PN-ABL-221

IRIS Reprint

The Center for Institutional Reform and the Informal Sector (IRIS) has two main purposes: expanding knowledge about institutions in economic development through research, and assisting reform efforts in the third world and in countries undergoing transitions to a market economy. The premise of the IRIS Center is that in unsuccessful economies the existing rules establish poor incentives, often forcing economic activity into the informal economy, and that appropriate reforms improve economic performance. IRIS is especially concerned with the legal and policy framework needed for democratic societies with competitive markets.

IRIS
7100 Baltimore Avenue, Suite 510
College Park, MD 20740
(301) 403-8153

CENTER FOR INSTITUTIONAL REFORM AND THE INFORMAL SECTOR

University of Maryland at College Park

Center Office: IRIS Center, 2105 Morrill Hall, College Park, MD 20705
Telephone (301) 405-3110 • Fax (301) 405-3020

The Credibility Game: Reputation and Rational Cooperation in a Changing Population

August, 1992

Joel M. Guttman

IRIS Reprint No. 29

This publication was made possible through support provided by the U.S. Agency for International Development, under Cooperative Agreement No. DHR-0015-A-00-0031-00.

Author: Joel Guttman, Bar-Ilan University, Israel
Reprinted from Journal of Comparative Economics, vol.16, pp. 619-632, 1992.

JOURNAL OF COMPARATIVE ECONOMICS 16, 619–632 (1992)

The Credibility Game: Reputation and Rational Cooperation in a Changing Population[1]

JOEL M. GUTTMAN

Bar-Ilan University, 52900 Ramat-Gan, Israel

Received August 21, 1992

Guttman, Joel M.—The Credibility Game: Reputation and Rational Cooperation in a Changing Population

Labor-managed firms often depend on voluntary cooperation to generate efficient levels of labor input. This paper shows how cooperation in Prisoner's Dilemma-like situations can be an equilibrium outcome, when rational individuals act so as to preserve reputations for cooperating. The theory developed here implies that (a) voluntary cooperation will be negatively related to population turnover, and (b) the relationship of community size to voluntary cooperation will have an inverted-U shape. *J. Comp. Econom.*, December 1992, **16**(4), pp. 619–632. Bar-Ilan University, 52900 Ramat-Gan, Israel. © 1992 Academic Press, Inc.

Journal of Economic Literature Classification Numbers: P32, C73, H41.

1. INTRODUCTION

In many labor-managed firms (LMFs), the remuneration of the worker is not simply a function of his work input. For example, in the Israeli Kibbutz, the net output is divided equally among members, regardless of the contribution of the member to collective output. In such situations, there is potentially a serious free-rider problem: Why should the worker contribute to collective output when his own income is independent of his work contribution? Various theories have been proposed to deal with this dilemma, but none have been entirely satisfactory.[2]

[1] This research began when I was an Olin Faculty Research Fellow at the Yale Law School, and has since been supported by a grant from the Agency for International Development to the Institutional Reform and the Informal Sector (IRIS) program of the University of Maryland. I thank Nava Kahana, Avi Weiss, and Michael Yalkut for helpful comments, while retaining responsibility for all remaining errors.

[2] See Guttman (1991b) for a critical survey.

0147-5967/92 $5.00
Copyright © 1992 by Academic Press, Inc.
All rights of reproduction in any form reserved.

The problem of explaining voluntary cooperation in LMFs like the Kibbutz can be divided into two subproblems. The first subproblem is explaining how individuals voluntarily provide each other with incentives to cooperate, particularly when there are many interacting individuals. This is the free-rider problem as the term is usually understood. The second problem, which arises even in two-person, binary choice situations, e.g., the Prisoners' Dilemma, may be called the credibility or self-enforcement problem. Briefly, the problem is explaining how individuals convince their fellow members that they will uphold their commitments, when, for any reason, the legal system will not intervene to enforce these commitments. For example, in the Israeli Kibbutz it would be very costly, and perhaps impossible for socio-institutional reasons, for the courts to force a member to honor his work commitments, even if some form of formal or informal agreement could be reached so as to solve the free-rider problem.

