are not available. I wish to consider the problem in its fundamental form.

1. There is no mechanism available to the players to make enforceable threats or commitments (Schelling, 1960). Since the players cannot make commitments, they must take into account all possible strategies which might be used by the other player, and they have all possible strategies available to themselves.

2. There is no way to be sure what the other player will do on a given move. This eliminates the possibility of metagame analysis (Howard, 1971) which allows such options as "make the same choice as the other player is about to make." It also eliminates the possibility of reliable reputations such as might be based on watching the other player interact with third parties.

3. There is no way to change the other player's utilities. The utilities already include whatever consideration each player has for the interests of the other (Taylor, 1976, pp. 69-83).

Under these conditions, words not backed by actions are so cheap as to be meaningless. The players can communicate with each other only through the sequence of their own behavior. This is the problem of the iterated Prisoner's Dilemma in its fundamental form.

Two things remain to be specified: how the payoff of a particular move relates to the payoff in a whole sequence, and the precise meaning of a strategy. A natural way to aggregate payoffs over time is to assume that later payoffs are worth less than earlier ones, and that this relationship is expressed as a constant discount per move (Shubik, 1959, 1970). Thus the next payoff is worth only a fraction, w, of the same payoff this move. A whole string of mutual defection would then have a "present value" of $P + wP + w^2P + w^3P \ldots = P/(1-w)$. The discount parameter, w, can be given either of two interpretations. The standard economic interpretation is that later consumption is not valued as much as earlier consumption. An alternative interpretation is that future moves may not actually occur, since the interaction between a pair of players has only a certain probability of continuing for another move. In either interpretation, or a combination of the two, w is strictly between zero and one. The smaller w is, the less important later moves are relative to earlier ones.

For a concrete example, suppose one player is following the policy of always defecting, and the other player is following the policy of *TIT FOR TAT*. *TIT FOR TAT* is the policy of cooperating on the first move and then doing whatever the other player did on the previous move. This means that *TIT FOR TAT* will defect once for each defection by the other player. When the other player is using *TIT FOR TAT*, a player who always defects will get T on the first move, and P on all the subsequent moves. The payoff to someone using *ALL D* when playing with someone using *TIT FOR TAT* is thus:

$$\begin{aligned} V(ALL\ D \mid TFT) &= T + wP + w^2P + w^3P \ldots \\ &= T + wP\,(1 + w + w^2 \ldots) \\ &= T + wP/(1-w). \end{aligned}$$

Both *ALL D* and *TIT FOR TAT* are strategies. In general, a *strategy* (or decision rule) is a function from the history of the game so far into a probability of cooperation on the next move. Strategies can be stochastic, as in the example of a rule which is entirely random with equal probabilities of cooperation and defection on each move. A strategy can also be quite sophisticated in its use of the pattern of outcomes in the game so far to determine what to do next. It may, for ex-

[1]A single player may be interacting with many others, but the player is interacting with them one at a time. The situations which involve more than pairwise interaction can be modeled with the more complex n-person Prisoner's Dilemma (Olson, 1965; G. Hardin, 1968; R. Hardin, 1971; Schelling, 1973). The principal application is to the provision of collective goods. It is possible that the results from pairwise interactions will help suggest how to undertake a deeper analysis of the n-person case as well, but that must wait. For a parallel treatment of the two-person and n-person cases, see Taylor (1976, pp. 29-62).

ample, use Bayesian techniques to estimate the parameters of some model it might have of the other player's rule. Or it may be some complex combination of other strategies. But a strategy must rely solely on the information available through this one sequence of interactions with the particular player.

The first question one is tempted to ask is, "What is the best strategy?" This is a good question, but unfortunately there is no best rule independent of the environment which it might have to face. The reason is that what works best when the other player is unconditionally cooperative will not in general work well when the other player is conditionally cooperative, and vice versa. To prove this, I will introduce the concept of a *nice* strategy, namely, one which will never be the first to defect.

THEOREM 1. *If the discount parameter, w, is sufficiently high, there is no best strategy independent of the strategy used by the other player.*

PROOF. Suppose A is a strategy which is best regardless of the strategy used by the other player. This means that for any strategies A' and B, $V(A|B) \geqslant V(A'|B)$. Consider separately the cases where A is nice and A is not nice. If A is nice, let $A' = ALL\ D$, and let $B = ALL\ C$. Then $V(A|B) = R/(1-w)$ which is less than $V(A'|B) = T/(1-w)$. On the other hand, if A is not nice, let $A' = ALL\ C$, and let B be the strategy of cooperating until the other player defects and then always defects. Eventually A will be the first to defect, say, on move n. The value of n is irrelevant for the comparison to follow, so assume that $n = 1$. To give A the maximum advantage, assume that A always defects after its first defection. Then $V(A|B) = T + wP/(1-w)$. But $V(A'|B) = R/(1-w) = R+wR/(1-w)$. Thus $V(A|B) < V(A'|B)$ whenever $w > (T-R)/(T-P)$. Thus the immediate advantage gained by the defection of A will eventually be more than compensated for by the long-term disadvantage of B's unending defection, assuming that w is sufficiently large. Thus, if w is sufficiently large, there is no one best strategy.

