Cooperation via Communication: Supporting Social Norms with Costly Messages.*

Kevin Hasker
Department of Economics
Bilkent University
Bilkent, Ankara, Turkey
hasker@bilkent.edu.tr

March 16, 2001

Abstract

This paper shows that if interaction is not anonymous then costly messages can establish a folk theorem in repeated matching games. This result holds for all population sizes and a broad class of matching rules—including many choice based rules. Cooperation is achieved through a formalization of a "reference" strategy.

1 Introduction

Trade is a cooperative activity, but how can selfish traders maintain cooperation? When economists address this question they usually assume exogenous institutions such as courts. Does this means that without courts there can be no trade? This paper shows that if people can send messages then there can be trade without courts when interaction is repeated and not anonymous.

*The author would like to thank Eddie Dekel, Juuso Valimakki, Herve Moulin, and Asher Wolinsky for their helpful comments and the Sloan Foundation for their generous support. Of course any remaining errors are the responsibility of the author.

Trade takes place within a two player repeated matching game and people communicate between periods by sending costly messages. The messages used in this paper are modeled on signed letters and face to face communication.

These messages enable cooperation with very minimal restrictions on interactions. Without exogenous institutions repeated interaction is needed to overcome short run incentives. A repeated *matching* game weakens this restriction to people only having to interact with *someone*. Furthermore, unless people have no incentive to lie, messages used to achieve cooperation can not be anonymous. In environments where these conditions are met I show that society can organize itself and achieve cooperation.

An example of an interaction described by this model is the landlord/tenant relationship. Let us assume each landlord has only one apartment to let, all apartments are basically the same, and the tenants all move at the same time of the year. The problem in this relationship is that while a landlord may require a deposit a tenant can cost more than this. On the tenant's side there is uncertainty about whether the landlord will keep up the maintenance and repay the deposit. The incentive to cooperate in this situation is commonly achieved by asking for references. Here an equilibrium is constructed where everyone asks for references and no one cheats.

This paper builds on the work of Kandori [8] and Hasker [7]. These papers use exogenous information transmission, and the latter paper achieves a folk theorem nearly as weak as in the standard repeated game. However the authors always acknowledged that relying on exogenous information transmission is a simplifying assumption. Kandori [8] wonders in his conclusion if the same results could be achieved with endogenous information transmission. If not then what is the difference between the exogenous transmission and a court? This paper shows that the results in Kandori [8] and Hasker [7] do not depend on exogenous information transmission. They are a natural outcome of a structural analysis of the interaction.

Two basic problems must be overcome to achieve cooperation based on references. First, why would anyone give a reference? A landlord will probably never rent to the same tenant again, and thus has no personal interest in giving a reference. On the other hand why would anyone check

a reference? In equilibrium all people will follow the rules, and thus the landlord knows what the reference will be before asking. I address both of these problems by requiring that not checking or not sending references equivalent to deviating in the stage game. In a world where reputations are made or ruined by a person's word, a person's word is as important as what she does.

To allow this paper to focus on communication, I only show that players can achieve cooperation in a symmetric prisoner's dilemma game with the option of not interacting. The matching rule is uniform, but I augment the standard payoffs to mimic the incentives players have if the matching rule is not uniform or even history dependent. The results can be shown to hold more generally but this unnecessarily complicates the analysis. I will discuss how to generalize the results in the final section, but until that time I will focus on this simple game.

One important restriction is that the results do not depend on a player's being *un*informed—a weakness of trigger strategies among others. While messages are all we need the results wouldn't be very compelling if we had to rule out newspapers and other sources of information. These other sources could result in some people having full information. Thus the equilibria found here—which use a limited amount of information—should work even if some people do have full information, this requirement is that equilibria are *other straightforward.*

The two fundamental assumptions of our model are that interaction is non-anonymous and players can send messages, related research indicates these may be necessary to achieve cooperation. Kandori [8] and Ellison [6] consider anonymous interaction. They show that if there is a dominant strategy and a uniform matching rule then cooperation can be achieved for small populations. Unfortunately their results fail as the population grows larger and if there is *any* probability of people making errors then everyone stops cooperating in finite time. Ahn and Suominen [2] consider what can be done with non-anonymous interactions and limited messages. They extend a variant of Kandori's model (with one sided moral hazard) by allowing people to send messages to their neighbors ("word of mouth"). They find that they can extend Kandori's results to large populations if the number of neighbors

is required to be large. The paper before you achieves much more general results, here the population size does not affect analysis and the results can be generalized to a large class of matching rules and stage games.