In order to make the distinction between these two subproblems clearer, it is worthwhile to review the study of Guttman and Schnytzer (1989), which addressed the following paradox. The conventional Cournot–Nash theory of the provision of public goods, when applied to LMFs (see, e.g., Sen, 1966), predicts that a purely egalitarian LMF such as the Israeli Kibbutz will suffer from an undersupply of labor, since each individual's labor supply is a contribution to the provision of a pure public good.[3] In contrast, the theory predicts that an LMF that divides the collective output according to relative work contributions, such as the classic Kolkhoz, will suffer from an *oversupply* of labor. Yet we observe that the Kibbutz, though presently going through a crisis, has historically succeeded in the sense of generating high labor inputs and producing efficiently, at least relative to the predictions of the conventional theory. The Kolkhoz, on the other hand, has been a failure by all standards.

It was found that by applying the theory of strategic matching (see Guttman, 1978, 1987), this paradox can be solved. In the theory of strategic matching, N players play a two-stage game. In the first stage, players choose matching rates, which, in the LMF context, link their labor contributions to the contributions of their fellow workers. These matching rates provide incentives to one's fellow workers to take account of the full marginal product of their labor inputs, and not only their own share of their marginal product, which, in the Kibbutz, is $1/N$ times the full marginal product. In the second stage, workers choose flat contributions, which are like the simple work contributions in the conventional Cournot–Nash model. These flat contributions are then matched at the matching rates determined in the first stage:

[3] The collective output is a public good in the sense that workers cannot be excluded from enjoying the benefits of their fellow workers' effort, regardless of whether they contribute to collective production. It is, however, a public good subject to congestion, since an increase in the number of members will decrease per-member benefits.

2

In both stages, the equilibrium is a Nash noncooperative equilibrium; each worker's matching rate is a best-response to his fellow workers' matching rates, and similarly for the flat contributions. In addition, each worker is assumed to predict the equilibrium in the second stage or subgame when determining his matching rate in the first stage, and to use this prediction in order to determine his optimal matching rate. Thus, the equilibrium in the overall game is subgame perfect.

As noted above, collective output in the Kibbutz is divided equally, so that combined work effort is a pure public good. In such a setting, the theory of strategic matching predicts, for identical members, that each member will set a matching rate equal to unity, and these matching rates induce all members to work up to the point at which their full marginal product is equal to the opportunity cost of their time. This efficient equilibrium is the unique equilibrium of the game, regardless of the number of members, although, as N increases, information about other members' matching and flat work contributions becomes less perfect, making the model less realistic. In the Kolkhoz, on the other hand, the theory of strategic matching predicts that the matching rates will rise above unity, as members try to increase their relative work inputs and thus increase their personal incomes. But matching rates above unity depress the flat work contributions below their Pareto-optimal levels, since, by working harder, a member only decreases his relative work contribution when his fellows match his work increase at a rate above unity. Thus the theory predicts suboptimal work effort in the Kolkhoz but not in the Kibbutz, which is consistent with observed behavior but in direct contradiction to the received theory.

While the theory of strategic matching thus appears to solve the first subproblem indicated earlier, explaining how members voluntarily provide each other with group-optimal incentives to contribute to the public good, and also solves the Kibbutz–Kolkhoz paradox, it is open to the following objection. The matching commitments made in the first stage of the game are assumed to be honored after the flat contributions are chosen in the second stage. But each player has an incentive to renege on his matching commitment and make only his flat contribution. What prevents him from doing so? Apart from the possibility of enforceable contracts, which is not very relevant to the Kibbutz and Kolkhoz, Guttman and Schnytzer (1989) suggested that the matching game is only part of a larger "credibility game" that embraces other aspects of the individual's ongoing interactions with his community. This credibility game can be modeled as a repeated Prisoners' Dilemma. Using the Folk Theorem, one equilibrium of this game would be the cooperative outcome, honoring one's matching commitments if the individual's last round in the community is unknown to his fellow members.

