

Penalties and Rewards As Inducements To Cooperate*

Cheng-Zhong Qin[†]

This Version: February 3, 2005; First Version: July 16, 2002

Abstract

This paper considers self-stipulated penalties for defection and rewards for cooperation as inducements to cooperate in the prisoner's dilemma. The paper explicitly characterizes penalties and rewards that are necessary and sufficient to induce the players to cooperate. The characterization results imply a partition of prisoner's dilemma games into four classes according to whether both penalties and rewards can induce the players to cooperate; penalties but not rewards can induce the players to cooperate; rewards but not penalties can induce the players to cooperate; and neither penalties nor rewards can induce the players to cooperate. The paper also discusses implications of the results to "penalty clauses" in the law of contracts.

KEYWORDS: The prisoner's dilemma, Nash equilibrium, subgame-perfect equilibrium. (JEL C72, K12)

1. Introduction

A fundamental characteristic of the prisoner's dilemma is that each of two players could capture substantial gains through mutual cooperation but is tempted by even greater gains should the player defect while the other player cooperates. For either player the worst case is to cooperate while the other defects. The result is that both players defect, even though

*An earlier version of this paper was titled "Credible Commitments: Ex Ante Deterrence and Ex Post Compensation". I wish to thank participants at the Midwest Theory Conference at University of Minnesota, October 13-15, 2000, the South West Economic Theory Conference at Caltech, March 2-4, 2001, and seminar participants at UBC, UCR, and the University of Hong Kong for helpful comments and suggestions.

[†]Department of Economics, University of California, Santa Barbara, CA 93106.

mutual defection leaves each player with less than the player could have obtained through mutual cooperation.

Economic agents often face similar incentive problems. Economic transactions create opportunities for one party to take the benefit of the other party's performance. For example, when two parties agree to exchange certain goods or services to their mutual benefit, each party must decide whether to defect by defaulting or whether to act in good faith and risk the other party defaulting. Both parties are better off if neither defaults than if both default. However, each would be best off getting something for nothing and each is afraid the other will reason the same way. The result may well be that the parties are unable to carry out the exchange as arranged.

In practice courts can enforce contracts. When a contract is enforced, the victim of a breached contract has the right to obtain a legally enforceable remedy from the breacher. Contract remedies give the parties the power at the beginning of their interaction to alter payoffs arising from particular action configurations, in a way that induces the parties to adopt strategies that are most mutually beneficial. One particular class of contract remedies is known as that of "liquidated remedies". A liquidated remedy is one under which, if one party breaches, the victim can recover an amount predetermined by the parties themselves. (Cooter and Ulen 2000, pp. 225-237). A liquidated remedy that appears to exceed the actual harm is however considered as a penalty, and will not be enforced by the common law courts. The law's refusal to enforce penalty clauses is one of the most important counterexamples to the efficiency theory of the common law (see Posner 1992, p. 255). In this paper we consider a process for players to determine liquidated remedies to induce them to cooperate in the prisoner's dilemma. To see if players stipulate remedies to induce them to cooperate that are larger than actual harms, we assume whatever amounts the players may stipulate will be enforced, and we call such remedies penalties because

they may exceed actual harms.

Specifically, we assume that before playing the prisoner's dilemma game, each player can offer to pay the other player an amount on a take-it-or-leave-it basis should he defect while the other player cooperates.¹ Offers once made will be enforced.² Take the example of exchanging goods between two parties. Mutual defection in those circumstances means that the parties simultaneously default. When that happens, neither party can reasonably accuse the other of default. It follows that we can reasonably assume that neither party needs to pay any damages in the event of mutual default. This provides a justification for our assumption that neither player pays the penalty he offered to pay when both players defect. We also consider an alternative treatment under which a player offers to pay a penalty whenever he defects.

We add a pre-play stage in which the players decide penalties to offer to pay. The players play the prisoner's dilemma game upon observing each other's offers. We consider *subgame-perfect equilibrium* as the solution concept. We say that a given penalty configuration "induces" the players to cooperate if there is a subgame-perfect equilibrium that involves the players offering to pay penalties that constitute the given configuration and cooperating conditional on that configuration.

The necessary and sufficient conditions for a penalty configuration to induce the players to cooperate turn out to require that the penalty player i offers to pay be, on the one hand, large enough to deter player i from defecting and, on the other hand, not so large that player $j \neq i$ would rather have player i defect than have both of them cooperate. These

¹In practice, most written contracts use standard forms that include terms without possibility of modification. A common standard-form contract is the consumer product warranty. The warranty is drawn up by the supplier and presented to the purchaser on a take-it-or-leave-it basis. See Cooter and Ulen (2000, p. 278). The purchaser can reject the offer by refusing to transact with the supplier.