Matsushima [12] and Kandori and Matsushima [11] also use messages in between the stage games in repeated game analysis. They use messages to allow people to correlate behavior when what happened in the stage game is not perfectly observable. Ben-Porath and Kahneman [4] also use this technique when what someone did in the stage game is only observable by some of the other players. However since both analyze repeated games the incentive problem is simpler. In both of these situations the incentive problem is getting truthful revelation, in repeated matching games the additional incentive problem of getting any revelation must also be solved.

In the next section of the paper the model and the stage game are presented. Following this I describe the strategy used to ensure cooperation. The fourth section then proves the results, and the final section discusses generalizations and possible paths for future research.

2 Model

Assume there is a finite population of players with an even number of members, call this set P where $\#(P) \geq 4$. These people will be matched in pairs repeatedly, with their payoff from the repeated matching game being the sum of their per-period payoffs geometrically discounted by $\delta \in (0,1)$. For arbitrary player x call $\mu(x)$ the person x is interacting with in period t, for the rest of the paper $\mu(i) = j$. Then in period $t+1$ i will be matched uniformly with all players in $P \backslash \{i,j\}$. After players observe who they are matched with each period they will play a *communication game* which will be followed by a *stage game.*

In the communication game players will send personal and joint messages. The model of a message is based upon two common methods of communication, letters and face to face communication. Both of these methods are *non-anonymous*—it is not possible for the sender to lie about who she is (letters must be signed)—and it is possible to send *joint* messages—if players i and j are interacting, then i and j can send a message from them both

(both must agree to this action). The receiver will observe who sent him the message and its contents—if it is a joint message then they will observe both senders. In any message the first element will be the identity of the sender, the second element will be the identity of the receiver, and the rest will be the contents. These contents will be identities of individuals in the population or one of five indicator functions, $\{d_{i,t}, d_{j,t}\} \in \{0,1,2\}$, $\{d_{i,t-1}, d_{j,t-1}\} \in \{0,1\}$ and $1(r_i \not\to j) \in \{0,1\}$. The indicator functions $d_{i,t}$ and $d_{j,t}$ will keep track of the number of deviations in the communication game. The indicator function $d_{i,t-1}$ equals one if last period i or i's last partner deviated, and $1(r_i \not\to j) = 1$ if r_i did not send a message to j. Given these functions the set of messages is: $M^0 = \left\{P \times P \times P^2 \times \{0,1\}^3 \times \{0,1,2\}^2\right\} \cup \emptyset$, where if $\emptyset$ is the message then this means a message was not sent. Note i can only send messages where i is the first element, so $M_i^0 = \left\{i \times P \times P^2 \times \{0,1\}^3 \times \{0,1,2\}^2\right\} \cup \emptyset$. The set of joint messages is the same as the set of messages except that the first two elements must be the identities of the senders, or $JM^0 = \left\{P^2 \times P \times P^2 \times \{0,1\}^2 \times \{0,1,2\}^2\right\} \cup \emptyset$ and $JM_{ij}^0 = \left\{\{i,j\} \times P \times P^2 \times \{0,1\}^2 \times \{0,1,2\}^2\right\} \cup \emptyset$. Players can send up to $\bar{k} \geq 4$ messages simultaneously (joint or not), and can send them in each of $\bar{s} \geq 5$ sub-periods before the stage game—indexed $t.1, t.2, t.3, ...t.\bar{s}$. Thus i's language is $M_i = \left((M_i^0)^{\bar{k}}\right)^{\bar{s}}$, and $\{i,j\}$'s joint language is $JM_{ij} = \left(\left(JM_{ij}^0\right)^{\bar{k}}\right)^{\bar{s}}$.

It will be indicated that a player $x \in P$ sends a message to $y \in P$ by a directed arrow: $x \to y$. For sets of players X, Y the notation $X \to Y$ indicates that $x \to y$ for all $x \in X$ and $y \in Y$. $X \leftrightarrows Y$ means $X \to Y$ and $Y \to X$. $X \not\to Y$ means that some player $x \in X$ did not send a message to some $y \in Y$, and $X \not\leftrightarrows Y$ means either $X \not\to Y$ or $Y \not\to X$. If something is done conditional on $x \to y$ then it is implicit x also must have made the correct statement. If a pair of players send a joint message then this is indicated by $\{x,z\} \overset{j}{\to} y$.