In the paradigm case of a legislature, this theorem says that if there is a large enough chance that a member of Congress will interact *again* with another member of Congress, then there is no one best strategy to use independently of the strategy being used by the other person. It would be best to be cooperative with someone who will reciprocate that cooperation in the future, but not with someone whose future behavior will not be very much affected by this interaction (see, for example, Hinckley, 1972). The very possibility of achieving stable mutual cooperation depends upon there being a good chance of a continuing interaction, as measured by the magnitude of w. Empirically, the chance of two members of Congress having a continuing interaction has increased dramatically as the biennial turnover rates in Congress have fallen from about 40 percent in the first 40 years of the Republic to about 20 percent or less in recent years (Young, 1966, pp. 87-90; Jones, 1977, p. 254; Patterson, 1978, pp. 143-44). That the increasing institutionalization of Congress has had its effects on the development of congressional norms has been widely accepted (Polsby, 1968, esp. n. 68). We now see how the diminished turnover rate (which is one aspect of institutionalization) can allow the development of reciprocity (which is one important part of congressional folkways).

But saying that a continuing chance of interaction is necessary for the development of cooperation is not the same as saying that it is sufficient. The demonstration that there is not a single best strategy still leaves open the question of what patterns of behavior can be expected to emerge when there actually is a sufficiently high probability of continuing interaction between two people.

The Tournament Approach

Just because there is no single best decision rule does not mean analysis is hopeless. For example, progress can be made on the question of which strategy does best in an environment of players who are also using strategies designed to do well. To explore this question, I conducted a tournament of strategies submitted by game theorists in economics, psychology, sociology, political science and mathematics (Axelrod, 1980a). Announcing the payoff matrix shown in Figure 1, and a game length of 200 moves, I ran the 14 entries and *RANDOM* against each other in a round robin tournament. The result was that the highest average score was attained by the simplest of all the strategies submitted, *TIT FOR TAT*.

I then circulated the report of the results and solicited entries for a second round. This time I received 62 entries from six countries.[2] Most of the contestants were computer hobbyists, but there were also professors of evolutionary biology,

[2]In the second round, the length of the games was uncertain, with an expected median length of 200 moves. This was achieved by setting the probability that a given move would not be the last one at $w = .99654$. As in the first round, each pair was matched in five games. See Axelrod (1980b) for a complete description.

physics, and computer science, as well as the five disciplines represented in the first round. *TIT FOR TAT* was again submitted by the winner of the first round, Anatol Rapoport of the Institute for Advanced Study (Vienna). And it won again. An analysis of the 3,000,000 choices which were made in the second round shows that *TIT FOR TAT* was a very robust rule because it was nice, provocable into a retaliation by a defection of the other, and yet forgiving after it took its one retaliation (Axelrod, 1980b).

The Ecological Approach

To see if *TIT FOR TAT* would do well in a whole series of simulated tournaments, I calculated what would happen if each of the strategies in the second round were submitted to a hypothetical next round in proportion to its success in the previous round. This process was then repeated to generate the time path of the distribution of strategies. The results showed that as the less-successful rules were displaced, *TIT FOR TAT* continued to do well with the rules which initially scored near the top. In the long run, *TIT FOR TAT* displaced all the other rules and went to what biologists call fixation (Axelrod, 1980b).

This is an ecological approach because it takes as given the varieties which are present and investigates how they do over time when interacting with each other. It provides further evidence of the robust nature of the success of *TIT FOR TAT*.

The Evolutionary Approach

A much more general approach would be to allow all possible decision rules to be considered, and to ask what are the characteristics of the decision rules which are stable in the long run. An evolutionary approach recently introduced by biologists offers a key concept which makes such an analysis tractable (Maynard Smith, 1974, 1978). This approach imagines the existence of a whole population of individuals employing a certain strategy, B, and a single mutant individual employing another strategy, A. Strategy A is said to *invade* strategy B if $V(A|B) > V(B|B)$ where $V(A|B)$ is the expected payoff an A gets when playing a B, and $V(B|B)$ is the expected payoff a B gets when playing another B. Since the B's are interacting virtually entirely with other B's, the concept of invasion is equivalent to the single mutant individual being able to do better than the population average. This leads directly to the key concept of the evolutionary approach. A strategy is *collectively stable* if no strategy can invade it.[3]

The biological motivation for this approach is based on the interpretation of the payoffs in terms of fitness (survival and fecundity). All mutations are possible, and if any could invade a given population it would have had the chance to do so. Thus only a collectively stable strategy is expected to be able to maintain itself in the long-run equilibrium as the strategy used by all.[4] Collectively stable strategies are important because they are the only ones which an entire population can maintain in the long run if mutations are introduced one at a time.

The political motivation for this approach is based on the assumption that all strategies are possible, and that if there were a strategy which would benefit an individual, someone is sure to try it. Thus only a collectively stable strategy can maintain itself as the strategy used by all—provided that the individuals who are trying out novel strategies do not interact too much with one another.[5] As we shall see later, if they do interact in clusters, then new and very important developments become possible.

A difficulty in the use of this concept of collective stability is that it can be very hard actually to determine which strategies have it and which do not. In most biological applications, including both the Prisoner's Dilemma and other types of interactions, this difficulty has been dealt with in

[3]Those familiar with the concepts of game theory will recognize this as a strategy being in Nash equilibrium with itself. My definitions of invasion and collective stability are slightly different from Maynard Smith's (1974) definitions of invasion and evolutionary stability. His definition of invasion allows $V(A|B) = V(B|B)$ provided that $V(B|A) > V(A|A)$. I have used the new definitions to simplify the proofs and to highlight the difference between the effect of a single mutant and the effect of a small number of mutants. Any rule which is evolutionarily stable is also collectively stable. For a nice rule, the definitions are equivalent. All theorems in the text remain true if "evolutionary stability" is substituted for "collective stability" with the exception of Theorem 3, where the characterization is necessary but no longer sufficient.