The question raised in the preceding paragraph is clearly the second subproblem posed earlier, the credibility problem. A more satisfactory answer to

this question would not rest content with an appeal to the Folk Theorem, which allows for a multiplicity of inefficient equilibria as well as the efficient, cooperative outcome. The present paper discusses two ways of ruling out many or all of these inefficient equilibria, thus providing a more convincing explanation of voluntary cooperation. The first approach, discussed in Section 2, follows Kreps et al. (1982) by arguing that the rational members of a community, e.g., a Kibbutz or Kolkhoz, that includes a small number of mechanical tit-for-tat (TFT) types will behave as if they themselves are TFT types. In other words, members will rationally try to maintain reputations for being TFT types, and these reputations are believable because some true TFT types actually exist in the community. Section 2 poses the question as to whether the mechanical TFT types could survive a social evolutionary process in which only the strategies receiving the highest average payoffs are reproduced. This question is answered in the affirmative if there is a small cost of rational optimizing, or if the community invests in promoting the TFT type through education. An empirical implication arising from this model is that communities having lower population turnover will tend to cooperate more extensively. This implication is supported by the empirical work of Guttman and Haruvi (1986) on the Israeli Moshav (cooperative village).

Section 3 develops a second approach, which does not assume that optimizing is costly and does not introduce irrational types. In the second model, players do not actively punish noncooperative play in the manner of TFT. Rather, they simply refuse to play with actors whose past behavior is blemished by noncooperation. In the second model, in common with the first, there is a need to introduce some form of uncertainty in order to rationalize cooperative reputations. While the first model relies on uncertainty regarding one's counterpart's type, rational or TFT, the second model assumes that players are uncertain as to when their opponent will leave the community, i.e., when his last round occurs. This uncertainty is sufficient to induce cooperative reputation-building in a manner somewhat similar to the behavior predicted by the first model. In addition to predicting the negative effect of turnover on cooperation implied by the first model, the second model also implies an inverted-U-shaped relationship between community size and the level of cooperation in the community.

2. RATIONAL OPTIMIZERS AND TIT-FOR-TAT TYPES

Consider the game depicted in Table 1. If the two players play C, e.g., work hard, they both receive 1 unit of utility. If they both play D, e.g., slack off, they both receive 0 units. If, say, player 1 plays D and player 2 plays C, then player 1 receives $a > 1$ units while player 2 receives $b < 0$. Table 1 thus depicts the classical Prisoners' Dilemma (PD). In order to make the coopera-

TABLE 1

PAYOFF MATRIX IN PRISONERS' DILEMMA

	C	D
C	(1, 1)	(b, a)
D	(a, b)	(0, 0)

tive outcome (C, C) the only Pareto-optimal strategy-pair, we stipulate further that $a + b < 2$.

Let the game be played T times, with T known to both players. Furthermore, we assume that there is no discounting: players maximize the expected sum of their payoffs over the whole game. If both players are rational, and if this rationality is common knowledge to both players, the only resulting equilibrium will be for both players to play D, defect, throughout the game. This can be shown by the familiar backward induction argument. In round T, they will both play D, since this will not affect rewards in future rounds as there are no future rounds, and D is the dominant strategy in the stage game. Knowing this, in round $T - 1$, both players will play D, since again there will be no punishment in the following round: Regardless of what strategy they play in round $T - 1$, they will defect in round T. The argument thus precedes backward to the first round of the game.

Now let each player attach a small probability $(1 - p)$ to the proposition that he is not playing with a rational opponent, but rather with a player who mechanically plays the TFT strategy. The TFT strategy starts out with the cooperative (C) move, and then always plays what the opponent played in the previous round. One may well ask, where does this uncertainty come from, particularly if both players in fact are rational? One answer (Guttman, 1991a) is that while these two players are rational, they are members of a larger population which includes a proportion $(1 - p)$ of TFT types. Players are randomly matched and do not know, at least initially, the type of player with whom they are playing. This would explain the uncertainty of each player, but raises the question, is it reasonable to assume that part of the larger population consists of irrational TFT types? This question can be answered affirmatively if either of the following conditions are met: (a) there is a cost to rational optimizing, which cancels out the evolutionary advantage that the rational players would otherwise have over the TFT types, or (b) there is social investment in promoting TFT types through education.[4]

[4] See Guttman et al. (1992) for an analysis of education to inculcate socially optimal tastes. Here, however, the suggestion is that tastes for TFT behavior, rather than simply altruistic tastes, are inculcated.

5

Given this uncertainty, the backward induction argument no longer holds, at least in the simple form given above. Since each player cannot be certain that his counterpart will defect in the last round, it is no longer necessarily optimal to defect in round $T - 1$. As Kreps et al. (1982) showed for the case of one-sided uncertainty (i.e., only one player is uncertain of his opponent's type, while the opponent is certain he is playing with a rational player) the number of rounds in which there may be defections is bounded by an expression that depends only on a, b, and p, and is independent of T. These defections will occur toward the end of the game. Thus, as T increases, the *proportion* of rounds in which there must be joint cooperation in any sequential equilibrium approaches unity.