The costs of taking these actions will be dependent on the individual and period, for all $m_i \in M_i$, $jm_{ij} \in M_{ij}$ there will be a cost $c_{it}(m_i, jm_{ij})$. All that will be common knowledge among all the players is that this cost function is weakly positive, cardinally dependent, and has bounded incremental cost. A *cardinally dependent* cost function depends only on the number of messages,

or for $m_i, \tilde{m}_i \in M_i$, $jm_{ij}, \widetilde{jm}_{ij} \in JM_{ij}$ if $\#(jm_i) + \#(jm_{ij}) = \#(\tilde{m}_i) + \#\left(\widetilde{jm}_{ij}\right)$ then $c_{it}(m_i, jm_{ij}) = c_{it}\left(\tilde{m}_i, \widetilde{jm}_{ij}\right)$. A cost function has *bounded incremental cost* if for any $m_i^0 \in M_i^0$ if $m_i \cup m_i^0 \in M_i$ then

$$c_{ij}(m_i, jm_{ij}) \leq c_{it}\left(m_i \cup m_i^0, jm_{ij}\right) \leq c_{it}(m_i, jm_{ij}) + \bar{c}$$

(if $m_i^0 = \emptyset$ then $c_{ij}(m_i, jm_{ij}) = c_{it}(m_i \cup m_i^0, jm_{ij})$). If a message is a *joint message* then the cost is incurred when both parties agree to send the message.

After playing the communication game each player and his partner will play a symmetric prisoner's dilemma with the option of not interacting. Their action set will be $A = \{0, 1, 2\}$. Let a_{it} be the action player i takes in period t, then $a_t = \{a_{it}\}_{i \in P}$ is the vector of actions taken by the entire population in period t. The payoffs in the stage game will then be:

		j		
		0	1	2
i	0	$7 + \rho_{it}(a_t)$	$0 + \rho_{it}(a_t)$	0
	1	$8 + \rho_{it}(a_t)$	$4 + \rho_{it}(a_t)$	0
	2	0	0	0

where $\rho_{it}(a_t) \in \{0, 1\}$ is independent of a_{it} and a_{jt} and known only to player i. These auxiliary payoffs—$\rho_{it}(a_t)$—ensures that the derived equilibrium will work for much more general matching rules, these auxiliary payoffs would in general be generated by the interaction between matching and actions. They also are the incentive players have to deviate in the communication game. Write $\pi(a_i, a_j) + \rho_{it}(a_t)$ for player i's total payoff when $\{a_i, a_j\}$ are the actions i and j takes and $\rho_{it}(\cdot)$ is as defined above, then player i's per-period payoff is $u_{it}(a_t, m_i, jm_{ij}) = \pi_{it}(a_t) + \rho_{it}(a_t) - c_{it}(m_i, jm_{ij})$.

It will be shown that playing $\{0, 0\}$ can be an equilibrium path of this repeated matching game. The equilibrium will be an *endogenous social norm*, which has four elements, $\{Z, \tau, \sigma, \lambda\}$. $Z = \{0, 1, 2\}$ is the set of social statuses, and each player will have a $z_{it} \in Z$. τ is the *transition rule* which is implicitly a function of what i and j did last time and their social status last time, but explicitly will be a function of $\{d_{i,t}, d_{i,t-1}, d_{j,t}, d_{j,t-1}\}$. The *social standard of behavior* is then a function from Z^2 to A^2 and will

tell them what action to take today given their social status. λ is the *communication protocol* which will be detailed below, its primary purpose is to transmit $\{d_{i,t}, d_{i,t-1}, d_{j,t}, d_{j,t-1}\}$ correctly. Thus to dispense with the first three elements,

$$\begin{aligned} z_{it} &= \tau\left(d_{i,t}, d_{i,t-1}, d_{j,t}, d_{j,t-1}\right) = \max\left\{d_{i,t}, d_{i,t-1}, d_{j,t}, d_{j,t-1}\right\} \\ a_{it} &= z_{it} = \sigma\left(z_{it}, z_{jt}\right) \end{aligned}$$

or $\{i, j\}$ will share the worst future of the possibilities $\{d_{i,t}, d_{i,t-1}, d_{j,t}, d_{j,t-1}\}$.