[4]For the application of the results of this article to biological contexts, see Axelrod and Hamilton (1981). For the development in biology of the concepts of reciprocity and stable strategy, see Hamilton (1964), Trivers (1971), Maynard Smith (1974), and Dawkins (1976).

[5]Collective stability can also be interpreted in terms of a commitment by one player, rather than the stability of a whole population. Suppose player Y is committed to using strategy B. Then player X can do no better than use this same strategy B if and only if strategy B is collectively stable.

one of two ways. One method has been to restrict the analysis to situations where the strategies take some particularly simple form such as in one-parameter models of sex-ratios (Hamilton, 1967). The other method has been to restrict the strategies themselves to a relatively narrow set so that some illustrative results could be attained (Maynard Smith and Price, 1973; Maynard Smith, 1978).

The difficulty of dealing with all possible strategies was also faced by Michael Taylor (1976), a political scientist who sought a deeper understanding of the issues raised by Hobbes and Hume concerning whether people in a state of nature would be expected to cooperate with each other. He too employed the method of using a narrow set of strategies to attain some illustrative results. Taylor restricted himself to the investigation of four particular strategies (including *ALL D*, *ALL C*, *TIT FOR TAT*, and the rule "cooperate until the other defects, and then always defect"), and one set of strategies which retaliates in progressively increasing numbers of defections for each defection by the other. He successfully developed the equilibrium conditions when these are the only rules which are possible.[6]

Before running the computer tournament I believed that it was impossible to make very much progress if all possible strategies were permitted to enter the analysis. However, once I had attained sufficient experience with a wide variety of specific decision rules, and with the analysis of what happened when these rules interacted, I was able to formulate and answer some important questions about what would happen if all possible strategies were taken into account.

The remainder of this article will be devoted to answering the following specific questions about the emergence of cooperation in the iterated Prisoner's Dilemma:

1. Under what conditions is *TIT FOR TAT* collectively stable?
2. What are the necessary and sufficient conditions for any strategy to be collectively stable?
3. If virtually everyone is following a strategy of unconditional defection, when can cooperation emerge from a small cluster of newcomers who introduce cooperation based on reciprocity?

[6]For the comparable equilibrium conditions when all possible strategies are allowed, see below, Theorems 2 through 6. For related results on the potential stability of cooperative behavior, see Luce and Raiffa (1957, p. 102), Kurz (1977) and Hirshleifer (1978).

TIT FOR TAT as a Collectively Stable Strategy

TIT FOR TAT cooperates on the first move, and then does whatever the other player did on the previous move. This means that any rule which starts off with a defection will get T, the highest possible payoff, on the first move when playing *TIT FOR TAT*. Consequently, *TIT FOR TAT* can only avoid being invadable by such a rule if the game is likely to last long enough for the retaliation to counteract the temptation to defect. In fact, no rule can invade *TIT FOR TAT* if the discount parameter, w, is sufficiently large. This is the heart of the formal result contained in the following theorem. Readers who wish to skip the proofs can do so without loss of continuity.

THEOREM 2. TIT FOR TAT *is a collectively stable strategy if and only if* $w \geq \max\left(\frac{T-R}{T-P}, \frac{T-R}{R-S}\right)$. *An alternative formulation of the same result is that* TIT FOR TAT *is a collectively stable strategy if and only if it is invadable neither by* ALL D *nor the strategy which alternates defection and cooperation.*

PROOF. First we prove that the two formulations of the theorem are equivalent, and then we prove both implications of the second formulation. To say that *ALL D* cannot invade *TIT FOR TAT* means that $V(ALL\ D|TFT) \leq V(TFT|TFT)$. As shown earlier, $V(ALL\ D|TFT) = T + wP/(1-w)$. Since *TFT* always cooperates with its twin, $V(TFT|TFT) = R + wR + w^2R \ldots = R/(1-w)$. Thus *ALL D* cannot invade *TIT FOR TAT* when $T + wP/(1-w) \leq R/(1-w)$, or $T(1-w) + wP \leq R$, or $T-R \leq w(T-P)$ or $w \geq \frac{T-R}{T-P}$. Similarly, to say that alternation of D and C cannot invade *TIT FOR TAT* means that $(T+wS)/(1-w^2) \leq R/(1-w)$, or $\frac{T-R}{R-S} \leq w$. Thus $w \geq \frac{T-R}{T-P}$ and $w \geq \frac{T-R}{R-S}$ is equivalent to saying that *TIT FOR TAT* is invadable by neither *ALL D* nor the strategy which alternates defection and cooperation. This shows that the two formulations are equivalent.

Now we prove both of the implications of the second formulation. One implication is established by the simple observation that if *TIT FOR TAT* is a collectively stable strategy, then no rule can invade, and hence neither can the two specified rules. The other implication to be proved is that if neither *ALL D* nor Alternation of D and C can invade *TIT FOR TAT*, then no strategy can.