This striking result can be demonstrated either for one-sided or two-sided uncertainty by first analyzing the optimal response of a rational player to a preemption, i.e., an unprovoked defection, in, say, round t by his opponent. Such a preemption reveals the opponent to be a rational player, since a TFT player, by definition, never preempts. Now, if the preempted rational player does not respond in the following round by defecting, he reveals himself to be rational, resulting in joint defection until the end of the game, because the rationality of the two players would now be known to both. In round $t + 1$ taken by itself, it is clearly the dominant strategy to defect. Thus, by defecting in round $t + 1$, the preempted player leaves open the possibility of cooperation in the future and plays the dominant strategy in the stage game. By similar reasoning, the preempted player would not "forgive" his opponent and return to cooperating before his opponent does so. In short, the defected rational player will mimic the TFT type, precisely in order to avoid the emergence of joint knowledge of rationality.

Given that his opponent would behave like a TFT type regardless of whether he is really a TFT type or a rational player, any rational would-be preemptor would, in general, not plan a preemption and a subsequent return to cooperation, but rather would plan to preempt and then to continue to defect until the end of the game. The reason for this is simply that by preempting and returning to cooperation in, say, the following round, he receives $a + b$ units of utility, whereas by simply cooperating over the same two rounds he would receive 2 units, and we have assumed that $a + b < 2$. Thus, a loop from C to D and back to C gives a lower payoff, over the course of the loop, than simply cooperating over the course of the loop; and the difference in payoffs grows as the loop increases in length. While a loop gives a lower payoff, over the course of the loop, than cooperating, it may be still better to preempt and then continuing to defect until the end of the game. The attractiveness of such a "preempt strategy" increases as the end of the game approaches. In the last round, it clearly is optimal to defect; but the optimal preemption round is, in general, before the last round.

Given these characteristics of the sequentially optimal behavior of two

6

rational players, we can conclude that the sequential equilibrium will generally have one of three forms: (a) joint defection throughout the game; (b) joint cooperation, then a preemption, followed by joint defection until the end of the game; or (c) a simultaneous switch from cooperation to defection, with joint defection continuing until the end of the game.[5]

Which of these three structures will occur in equilibrium depends on the proportion of rational actors in the community, p, on the length of the game T, and on the payoff parameters, a and b. In particular, if $p > 1/a$, the equilibrium will be of forms (a) or (b), while if $p < 1/a$, the equilibrium, if it is a pure strategy equilibrium, will be of form (c) (Guttman, 1991a, Proposition 3).

Let us first consider the case in which $p > 1/a$. When p is very high, or if the game is of sufficiently short duration, the equilibrium will be of the degenerate joint defect type from the beginning of the game. But when p is lower, one can calculate, in the spirit of Kreps et al. (1982), a maximum number of rounds before the last round in which defection may occur. This maximum is calculated by comparing the expected payoff of the preempt strategy to that of a "wait strategy" as the game progresses. The wait strategy instructs a player to let himself be preempted and then defect in the following round, continuing to defect until his opponent returns to cooperation. In the last round, the strategy specifies defection under all circumstances. The wait strategy has the advantage that, against a TFT opponent who, by definition, never preempts, the player cooperates until round $T - 1$, with the consequent gain in payoff over the game. The risk, of course, is that the opponent is rational and preempts at some round t before the end of the game, giving the actor playing the wait strategy a one-time negative payoff of b. It is easy to verify that the expected payoff of the preempt strategy is $t - 2 + a$, where the preemption is planned in round $t - 1$. The expected payoff of the wait strategy is

$$EU_{\text{wait}} = (t - 1 + b)p + (T - 1 + a)(1 - p),$$

where the opponent, if rational, is expected to preempt in round t.

Figure 1 shows the expected payoff of the wait strategy as a function of t, and the expected payoff of the preemptor when he preempts in round $t - 1$.