This equilibrium will be an *other-straightforward* sequential equilibrium. An equilibrium is *other-straightforward* if some players having full information does not change the vector of statuses. Full information is the history of the entire game up to the current period—including who is matched together this period—and $\{c_{it}(\cdot), \rho_{it}(\cdot)\}$ for all i and t. Define $Z^t = \{z_{it}\}_{i \in P}$ $FI^t \subseteq P$ as the subset of players with full information in period t, and $X|y$ as the expectation of X given y.

Definition 1 *An endogenous social norm is* straightforward *if for all* $FI^t \subseteq P$ $Z^t|FI^t = Z^t|\emptyset$.

Comparing this with *straightforward* (as used in Kandori [8] and Hasker [7]) the only difference is that here some players might not send messages. These messages must have no affect (since no one's social status can be changed) but the equilibrium will still ask for them to be sent.

3 The Communication Protocol

Repeated matching games are very common and have a common problem. Two classic examples are employees with their employers and landlords with their tenants. In both of these situations most problems happen right before the relationship ends. At this point the intertemporal incentives are weakest and people frequently either shirk at work or do not pay rent. What mitigates these problems? The expectation that the employee's next employer or the tenant's next landlord will ask for a reference. Here an equilibrium is

constructed where the next employer always checks the reference and nobody shirks.

The strategy has two basic phases. First everyone gets a reference in the *reference phase*. This phase is what provides the intertemporal incentives for people to behave cooperatively. If someone deviates in this phase this information is shared with her current partner in the *reporting phase*. This phase ensures that any deviation is common knowledge among all affected parties before the stage game takes place.

The *reference phase* is a fairly straightforward strategy. To describe it concisely, consider an arbitrary player i. Think of this player as a tenant who wants to rent an apartment from the landlord $j = \mu(i)$. The new landlord will require a reference, so first i will send a letter to his last landlord requesting one—call this player r_i. r_i will then send off a letter for i, then i and j will send messages to each other to make sure everyone did what they were supposed to.

Now if r_i deviates—for example by not sending a reference—he must be punished but his partner $(\mu(r_i))$ clearly won't know this at this time. Informing $\mu(r_i)$ will be done in the *reporting phase*, which will be structured as a two stage declaration game. First i and j will send a joint message to both r_i and $\mu(r_i)$ (likewise $\{r_i, \mu(r_i)\}$ will send joint messages to $\{i, j\}$). Since all parties know that the their partner will be informed regardless of their action they will cooperate during this phase. In the second stage i and j will exchange messages to make sure everything went fine.

This strategy is an equilibrium with a few restrictions on beliefs. Notice that only limited restrictions are possible. Beliefs about what happened last period must allow for the fact that some players can have full information. Thus they can only be restricted to assuming that other players follow the social norm unless they know that this didn't happen. Beliefs during this period can still be restricted, these beliefs must be consistent with:

Off Path Beliefs Off the equilibrium path, players' beliefs must be consistent with:

1. $k+1$ deviations are infinitely less likely than k .

2. A deviation to cover up a previous deviation is infinitely more likely than any other single deviation.

In the rest of this section I will give a precise description of the strategy, the proof that it is an equilibrium will be in the following section.

3.1 The Reference phase.

First let me mention that any time that i expects $d_{i,t} = 2$ or $d_{j,t} = 2$, i stops sending all messages. The reference phase is a straightforward strategy, formally first (in $t.1$) i asks for a reference then in $t.2$ i sends off a reference. In $t.3$ i reveals everything that happened to j, including enough information for j to know what to do if there has been a deviation. To formalize this strategy it is easiest to write it down in a table form. The subperiod in which the message is sent will be written first, who is to send a message to whom is written second, then the contents of the message, and finally what condition must be met to send the message. Note that all actions are symmetric, for example $j \to r_j$ in $t.1$ and $i \to \mu(r_i)$ in $t.2$.