TIT FOR TAT has only two states, depending on what the other player did the previous move (on the first move it assumes, in effect, that the other player has just cooperated). Thus if *A* is interacting with *TIT FOR TAT*, the best which any strategy, *A*, can do after choosing *D* is to choose *C* or *D*. Similarly, the best *A* can do after choosing *D* is to choose *C* or *D*. This leaves four possibilities for the best *A* can do with *TIT FOR TAT:* repeated sequences of *CC, CD, DC,* or *DD*. The first does the same as *TIT FOR TAT* does with another *TIT FOR TAT*. The second cannot do better than both the first and the third. This implies if the third and fourth possibilities cannot invade *TIT FOR TAT*, then no strategy can. These two are equivalent, respectively, to Alternation of *D* and *C*, and *ALL D*. Thus if neither of these two can invade *TIT FOR TAT*, no rule can, and *TIT FOR TAT* is a collectively stable strategy.

The significance of this theorem is that it demonstrates that if everyone in a population is cooperating with everyone else because each is using the *TIT FOR TAT* strategy, no one can do better using any other strategy *provided* the discount parameter is high enough. For example, using the numerical values of the payoff parameters given in Figure 1, *TIT FOR TAT* is uninvadable when the discount parameter, w, is greater than 2/3. If w falls below this critical value, and everyone else is using *TIT FOR TAT*, it will pay to defect on alternative moves. For w less than 1/2, *ALL D* can also invade.

One specific implication is that if the other player is unlikely to be around much longer because of apparent weakness, then the perceived value of w falls and the reciprocity of *TIT FOR TAT* is no longer stable. We have Caesar's explanation of why Pompey's allies stopped cooperating with him. "They regarded his [Pompey's] prospects as hopeless and acted according to the common rule by which a man's friends become his enemies in adversity" (trans. by Warner, 1960, p. 328). Another example is the business institution of the factor who buys a client's accounts receivable. This is done at a very substantial discount when the firm is in trouble because

> once a manufacturer begins to go under, even his best customers begin refusing payment for merchandise, claiming defects in quality, failure to meet specifications, tardy delivery, or what-have-you. The great enforcer of morality in commerce is the continuing relationship, the belief that one will have to do business again with this customer, or this supplier, and when a failing company loses this automatic enforcer, not even a strong-arm factor is likely to find a substitute (Mayer, 1974, p. 280).

Similarly, any member of Congres who is perceived as likely to be defeated in the next election may have some difficulty doing legislative business with colleagues on the usual basis of trust and good credit.[7]

There are many other examples of the importance of long-term interaction for the stability of cooperation in the iterated Prisoner's Dilemma. It is easy to maintain the norms of reciprocity in a stable small town or ethnic neighborhood. Conversely, a visiting professor is likely to receive poor treatment by other faculty members compared to the way these same prople treat their regular colleagues.

Another consequence of the previous theorem is that if one wants to prevent rather than promote cooperation, one should keep the same individuals from interacting too regularly with each other. Consider the practice of the government selecting two aerospace companies for competitive development contracts. Since firms specialize to some degree in either air force or in navy planes, there is a tendency for firms with the same specialties to be frequently paired in the final competition (Art, 1968). To make tacit collusion between companies more difficult, the government should seek methods of compensating for the specialization. Pairs of companies which shared a specialization would then expect to interact less often in the final competitions. This would cause the later interactions between them to be worth relatively less than before, reducing the value of w. If w is sufficiently low, reciprocal cooperation in the form of tacit collusion ceases to be a stable policy.

Knowing when *TIT FOR TAT* cannot be invaded is valuable, but it is only a part of the story. Other strategies may be, and in fact are, also collectively stable. This suggests the question of what a strategy has to do to be collectively stable. In other words, what policies, if adopted by everyone, will prevent any one individual from benefiting by a departure from the common strategy?

The Characterization of Collectively Stable Strategies

The characterization of all collectively stable strategies is based on the idea that invasion can be prevented if the rule can make the potential invader worse off than if it had just followed

[7]A countervailing consideration is that a legislator in electoral trouble may receive help from friendly colleagues who wish to increase the chances of reelection of someone who has proven in the past to be cooperative, trustworthy, and effective. Two current examples are Morris Udall and Thomas Foley. (I wish to thank an anonymous reviewer for this point.)

the common strategy. Rule B can prevent invasion by rule A if B can be sure that no matter what A does later, B will hold A's total score low enough. This leads to the following useful definition: B has a *secure position* over A on move n if no matter what A does from move n onwards, $V(A|B) < V(B|B)$, assuming that B defects from move n onwards. Let $V_n(A|B)$ represent A's discounted cumulative score in the moves before move n. Then another way of saying that B has a secure position over A on move n is that

$$V_n(A|B) + w^{n-1}P/(1-w) < V(B|B),$$

since the best A can do from move n onwards if B defects is get P each time. Moving the second term to the right side of the inequality gives the helpful result that B has a secure position over A on move n if $V_n(A|B)$ is small enough, namely if

$$V_n(A|B) < V(B|B) - w^{n-1}P/(1-w). \qquad (1)$$

The theorem which follows embodies the advice that if you want to employ a collectively stable strategy, you should only cooperate when you can afford an exploitation by the other side and still retain your secure position.