[5] In Guttman (1991a), a difficulty with this point is addressed. While a preemptor initiating a loop clearly loses over the course of the loop, it remains to be shown that he gains nothing in the continuation of the game. If the preemptor *did* gain in the continuation more than he lost over the course of the loop, such a loop would be part of the equilibrium path. In an Appendix to Guttman (1991a), it is shown that this difficulty would not change the equilibrium of the game significantly, even if the initiator of a loop did gain in the continuation enough to make such a loop worthwhile. In fact, however, an example of the initiator of a loop gaining in the continuation of the game has yet to be found.

7

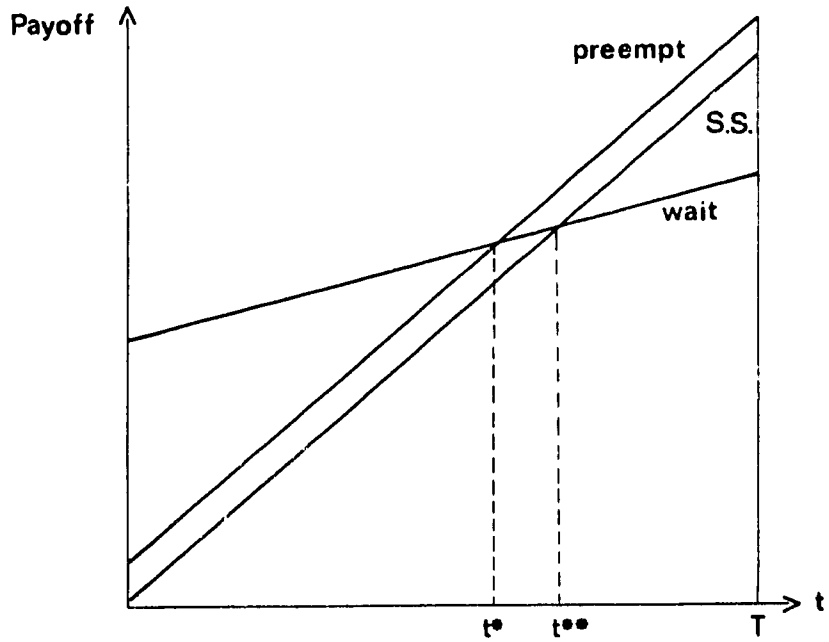

FIG. 1. Expected payoffs in finitely repeated Prisoners' Dilemma.

The figure also shows the expected payoff of the simultaneous switch (SS) strategy, but this does not concern us at present. As the figure shows, before round t^* the wait strategy gives a higher expected payoff, while after t^* the preempt strategy gives the higher expected payoff. By equating the two expected payoffs,

$$t^* = T - [(a - b)p - 1]/[1 - p].$$

We are now in a position to note the following fact regarding this game: when $p > 1/a$, there is, in general, no pure strategy equilibrium. Suppose, for example, that player 1 expected player 2, if rational, to preempt at some round $t > t^*$. Player 1's best response is to preempt in the previous round. Player 2, expecting player 1, if rational, to preempt in round $t - 1$, would then want to preempt in round $t - 2$. And thus it continues, until the expected preemption round of the opponent is before t^*. At this point, the best response is to play the wait strategy. But the best response to the wait strategy is to preempt in round $T - 1$. And then the cycle starts again: the best response is to preempt in round $T - 2$, and so forth. Now suppose that player 1 initially expects that his opponent, if rational, will preempt in some round $t < t^*$. Then player 1's best response is to play the wait strategy. But then player 2's best response is to preempt in round $T - 1$, and then the cycle starts. We conclude that the game, in general, has no pure strategy equilibrium.

Nevertheless, the game must have a mixed strategy equilibrium. In Guttman (1991a, Prop. 4), it is shown that in this mixed strategy equilibrium

8

there is a zero probability of defections at or before round $t^* - 1$. Since $T - t^*$ is independent of T, we obtain the main result of Kreps et al. (1982) that the number of rounds at the end of the game in which there may be defections is independent of the length of the game. Thus, as noted above, as T increases, the *proportion* of cooperative play approaches unity. But the upper bound derived here on the number of rounds in which there may be defections is considerably lower than that obtained by Kreps et al. (1982). The reason for this tightening of the upper bound is the additional information incorporated regarding the structure of the equilibrium path.