Subperiod	Action	Contents	Condition
t.1	$i \to r_i$	j	
t.2	$r_i \to j$	$d_{i,t-1}$	if $i \leftrightarrows r_i$ in $t.1$
t.3	$i \to j$	$\{r_i, \mu(r_i)\}; d_{i,t-1}, d_{j,t-1}, d_{i,t}, d_{j,t}$	

Now I will specify how $d_{i,t}$ changes in these three periods. For clarity I will write $d_{i,t}(x,s)$ and $d_{i,t-1}(x,s)$ where this is the value of $d_{i,t}$ ($d_{i,t-1}$) known by individual x at the end of subperiod s, given this convention in sub-period $t.3$ the variables $d_{i,t}(i,2)$ and $d_{j,t}(i,2)$ are equal to:

$$d_{i,t}(i,2) = \begin{cases} 2 & \text{if } i \not\leftrightarrows r_i \text{ in } t.1 \\ 1 & \text{if } i \not\to \mu(r_i) \text{ in } t.2 \\ 0 & \text{else} \end{cases}$$

$$d_{j,t}(i,2) = \begin{cases} 1 & \text{if } r_j \not\to i \text{ in } t.2 \\ 0 & \text{else} \end{cases}$$

and $d_{j,t-1}$ is based on j's reference. At the end of $t.3$:

$$d_{i,t}(i,3) = \begin{cases} 2 & \text{if } i \not\to j \text{ in } t.3 \\ \max\{d_{i,t}(i,3), d_{i,t}(j,3), |d_{i,t-1}(i,3) - d_{i,t-1}(j,3)|\} & \text{else} \end{cases}$$

and if $d_{i,t}(i,3) + d_{j,t}(i,3) = 1$ then the players send messages in $t.4$ in the reporting phase, if $d_{i,t}(i,3) + d_{j,t}(i,3) = 0$ they only have to send one more message, in $t.5$.

Before continuing on to this phase, let me consider a natural question. Why does i ask for a reference for herself instead of j? For example a landlord usually ask a tenant's last landlord for a reference, here the tenant does. There is a more complicated equilibrium where j asks for the reference. In in that equilibrium you must overcome the problem of j just not asking for the reference—which he knows the value of. Since r_i only knows who i is, he must first ask i who j is and then report the deviation to j.

3.2 The Reporting phase.

At this point if $d_{i,t}(i,3) + d_{j,t}(i,3) = 0$ then players only exchange one more message in $t.5$. If $d_{i,t}(i,3) + d_{j,t}(i,3) = 1$ then players send a joint message to both r_i and $\mu(r_i)$. Since these are joint messages if either does not send it then both observe this, since both r_i and $\mu(r_i)$ receive messages both knows it is worthless to try to cover them up. Thus everyone does what they are supposed to, and if there was a deviation in $t.2$ then this is verified. In table form this strategy is:

Subperiod	Action	Contents	Condition
t.4	$\{i,j\} \xrightarrow{j} \{r_i, \mu(r_i)\}$	$1(r_i \not\to j)$	$d_{i,t-1}(i) \neq d_{i,t-1}(j)$
t.5	$i \to j$	$d_{i,t}, d_{j,t}$	

In $t.5$

$$d_{i,t}(i,4) = \begin{cases} 2 & \text{if } d_{i,t}(i,3) = 1 \text{ and } \{i,j\} \overset{j}{\not\to} \{r_i, \mu(r_i)\} \\ & \text{or if } d_{i,t}(i,3) = 1 \text{ and} \\ & \qquad 1(r_i \not\to j) + 1(i \not\to \mu(r_i)) > 1 \\ & \text{or if } \ d_{i,t}(i,3) = 0 \text{ and } \{r_i, \mu(r_i)\} \xrightarrow{j} i \\ \min\{d_{i,t}(i,3) + d_{j,t}(i,3), 2\} & \text{else} \end{cases}$$

$d_{j,t}(j,4)$ can be found by symmetry. At the end of $t.5$

$$d_{i,t}(i,5) = \begin{cases} 2 & \text{if } i \not\to j \text{ in } t.5 \\ \max\{d_{i,t}(i,4), d_{i,t}(j,4)\} & \text{else} \end{cases}$$

Note that the messages in $t.5$ are there because the *absence* of a message will signal that something is wrong. Since people are supposed to reassure each other that everything is fine, in $t.5$ the lack of a message is the signal that $d_{i,t} = 2$. Another point is that messages will be sent in $t.4$ only if the original deviation really was in $t.2$. If someone lied in $t.3$ then they will not send the messages in $t.4$ since the liar knows $\{r_i, \mu(r_i)\} \overset{j}{\not\mapsto} \{i, \mu(i)\}$.