THEOREM 3. THE CHARACTERIZATION THEOREM. B *is a collectively stable strategy if and only if* B *defects on move* n *whenever the other player's cumulative score so far is too great, specifically when* $V_n(A|B) > V(B|B) - w^{n-1}(T+wP/(1-w))$.

PROOF. First it will be shown that a strategy B which defects as required will always have a secure position over any A, and therefore will have $V(A|B) < V(B|B)$ which in turn makes B a collectively stable strategy. The proof works by induction. For B to have a secure position on move 1 means that $V(A|ALL\ D) < V(ALL\ D|ALL\ D)$ according to the definition of secure position applied to n=1. Since this is true for all A, B has a secure position on move 1. If B has a secure position over A on move n, it has a secure position on move n+1. This is shown in two parts.

First, if B defects on move n, A gets at most P, so

$$V_{n+1}(A|B) < V_n(A|B) + w^{n-1}P.$$

Using Equation (1) gives:

$$V_{n+1}(A|B) < V(B|B) - w^{n-1}P/(1-w) + w^{n-1}P$$

$$V_{n+1}(A|B) < V(B|B) - w^nP/(1-w).$$

Second, B will only cooperate on move n when

$$V_n(A|B) < V(B|B) - w^{n-1}(T + wP/(1-w)).$$

Since A can get at most T on move n, we have

$$V_{n+1}(A|B) < V(B|B) - w^{n-1}(T + wP/(1-w)) + w^{n-1}T$$

$$V_{n+1}(A|B) < V(B|B) - w^nP/(1-w).$$

Therefore, B always has a secure position over A, and consequently B is a collectively stable strategy.

The second part of the proof operates by contradiction. Suppose that B is a collectively stable strategy and there is an A and an n such that B does not defect on move n when

$$V_n(A|B) > V(B|B) - w^{n-1}(T + wP/(1-w)),$$

i.e., when

$$V_n(A|B) + w^{n-1}(T + wP/(1-w)) > V(B|B). \qquad (2)$$

Define A' as the same as A on the first $n-1$ moves, and D thereafter. A' gets T on move n (since B cooperated then), and at least P thereafter. So,

$$V(A'|B) \geq V_n(A|B) + w^{n-1}(T + wP/(1-w)).$$

Combined with (2) this gives $V(A'|B) > V(B|B)$. Hence A' invades B, contrary to the assumption that B is a collectively stable strategy. Therefore, if B is a collectively stable strategy, it must defect when required.

Figure 2 illustrates the use of this theorem. The dotted line shows the value which must not be exceeded by any A if B is to be a collectively stable strategy. This value is just $V(B|B)$, the expected payoff attained by a player using B when virtually all the other players are using B as well. The solid curve represents the critical value of A's cumulative payoff so far. The theorem simply says that B is a collectively stable strategy if and only if it defects whenever the other player's cumulative value so far in the game is above this line. By doing so, *B* is able to prevent the other player from eventually getting a total expected value of more than rule *B* gets when playing another rule *B*.

The characterization theorem is "policy-relevant" in the abstract sense that it specifies what a strategy, *B*, has to do at any point in time as a function of the previous history of the interaction in order for *B* to be a collectively stable strategy.[8] It is a complete characterization because this requirement is both a necessary and a sufficient condition for strategy *B* to be collectively stable.

Two additional consequences about collectively stable strategies can be seen from Figure 2. First, as long as the other player has not accumulated too great a score, a strategy has the flexibility to either cooperate or defect and still be collectively stable. This flexibility explains why there are typically many strategies which are collectively stable. The second consequence is that a nice rule (one which will never defect first) has the most flexibility since it has the highest possible score when playing an identical rule. Put another way, nice rules can afford to be more generous than other rules with potential invaders because nice rules do so well with each other.

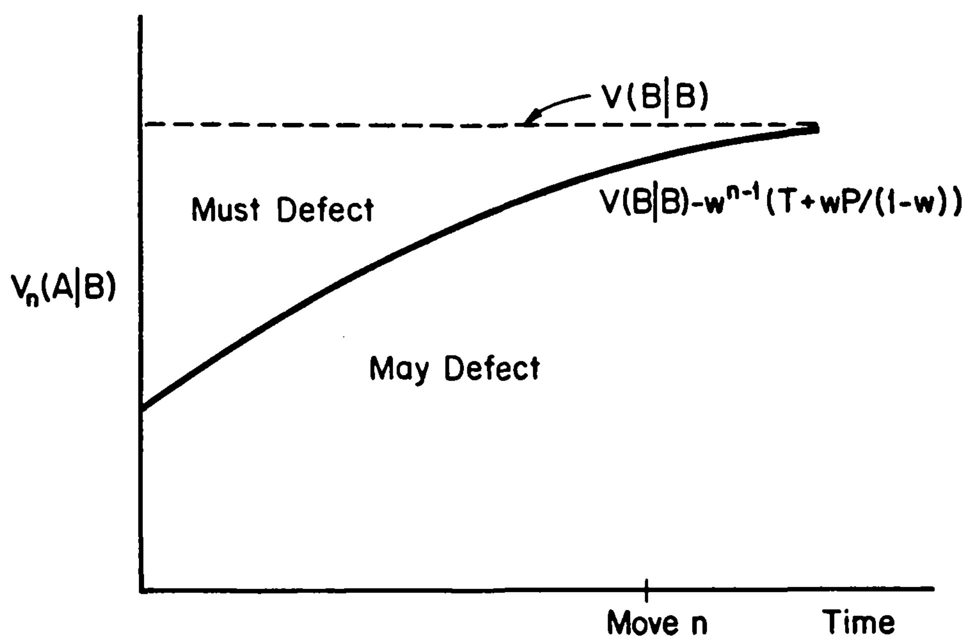
The flexibility of a nice rule is not unlimited, however, as shown by the following theorem. In fact, a nice rule must be *provoked* by the very first defection of the other player, i.e., on some later move the rule must have a finite chance of retaliating with a defection of its own.