We now turn to the case in which $p < 1/a$. In this case, the expected payoff of switching simultaneously from C to D with one's opponent is higher than the expected payoff of preempting. Figure 1, as noted earlier, also shows the expected payoff of a simultaneous switch (SS) at round t. Since the figure is drawn on the assumption that $p > 1/a$, the SS curve is below the preempt curve; in the opposite case which we are now considering, the relative position of these two curves is reversed. Since the expected payoff of a simultaneous switch at stage t is $t - 1 + (1 - p)a$, the SS curve intersects the wait curve at stage

$$t^{**} = T + [bp/(1 - p)].$$

Before t^{**}, the wait strategy gives a higher expected payoff than the SS strategy, while after t^{**}, the opposite holds. Consequently, before t^{**} there will be no defections in equilibrium, since the best response to an expected preemption by one's opponent, if he is rational, at any stage $t < t^{**}$ is to wait. After t^{**}, however, there may be simultaneous switch equilibria, and at the last round T there will always be such an equilibrium. Note that the earliest possible switch-point from C to D, t^{**}, is separated from stage T by a number of rounds that is independent of T, so that the result of Kreps et al. (1982) is again obtained.

In Guttman (1991a) an additional result was obtained in order to assess the evolutionary viability of the TFT type. It was found that (a) the rational players, not surprisingly, receive a higher expected payoff than the mechanical TFT types, but (b) the evolutionary advantage of the rational actors over the whole PD game stays constant as T increases. Now let us introduce a cost of optimizing. As T increases, the complexity of the rational player's decision problem increases rapidly, so that this cost of optimizing would also increase. We thus obtain the result that, for sufficiently large T, the TFT types may actually have a net evolutionary advantage. For intermediate values of T, the two types may coexist in mixed evolutionary equilibrium, while for lower values of T, the rational actors retain their evolutionary advantage and drive the TFT types to extinction. Thus we find that a community in which PD-type interactions are relatively long lived would tend to behave more cooperatively than a community in which these interactions have a relatively

short duration. In any case, it was found that the assumption underlying the model of Kreps et al., that one player suspects his opponent of being a mechanical TFT type, can be justified only if there is a cost of optimizing and if T is sufficiently large.

3. REPUTATIONS IN A CHANGING POPULATION

We now study a model that shows promise of explaining the emergence of cooperative reputations without assuming a cost of optimizing. As indicated in the Introduction, this model explicitly allows for turnover in the community whose behavior is being studied. Consider a community consisting of a large number of individuals, each of whom stays in the community for an exogenously determined T time periods. In each of these time periods, each individual has the opportunity of playing the PD game once, with a randomly chosen counterpart. This PD interaction need not be in the productive sector of the community. It may equally well be in the consumptive sector, e.g., showing up on time to play tennis. The idea of the model is that reputations are formed and preserved in all of the individual's interpersonal interactions, and these reputations are then used to enforce matching commitments in the productive sector, i.e., the LMF.

While the individual's counterpart is assumed to be chosen randomly, we assume that each individual can refuse to play the game with the individual with whom he is matched, and then is costlessly matched with a new individual. The date of each individual's entry into the community is private information to that individual. Thus, even though T is the same for all individuals, and this is known by all individuals, the precise date of any counterpart's departure is unknown to everyone else. Individuals are assumed to be continuously entering and exiting the community; thus, in effect, we have a model of overlapping generations. The rate of entry and exit is assumed to be constant over time, so that, at any point in time, all cohorts are the same size and comprise $1/T$ of the population. If any individual plays D in one of these PD encounters, this fact becomes immediately known to all individuals in the community.[6]

[6] Note that we are assuming perfect information regarding an individual's history of play, but not of the date at which he entered the community. The uncertainty assumed regarding the individual's entry date is simply a device for modeling uncertainty regarding the individual's *departure* date, since the length of the individual's career in the community, T, is assumed, for simplicity, to be the same for all players. A formally more descriptive, but more complicated, treatment would make T a random variable as viewed by all other players; this would permit making the entry date common knowledge. An alternative resolution of this difficulty would be to assume that individuals have perfect information of each player's move in the previous period only. When an individual defects, this results in no one playing the game with him in the following period. In subsequent periods, players observe that the defector has no one with whom to play the game and infer that in the past the individual must have defected.