4 The Theorem.

The choice of the stage game makes the theorem straightforward.

Theorem 2 *If $\delta \geq \frac{1}{2}$ and $\bar{c} \leq 1$ then $\{0,0\}$ is an other-straightforward sequential equilibrium of the repeated matching game.*

Before proving this result let me mention what could be done with correlated actions. As in other folk theorems all that is really required about the correlated payoff (call this $\pi(a)$) is that $\pi(a) > \pi(1,1)$ and $\pi(1,1) - 4\bar{c} > \pi(2,2) = 0$. Thus the folk theorem holds as long as $\pi(a) - 4\bar{c} > 0$, where zero is the minmax in the game. Notice that along the equilibrium path four messages have to be sent every period, thus any payoff that is individually rational *given* communication can be supported. To prove that $\{0,0\}$ is an equilibrium I will first prove that if players intend to be honest when sending messages they will not deviate during the stage game. Then I will show that given this result players can not manipulate their payoff by lying.

Lemma 3 *As long as players intend to be honest, they will only send messages they are supposed to and not deviate during the stage game if $\delta \geq \frac{1}{2}$ and $\bar{c} \leq 1$.*

Proof. player will not deviate during the stage game when her status is 0 if:

$$\begin{aligned} \pi(1,0) - \pi(0,0) &\leq \delta\left(\pi(0,0) + \rho_{it+1}(a_{t+1}|a_{it}=0)\right) \\ &\quad -\delta\left(\pi(1,1) + \rho_{it+1}(a_{t+1}|a_{it}=1)\right) \\ \frac{\pi(1,0) - \pi(0,0)}{\pi(0,0) - (\pi(1,1)+1)} &\leq \delta \end{aligned}$$

and this is true if $\delta \geq \frac{1}{2}$. If she currently expects to receive $\pi(1,1)$ or $\pi(0,0)$ then she will gain nothing and the loss will be the same, thus she will follow the strategy during the stage game.

If they do not send the message in $t.1$ then they will not have to send four messages but they will receive $\pi(2,2)$ the worst possible future they can face is to expect to receive $\pi(1,1)$ currently, so they will not deviate if

$$\pi(1,1) - 4\bar{c} \geq \pi(2,2)$$

which is true as long as $\bar{c} \leq 1$. In all future sub-periods if they did not send a message in $t.1$ then there is no benefit to sending any further messages since their payoff will be $\pi(2,2)$, thus they will never intend to send these messages. The same result holds if they realize that they unintentionally lied about who j was in $t.1$.

In $t.2$ first we will deal with the case where (due to a player having full information) they realize that r_i lied about $\mu(r_i)$. In this case regardless of whether i sends a message or not i will not receive a reference, and if i sends messages in $t.4$ she will not receive messages, thus regardless of what i does they will always receive $\pi(2,2)$. This leaves only the case where everything went fine in $t.1$. In this case if i does not send the message in $t.2$ then i will receive $\pi(1,1) + \rho_{it}(a_t)$ if i follows the rest of the strategy, and the worst i could expect to receive if she does send the message is $\pi(1,1) + \rho_{it}(a_t)$, where a_t is the same in both cases. However, if i does not send the message she gains one message by doing this but will have to send two in $t.4$, thus she will not do this. If she decides not to send any messages then she will gain at most $3\bar{c}$ by doing this, since this is less than she would have received by following the same strategy in $t.1$ she will not do this.

In $t.3$ the possible states are either she deviated herself, she observed a deviation by r_j, both or neither. If neither then by not sending a message she will receive $\pi(2,2)$, and she will gain at most one message, thus since she wanted to send messages in $t.1$ she wants to send in this case. If both then either j will not send a message or in $t.4$ there will be two declarations that someone deviated, thus in both cases she will receive $\pi(2,2)$ and she will not send the message. If she observed only one deviation in $t.2$ the worst case is when it was an apparent deviation by r_j. In this case it could have been j not sending a message, but it could also have been that r_j just not sending a message (or lying). Thus she will do it since $\frac{1}{2}(\pi(1,1) + \rho_{it}(a_t) - c) + \frac{1}{2}\pi(2,2) \geq \pi(2,2)$.