THEOREM 4. *For a nice strategy to be collectively stable, it must be provoked by the first defection of the other player.*

PROOF. If a nice strategy were not provoked by a defection on move *n*, then it would not be collectively stable because it could be invaded by a rule which defected only on move *n*.

Besides provocability, there is another requirement for a nice rule to be collectively stable. This requirement is that the discount parameter *w*, be sufficiently large. This is a generalization of the second theorem, which showed that for *TIT FOR TAT* to be collectively stable, *w* has to be large enough. The idea extends beyond just nice rules to any rule which might be the first to cooperate.

[8]To be precise, $V(B|B)$ must also be specified in advance. For example, if B is never the first to defect, $V(B|B) = R/(1-w)$.



Source: Drawn by the author.

Figure 2. Characterization of a Collectively Stable Strategy

THEOREM 5. *Any rule,* B, *which may be the first to cooperate is collectively stable only when* w *is sufficiently large.*

PROOF. If B cooperates on the first move, $V(ALL\ D|B) \geqslant T + wP/(1-w)$. But for any B, $R/(1-w) \geqslant V(B|B)$ since R is the best B can do with another B by the assumptions that $R>P$ and $R>(S+T)/2$. Therefore $V(ALL\ D|B) > V(B|B)$ is so whenever $T + wP/(1-w) > R/(1-w)$. This implies that $ALL\ D$ invades a B which cooperates on the first move whenever $w<\frac{T-R}{T-P}$. If B has a positive chance of cooperating on the first move, then the gain of $V(ALL\ D|B)$ over $V_1(B|B)$ can only be nullified if w is sufficiently large. Likewise, if B will not be the first to cooperate until move n, $V_n(ALL\ D|B) = V_n(B|B)$ and the gain of $V_{n+1}(ALL\ D|B)$ over $V_{n+1}(B|B)$ can only be nullified if w is sufficiently large.

There is one strategy which is *always* collectively stable, that is regardless of the value of w or the payoff parameters T, R, P, and S. This is $ALL\ D$, the rule which defects no matter what.

THEOREM 6. ALL D *is always collectively stable.*

PROOF. $ALL\ D$ is always collectively stable because it always defects and hence it defects whenever required by the condition of the characterization theorem.

This is an important theorem because of its implications for the evolution of cooperation. If we imagine a system starting with individuals who cannot be enticed to cooperate, the collective stability of $ALL\ D$ implies that no single individual can hope to do any better than just to go along and be uncooperative as well. A world of "meanies" can resist invasion by anyone using any other strategy—provided that the newcomers arrive one at a time.

The problem, of course, is that a single newcomer in such a mean world has no one who will reciprocate any cooperation. If the newcomers arrive in small clusters, however, they will have a chance to thrive. The next section shows how this can happen.

The Implications of Clustering

To consider arrival in clusters rather than singly, we need to broaden the idea of "invasion" to include invasion by a cluster.[9] As before, we will suppose that strategy B is being used by virtually everyone. But now suppose that a small group of individuals using strategy A arrives and interacts with both the other A's and the native B's. To be specific, suppose that the proportion of the interactions by someone using strategy A with another individual using strategy A is p. Assuming that the A's are rare relative to the B's, virtually all the interactions of B's are with other B's. Then the average score of someone using A is $pV(A|A) + (1-p)V(A|B)$ and the average score of someone using B is $V(B|B)$. Therefore, a p*-cluster of* A *invades* B if $pV(A|A) + (1-p)V(A|B) > V(B|B)$, where p is the proportion of the interactions by a player using strategy A with another such player. Solving for p, this means that invasion is possible if the newcomers interact enough with each other, namely when

$$p > \frac{V(B|B) - V(A|B)}{V(A|A) - V(A|B)}. \qquad (3)$$

Notice that this assumes that pairing in the interactions is not random. With random pairing, an A would rarely meet another A. Instead, the clustering concept treats the case in which the A's are a trivial part of the environment of the B's, but a nontrivial part of the environment of the other A's.

The striking thing is just how easy invasion of $ALL\ D$ by clusters can be. Specifically, the value of p which is required for invasion by *TIT FOR TAT* of a world of $ALL\ D$'s is surprisingly low. For example, suppose the payoff values are those of Figure 1, and that $w = .9$, which corresponds to a 10-percent chance two interacting players will never meet again. Let A be *TIT FOR TAT* and B be $ALL\ D$. Then $V(B|B) = P/(1-w) = 10$; $V(A|B) = S + wP/(1-w) = 9$; and $V(A|A) = R/(1-w) = 30$. Plugging these numbers into Equation (3) shows that a p-cluster of *TIT FOR TAT* invades $ALL\ D$ when $p > 1/21$. Thus if the newcomers using *TIT FOR TAT* have any more than about 5 percent of their interactions with others using *TIT FOR TAT*, they can thrive in a world in which everyone else refuses ever to cooperate.