As shown below, for sufficiently large T and for b sufficiently small in absolute value, one equilibrium in the game is for all individuals to play cooperatively over their entire terms in the community, defecting only in the stage in which they plan to leave. In such an equilibrium, a deviating individual, who chose to play D even once before his last period in the community, would induce all other players to refuse to play with him in future periods. Thus he would receive a payoff of zero in all future rounds, just as if all players were using a "grim" strategy of playing D forever against any one-time deviator. The rationale for this "punishment" is simple: by playing D, the deviating individual shows that his own strategy is not to play C consistently. Since it costs players nothing to be rematched against a different player, there is no reason why they should not do so, leaving the deviator an "outcast" without anyone with whom to play the game.

The claim of the previous paragraph can be demonstrated as follows. Let the stage at which each player plans to defect be T, i.e., his last period in the community. In each period there is a probability of $1/T$ that the player's opponent will preempt him, since there is a probability of $1/T$ that his opponent is in the cohort presently exiting the community. Now let us express the expected payoff of deviating, by calculating the expected payoff of preempting in stage $t \leqslant T$ of the player's career. In each of the $t-1$ stages preceeding stage t, he has a probability of $(1/T)$ of being preempted, which would give him a payoff of $b < 0$, and a probability of $(T-1)/T$ of not being preempted, giving him a payoff of unity. In stage t, he has a probability of $(1/T)$ of meeting an opponent who simultaneously switches to D, giving him a payoff of zero, and a probability of $(T-1)/T$ of meeting an opponent who cooperates, giving him a payoff of a. In all following stages, no one plays with him, giving him a payoff of zero. Thus, the expected payoff of preempting at stage t is

$$EU_t = [(t-1)/T]b + [(t-1)(T-1)/T] + [(T-1)/T]a.$$

Taking first differences with respect to t,

$$\Delta EU_t/\Delta t = (1/T)(b + T - 1).$$

If $T - 1 > -b$, this expression will be positive. Therefore, the effect on the player's expected payoff of deviating by defecting before round T is negative. All players thus will preempt at the latest possible stage, i.e., stage T.

In the equilibrium just described, all players rationally expect their opponents to cooperate. If, on the other hand, all players were expected to defect throughout their careers, this would also be an equilibrium, since there would be no punishment for defecting. Can an expectation of cooperation by other players be shown to be more reasonable than an expectation of defection? Let us assume that the expectations of players are adaptive: the past behavior of the community as a whole serves as a focal point that generates

11

expectations of future equilibrium behavior. Thus, in a community that, in the past, was in an equilibrium in which people cooperated, players would expect their counterparts to cooperate in the present. This expectation of cooperation, together with the preference of each player to play the game with opponents whose past behavior is cooperative, would be sufficient to generate the cooperative outcome described above. This possibility suggests an interesting hysteresis effect: cooperatives that were strong (cooperative) in the past will tend to stay strong, while cooperatives that were weak in the past will stay weak.[7]

It is, in fact, sufficient that a nucleus of cooperators exist in the community, in order for an individual to expect to be punished for defecting by not finding individuals with whom to play the game and rewarded for cooperating by being able to find players with whom to play the game.[8] This nucleus can be small relative to the size of the community, if b is small in absolute value, provided that the *absolute* number of cooperators is large enough to create this expectation of reward and punishment. What is the minimum proportion of cooperators in the community that would make it worthwhile for all players to cooperate? Suppose that each player initially assigns a probability q to the proposition that his counterpart will cooperate in the current period, and $(1 - q)$ to the proposition that he will defect. The expected utility of preempting at round t then becomes

$$EU_t = (t - 1)(1 - q)b + (t - 1)q + aq.$$

Taking first differences with respect to t,

$$\Delta EU_t/\Delta t = b(1 - q) + q = b + q(1 - b),$$

which is positive iff $q > -b/(1 - b)$. Thus, if q is large enough relative to $-b$, the equilibrium that would emerge would be cooperation by all members until their last period in the community.

It should be noted that no assumption was made that players naively expect their *individual* counterparts to play as they have played in the past. Such an expectation would be naive unless mechanical types like the TFT type in the model of Section 2 existed in the community, and we have avoided introducing such types in the present model. Rather, the assumption is that the equilibrium behavior of the community as a whole may

[7] See Haruvi and Kislev (1984), who argue that such a polarization effect exists among Israeli Moshavim.