In $t.4$ if she is supposed to send messages and has not observed a deviation then she will only bear the cost if j also sends the message, thus even if she is unsure of whether it was j's lie in $t.3$ or r_i's deviation that caused the problem she will agree to send off the message. If she observed a deviation then by sending off two messages she will achieve the payoff $\pi(1,1) + \rho_{it}(a_t)$ if she does not then she will receive $\pi(2,2)$ thus she will send the messages. If she lied in $t.3$ then even if she sends off messages she will not receive them, thus she will not.

In $t.5$ if she is supposed to send a message the worst she can expect is $\pi(1,1)+\rho_{it}(a_t)$, thus she will do it. If she is not supposed to then the worst case is when her partner claimed there was no deviation but she received a message in $t.4$. In this case the two possible states are that her partner deviated then lied to cover it up, or that r_j lied in $t.3$ then sent off a message in $t.4$ for no reason. Since it is infinitely more likely that someone lies to cover up a deviation than any other deviation she will believe that j deviated and then covered it up, and will not send a message. ■

Now I will show that truth telling is always optimal, or that the communication protocol is status revealing.

Lemma 4 *No player can manipulate the outcome by lying..*

Proof. In $t.1$ if they lie then they will not receive a reference in $t.3$. If i doesn't send a message in $t.2$ then covers up their original deviation in $t.3$ then $1(r_i \not\rightarrow j)+1(i \not\rightarrow \mu(r_i))=2$ and i's status will be 2 regardless. If i does send a message in $t.2$ then in $t.3$ j will not receive a reference and $\{r_i,\mu(r_i)\}$ will not send messages in $t.4$ thus i's status will be two regardless. Thus in every possible future i can not lie to cover up her original deviation, and she will be honest.

In $t.2$ they will be honest because if they lie then they will have to send two more messages or get status two. Since it is not worth getting status two at this point she will not lie.

In $t.3$ if she has detected no deviations she will not lie to make the situation worse. If she does lie then she will get status two, right now the worst she can expect is status one. If she has detected a deviation in $t.1$ then since the deviator will not send any messages if she lies to cover this up it will still result in her getting status two. If she has detected one deviation in $t.2$ then if she covers it up it will either be known to her partner (her partner not sending a message in $t.1$) or it will be reported in $t.4$ thus if she does not declare the deviation then she will have status two, if she does she will have status one and thus she will be honest. If she has detected two deviations in $t.2$ either her partner will declare it (he, i, or r_i deviated in $t.1$), or both $\{r_j,\mu(r_j)\}$ and $\{r_i,\mu(r_i)\}$ will declare the deviation in $t.4$. In either case she will get status two if she lies or not, thus she will be honest. In $t.4$ since all messages are joint messages lying in a message will be observed by i's partner and result in status two, thus she will not do it. If she is not supposed to send messages then sending them will not change her status, thus she will not do it. In $t.5$ if she lies in the message then she will get status two, if she already expects this status no lie will change it so she will be honest. If she expects status one then lying will just decrease her payoff and she will not do it.

Thus in every sub-period she will be honest. ■

5 Discussion.

The results are easily generalized with multi-period two-tiered punishments. The first punishment is for deviating in the stage game (or once in the communication protocol), the second punishment is for players who deviate multiple times during the communication protocol. When you allow this change the results extend to any two player stage game with a two dimensional payoff space. This requirement is met in any stage game that satisfies *non-equivalent utilities* and *pareto ranked payoffs*. Non-equivalent utilities rules out games of pure coordination (players' payoffs are affinely equivalent). In games of pure coordination punishing others is equivalent to being punished. In a repeated matching game this removes a player's incentive to cooperate if she must punish in future periods. Pareto ranked payoffs rules out generalized zero sum games. In zero sum games one person's gain is the other's loss—thus players will always lie about the past. To support every individually rational payoff the cost of messages must go to zero, but as was already pointed out any individual rational payoff *given* the costs of communication can always be supported.

The result also holds for a large range of matching rules. Choice base matching can be allowed as long as the matching is *only* affected by actions taken in the *stage game*. Messages can affect the matching by affecting social status, but messages that have no direct effect on status must be ignored. Most other maintained assumptions can be weakened.. For example if the complexity of the strategy is increased then players payoffs can still be affected by the actions of others when they do not interact. At a cost of even greater complexity joint messages do not have to be used. Perhaps the strongest assumption that must be maintained is that costs are only affected by the number of messages sent—not the identity of the sender and receiver. If this assumption is relaxed then a general matching rule could have people who deviate staying in the same small town forever and people who cooperate bouncing around the globe—needing to send very costly messages.