In this way, a world of meanies can be invaded by a cluster of *TIT FOR TAT*—and rather easily at that. To illustrate this point, suppose a business school teacher taught a class to give cooperative behavior a chance, and to reciprocate cooperation

[9]For related concepts from biology, see Wilson (1979) and Axelrod and Hamilton (1981).

from other firms. If the students did, and if they did not disperse too widely (so that a sufficient proportion of their interactions were with others from the same class), then the students would find that their lessons paid off.

When the interactions are expected to be of longer duration (or the time discount factor is not as great), then even less clustering is necessary. For example, if the median game length is 200 moves (corresponding to $w = .99654$) and the payoff parameters are as given in Figure 1, even one interaction out of a thousand with a like-minded follower of *TIT FOR TAT* is enough for the strategy to invade a world of *ALL D*'s. Even when the median game length is only two moves ($w = .5$), anything over a fifth of the interactions by the *TIT FOR TAT* players with like-minded types is sufficient for invasion to succeed and cooperation to emerge.

The next result shows which strategies are the most efficient at invading *ALL D* with the least amount of clustering. These are the strategies which are best able to discriminate between themselves and *ALL D*. A strategy is *maximally discriminating* if it will eventually cooperate even if the other has never cooperated yet, and once it cooperates it will never cooperate again with *ALL D* but will always cooperate with another player using the same strategy.

THEOREM 7. *The strategies which can invade* ALL D *in a cluster with the smallest value of* p *are those which are maximally discriminating, such as* TIT FOR TAT.

PROOF. To be able to invade *ALL D*, a rule must have a positive chance of cooperating first. Stochastic cooperation is not as good as deterministic cooperation with another player using the same rule since stochastic cooperation yields equal probability of S and T, and $(S+T)/2<R$ in the Prisoner's Dilemma. Therefore, a strategy which can invade with the smallest p must cooperate first on some move, n, even if the other player has never cooperated yet. Employing Equation (3) shows that the rules which invade B=*ALL D* with the lowest value of p are those which have the lowest value of p^*, where $p^* = [V(B|B) - V(A|B)]/[V(A|A) - V(A|B)]$. The value of p^* is minimized when $V(A|A)$ and $V(A|B)$ are maximized (subject to the constraint that A cooperates for the first time on move n) since $V(A|A) > V(B|B) > V(A|B)$. $V(A|A)$ and $V(A|B)$ are maximized subject to this constraint if and only if A is a maximally discriminating rule. (Incidentally, it does not matter for the minimal value of p just when A starts to cooperate.) *TIT FOR TAT* is such a strategy because it always cooperates for n=1, it cooperates only once with *ALL D*, and it always cooperates with another *TIT FOR TAT*.

The final theorem demonstrates that nice rules (those which never defect first) are actually better able than other rules to protect themselves from invasion by a cluster.

THEOREM 8. *If a nice strategy cannot be invaded by a single individual, it cannot be invaded by any cluster of individuals either.*

PROOF. For a cluster of rule A to invade a population of rule B, there must be a $p \leq 1$ such that $pV(A|A) + (1-p)V(A|B) > V(B|B)$. But if B is nice, then $V(A|A) \leq V(B|B)$. This is so because $V(B|B) = R/(1-w)$, which is the largest value attainable when the other player is using the same strategy. It is the largest value since $R>(S+T)/2$. Since $V(A|A) \leq V(B|B)$, A can invade as a cluster only if $V(A|B) > V(B|B)$. But that is equivalent to A invading as an individual.

This shows that nice rules do not have the structural weakness displayed in *ALL D*. *ALL D* can withstand invasion by any strategy, as long as the players using other strategies come one at a time. But if they come in clusters (even in rather small clusters), *ALL D* can be invaded. With nice rules, the situation is different. If a nice rule can resist invation by other rules coming one at a time, then it can resist invasion by clusters, no matter how large. So nice rules can protect themselves in a way that *ALL D* cannot.[10]

In the illustrative case of the Senate, Theorem 8 demonstrates that once cooperation based on reciprocity has become established, it can remain stable even if a cluster of newcomers does not respect this senatorial folkway. And Theorem 6 has shown that without clustering (or some comparable mechanism) the original pattern of mutual "treachery" could not have been overcome. Perhaps these critical early clusters were based on the boardinghouse arrangements in the capital during the Jeffersonian era (Young, 1966). Or perhaps the state delegations and state party delegations were more critical (Bogue and Marlaire, 1975). But now that the pattern of reciprocity is established, Theorems 2 and 5 show that it is collec-

[10]This property is possessed by population mixes of nice rules as well. If no single individual can invade a population of nice rules, no cluster can either.

tively stable, as long as the biennial turnover rate is not too great.

Thus cooperation can emerge even in a world of unconditional defection. The development cannot take place if it is tried only by scattered individuals who have no chance to interact with each other. But cooperation can emerge from small clusters of discriminating individuals, as long as these individuals have even a small proportion of their interactions with each other. Moreover, if nice strategies (those which are never the first to defect) eventually come to be adopted by virtually everyone, then those individuals can afford to be generous in dealing with any others. The population of nice rules can also protect themselves against clusters of individuals using any other strategy just as well as they can protect themselves against single individuals. But for a nice strategy to be stable in the collective sense, it must be provocable. So mutual cooperation can emerge in a world of egoists without central control, by starting with a cluster of individuals who rely on reciprocity.