[8] Here, again, the model of Guttman et al. (1992) provides a starting point for modeling collective investment in education so as to create such a nucleus of cooperators. The collective educational investment can be modeled as a noncooperative contributions game, whose Nash equilibrium point would yield a small, but positive level of educational investment, sufficient to create a small nucleus of cooperators.

generate expectations that this behavior will continue as long as outside influences stay constant.

The above analysis purposefully avoided introducing grim or TFT punishment strategies, which would easily explain cooperative play. The reason is that in a game with an uncertain endpoint, there are many strategies that can coexist in equilibrium. It seems more plausible to assume that players simply avoid playing with individuals who have once defected (cf. Telser, 1980; Tullock, 1985) than to single out any specific strategy of active punishment of defectors.

Two empirical implications can be derived from this analysis. The first is that the larger is T, implying a relatively low turnover in the community, the more cooperative play would be expected. Note, however, that members of a community with *zero* turnover would know each other's last round with a high degree of certainty, and this would generate mutual defection over all periods. For, if an actor's last round is known, no one would play with him in that round. Then, in his next-to-last round, he would again have no incentive to cooperate, and, knowing this, no one would play with him in this round as well. Thus, by backward induction, no one would play the game.

The second implication results from weakening the assumption that actors can costlessly be matched up with a new partner if they refuse to play with the current partner. This "perfect competition" assumption would be less realistic in small communities, in which a given PD interaction can be performed by only a small number of counterparts. If, in addition, we recognize that the assumption of perfect information of past defections becomes less realistic in very large communities, we obtain an expectation that the level of cooperation will be related to the size of a community in an inverted-U shape. Very large and very small communities will tend to have relatively low levels of cooperation, while medium-sized communities will be large enough to generate nearly perfect "competition" while preserving high levels of awareness of past behavior.

4. CONCLUDING REMARKS

This paper has developed two models of the credibility game, designed to explain the emergence of cooperative reputations in an ongoing community. This, in turn, would tend to explain why LMFs sometimes succeed. The overall implication, aside from the specific empirical implications noted above, is that LMFs can thrive if they are set in an ongoing community whose multifaceted interactions serve to provide the basis of reputations for credibility, which in turn can support strategic matching of work contributions in the production activities of the community.

REFERENCES

Guttman, Joel M., "Understanding Collective Action: Matching Behavior." *Amer. Econom. Rev.* **68**, 2:251–55, May 1978.

Guttman, Joel M., "A Non-Cournot Model of Voluntary Collective Action." *Economica* **54**, 1:1–19, Feb. 1987.

Guttman, Joel M., "Rational Actors, Tit-for-Tat Types, and the Evolution of Cooperation." Unpublished, Bar-Ilan University, 1991a.

Guttman, Joel M., "Voluntary Collective Action." In A. L. Hillman, Ed., *Markets and Politicians: Politicized Economic Choice*, Boston: Kluwer, 1991b.

Guttman, Joel M., and Haruvi, Nava, "Cooperation and Part-Time Farming in the Israeli Moshav." *Amer. J. Agric. Econom.* **67**, 1:77–87, Feb. 1986.

Guttman, Joel M., and Schnytzer, Adi, "Strategic Work Interactions and the Kibbutz–Kolkhoz Paradox." *Econom. J.* **99**, 397:686–99, Sept. 1989.

Guttman, Joel M., Nitzan, Shmuel, and Spiegel, Uriel, "Rent Seeking and Social Investment in Taste Change." *Econom. Politics* **4**, 1:31–42, Mar. 1992.

Haruvi, Nava, and Kislev, Yoav, "Cooperation in the Moshav." *J. Comp. Econom.* **8**, 1:54–73, Mar. 1984.

Kreps, David M., Milgrom, Paul, Roberts, John, and Wilson, Robert, "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma." *J. Econom. Theory* **27**, 2:245–52, Aug. 1982.

Sen, Amartya. "Labour Allocation in a Cooperative Enterprise." *Rev. Econom. Stud.* **33**:361–371, 1966.

Telser, Lester G., "A Theory of Self-Enforcing Agreements." *J. Business* **53**, 1:27–44, Jan. 1980.

Tullock, Gordon, "Adam Smith and the Prisoners' Dilemma." *Quart. J. Econom.* **100**, 402:1073–1081, Supplement 1985.