I would like to explain two elements of the communication protocol that initially seem counter intuitive. First even in games of one sided moral hazard both people get references. This is because the stage game *with*

communication has two sided moral hazard. Second, lying during the communication game is always as bad as deviating during the stage game. In a communication protocol people's reputations are made or broken by other people's word. In a world where someone's life can be ruined by a word saying the wrong thing is naturally as bad as doing the wrong thing.

Clearly there are other mechanisms that can be used to enforce cooperation. Checking references—the reference phase—is common, but other methods are often used in case of a deviation. Often when there is a physical record of the interaction a court can be used to verify what occurred—this was the method Milgrom, North and Weingast [13] found was used in the champagne fairs of medieval Europe. What this paper shows is that such records are unnecessary, even without them there can be cooperation.

One element that this paper does not illustrate is the possible interaction between economic geography and cooperation. Ahn and Suominen [2] could be considered a first step in this direction, but they do not consider how their "neighborhoods" develop or analyze them as an equilibrium. Recently attention has been paid to the development of social networks and their effect on equilibrium. Jackson and Wolinsky [9] and Bala and Goyal [3] have both developed interesting models of social networks, and Jackson and Watts [10] have combined this with evolutionary game theory to consider the development of networks. It would be natural to extend these papers to consider the case where the reward from forming networks interacts with the amount of cooperation that can be achieved..

Another question that should be explored is when these equilibria can arise. While many examples of social norms can be found, one can also find equivalent situations where social norms are not used. What can account for this heterogeneity? A recent work that lends insight into this question is Chwe [5]. He analyzes minimal social networks such that "revolutions" can occur. Here the "revolution" would be adaptation of a social norm, and weak social networks might lead to social norms not being used.

This paper presents the first self contained model of social cooperation. It develops a lower bound on the benefits of cooperation that can make self enforcing social agreements possible. When will this potential gain be realized, and how? Since cooperative behavior is integral to a lot of economic

interaction a deeper understanding of this point should be developed.

References

[1] Abreu, Dilip; Prajit K. Dutta; and Lones Smith. (1994) "The Folk Theorem for Repeated Games: A NEU Condition." *Econometrica*. 62:939-48.

[2] Ahn, Illtae and Matti Suominen. (1997) "Word-of-Mouth Communication and Community Enforcement." *Working Paper*.

[3] Bala, Venkatesh and Sanjeev Goyal: . (2000) "A Non-Cooperative Model of Network Formation." *Econometrica.* 68:1181-1230.

[4] Ben-Porath, Elchanan and Michael Kahneman. (1996) "Communication in Repeated Games with Private Monitoring." *Journal of Economic Theory*. 70:281-297.

[5] Chwe, Michael. (2000) "Communication and Coordination in Social Networks." *Review of Economic Studies.* 67: 1-16.

[6] Ellison, Glenn. (1994) "Cooperation in the Prisoner's Dilemma with Anonymous Random Matching." *Review of Economic Studies*. 61:567-588

[7] Hasker, Kevin. (2000) "Social Norms and Choice: A General Folk Theorem for Repeated Matching Games." *Working Paper.*

[8] Kandori, Michihiro. (1992) "Social Norms and Community Enforcement." *Review of Economic Studies.* 59: 63-80.

[9] Jackson, Matthew and Asher Wolinsky. (1996) "A Strategic Model of Economic and Social Networks." *Review of Economic Studies.* 71:44-74.

[10] Jackson, Matthew and Alison Watts. (2000) "The Evolution of Social and Economic Networks." *Working Paper.*

[11] Kandori, Michihiro; and Hitoshi Matsushima. (1998) "Private Observation, Communication and Collusion." *Econometrica*. 66:627-652.

[12] Matsushima, Hitoshi. (1991) "On the Theory of Repeated Games with Private Information." *Economics Letters.* 35:253-256.

[13] Milgrom, P., D. North and B. Weingast. (1995) "The Role of Institution in the Revival of Trade: The Law Merchant, Private Judges, and the Champagne Fairs." *Economics and Politics.* 2: 1-23