References

Art, Robert J. (1969). *The TFX Decision: McNamara and the Military.* Boston: Little, Brown.

Axelrod, Robert (1980a). "Effective Choice in the Prisoner's Dilemma." *Journal of Conflict Resolution* 24: 3-25.

——— (1980b). "More Effective Choice in the Prisoner's Dilemma." *Journal of Conflict Resolution* 24: 379-403.

———, and William D. Hamilton (1981). "The Evolution of Cooperation." *Science* 2: 1390-96.

Bogue, Allan G., and Mark Paul Marlaire (1975). "Of Mess and Men: The Boardinghouse and Congressional Voting, 1821-1842." *American Journal of Political Science* 19: 207-30.

Dawkins, Richard (1976). *The Selfish Gene.* New York: Oxford University Press.

Hamilton, William D. (1964). "The Genetical Theory of Social Behavior (I and II)." *Journal of Theoretical Biology* 7: 1-16; 17-32.

——— (1967). "Extraordinary Sex Ratios." *Science* 156: 477-88.

Hardin, Garrett (1968). "The Tragedy of the Commons." *Science* 162: 1243-48.

Hardin, Russell (1971). "Collective Action as an Agreeable *n*-Prisoner's Dilemma." *Behavioral Science* 16: 472-81.

Hinckley, Barbara (1972). "Coalitions in Congress: Size and Ideological Distance." *Midwest Journal of Political Science* 26: 197-207.

Hirshleifer, J. (1978). "Natural Economy versus Political Economy." *Journal of Social and Biological Structures* 1: 319-37.

Howard, Nigel (1971). *Paradoxes of Rationality: Theory of Metagames and Political Behavior.* Cambridge, Mass.: MIT Press.

Jervis, Robert (1978). "Cooperation Under the Security Dilemma." *World Politics* 30: 167-214.

Jones, Charles O. (1977). "Will Reform Change Congress?" In Lawrence C. Dodd and Bruce I. Oppenheimer (eds.), *Congress Reconsidered.* New York: Praeger.

Kurz, Mordecai (1977). "Altruistic Equilibrium." In Bela Balassa and Richard Nelson (eds.), *Economic Progress, Private Values, and Public Policy.* Amsterdam: North Holland.

Laver, Michael (1977). "Intergovernmental Policy on Multinational Corporations, A Simple Model of Tax Bargaining." *European Journal of Political Research* 5: 363-80.

Luce, R. Duncan, and Howard Raiffa (1957). *Games and Decisions.* New York: Wiley.

Lumsden, Malvern (1973). "The Cyprus Conflict as a Prisoner's Dilemma." *Journal of Conflict Resolution* 17: 7-32.

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

Mayer, Martin (1974). *The Bankers.* New York: Ballantine Books.

Mayhew, David R. (1975). *Congress: The Electoral Connection.* New Haven, Conn.: Yale University Press.

Maynard Smith, John (1974). "The Theory of Games and the Evolution of Animal Conflict." *Journal of Theoretical Biology* 47: 209-21.

——— (1978). "The Evolution of Behavior." *Scientific American* 239: 176-92.

———, and G. R. Price (1973). "The Logic of Animal Conflict." *Nature* 246: 15-18.

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

Ornstein, Norman, Robert L. Peabody, and David W. Rhode (1977). "The Changing Senate: From the 1950s to the 1970s." In Lawrence C. Dodd and Bruce I. Oppenheimer (eds.), *Congress Reconsidered.* New York: Praeger.

Patterson, Samuel (1978). "The Semi-Sovereign Congress." In Anthony King (ed.), *The New American Political System.* Washington, D.C.: American Enterprise Institute.

Polsby, Nelson (1968). "The Institutionalization of the U.S. House of Representatives." *American Political Science Review* 62: 144-68.

Rapoport, Anatol (1960). *Fights, Games, and Debates.* Ann Arbor: University of Michigan Press.

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

——— (1973). "Hockey Helmets, Concealed Weapons, and Daylight Savings: A Study of Binary Choices with Externalities." *Journal of Conflict Resolution* 17: 381-428.

——— (1978). "Micromotives and Macrobehavior." In Thomas Schelling (ed.), *Micromotives and Macrobehavior.* New York: Norton, pp. 9-43.

Shubik, Martin (1959). *Strategy and Market Structure.* New York: Wiley.

——— (1970). "Game Theory, Behavior, and the Paradox of the Prisoner's Dilemma: Three Solutions." *Journal of Conflict Resolution* 14: 181-94.

Smith, Margaret Bayard (1906). *The First Forty Years* *of Washington Society*. New York: Scribner's.

Snyder, Glenn H. (1971). " 'Prisoner's Dilemma' and 'Chicken' Models in International Politics." *International Studies Quarterly* 15: 66-103.

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

Trivers, Robert L. (1971). "The Evolution of Reciprocal Altruism." *Quarterly Review of Biology* 46: 35-57.

Warner, Rex, trans. (1960). *War Commentaries of Caesar*. New York: New American Library.

Wilson, David Sloan (1979). *Natural Selection of Populations and Communities*. Menlo Park, Calif.: Benjamin/Cummings.

Young, James Sterling (1966). *The Washington Community, 1800-1828*. New York: Harcourt, Brace & World.