²For example, each player may post the amount he offered to pay for defection in an escrow, as in the case of depositing a bond, held by a neutral third party. In the field, this is observed in real estate and construction matters, where performance bonds and escrows are the rule.

upper and lower bounds correspond nicely to the *expectation remedies* and the *disgorgement remedies*, respectively. Since expectation damages measure actual harms, our result provides a theoretical support for penalty clauses in situations modeled by the prisoner’s dilemma (see Section 3.1.C for more discussion). Under the alternative treatment, the lower bound remains unchanged but the upper bound may sometimes be decreased.

We also consider another way to induce mutual cooperation in a prisoner’s dilemma game by letting each player make cooperation more beneficial to the other player. Specifically, as with the compensation mechanism in Varian (1994), we assume that before playing the prisoner’s dilemma game, each player can offer to pay the other player for cooperating on a take-it-or-leave-it basis.³ Offers once made will be enforced.⁴

We consider exactly the same timing of moves as with using penalties to induce the players to cooperate, except that what players now offer to pay are rewards for cooperation. We say that a given reward configuration “induces” the players to cooperate if there is a subgame-perfect equilibrium that involves the players offering to pay rewards that constitute the given configuration and cooperating conditional on that configuration.

The necessary and sufficient conditions for a reward configuration to induce the players to cooperate turn out to require that the reward from player i be, on the one hand, large enough to make it desirable for player $j \neq i$ to cooperate given that player i cooperates and, on the other hand, not so large as to cause one or both of the following possibilities.

³Varian (1994, pp. 1279) argues that the compensation mechanism provides a structure for negotiation and hence can be viewed as complementary to the *Coase Theorem*. He shows, among other things, that the compensation mechanism implements the efficient outcome of the prisoner’s dilemma with certain specifications of the payoffs in subgame-perfect equilibrium. Ziss (1997) shows that the efficient outcome is not among the set of possible subgame-perfect equilibrium outcomes for certain other specifications of the payoffs. In a recent experimental study of whether subjects actually manage to achieve mutual cooperation with the help of the compensation mechanism, Andreoni and Varian (1999) finds support for the mechanism. None of these papers characterizes rewards that are necessary and sufficient for inducing the players to cooperate.

⁴For example, each player may post the reward he offered to pay the other player for cooperation in an escrow held by a neutral third party, as in the case of depositing a bond.

First, the reward is so large that player i would rather have player j defect than have him cooperate. Second, the reward is so large relative to the reward from player j that player i prefers mutual defection to mutual cooperation. An implication is that with a given negotiation process, the result that costless negotiation leads to an efficient outcome no matter how the law assigns responsibility for damages as concluded in the Coase Theorem is sometimes invalid due to strategic behavior.

Our characterizations of penalties and rewards inducing the players to cooperate implies a partition of the prisoner's dilemma games into four classes according to whether both penalties and rewards can induce the players to cooperate; penalties but not rewards can induce the players to cooperate; rewards but not penalties can induce the players to cooperate; and neither penalties nor rewards can induce the players to cooperate.⁵

The rest of the paper is organized as follows. Section 2 introduces the prisoner's dilemma. Section 3 establishes respective necessary and sufficient conditions for penalties and rewards to induce the players to cooperate. Section 4 compares penalties with rewards as inducements to cooperate and section 5 summarizes the paper.

2. Prisoner's Dilemma

A generic prisoner's dilemma game has two players each of whom can either cooperate (action C) or defect (action D). Payoffs are as in Figure 1. The pair (D, D) is the only Nash equilibrium for the prisoner's dilemma game. This Nash equilibrium yields player i a payoff of P_i which is less than payoff R_i that player i could have obtained from mutual cooperation. We assume payoffs are transferable.

⁵Williamson (1983, pp 537-538) discusses the merit of crafting *ex ante* incentive structures for prisoner's dilemma.

		Player 2	
		<i>C</i>	<i>D</i>
	<i>C</i>	R_1, R_2	S_1, T_2
Player 1	<i>D</i>	T_1, S_2	P_1, P_2

Figure 1: A Generic Prisoner’s Dilemma Game with $S_k < P_k < R_k < T_k$, $k = 1, 2$.

The following payoff differences are used in our characterizations of penalties and rewards inducing the players to cooperate.⁶ First, player i can guarantee himself at least payoff P_i by defecting. In cooperating and in trusting player j to cooperate, he may get $L_i = P_i - S_i$ units less than he could guarantee himself by defecting. This payoff decrease is player i ’s “loss” from unilaterally cooperating. Second, if player j is going to cooperate, player i gets $G_i = T_i - R_i$ units more by defecting than he receives when he also cooperates. This payoff increase is player i ’s “gain” from unilaterally defecting.

3. Promoting Cooperation via Penalties for Defection and Rewards for Cooperation

In this section we characterize penalties and rewards that induce the players to cooperate in the prisoner’s dilemma.

3.1. A Penalty Scheme

Suppose before playing a prisoner’s dilemma game, each player can offer to pay a penalty to the other player on a take-it-or-leave-it basis should he defect while the other player cooperates. Suppose further payoffs from the play of the prisoner’s dilemma game and

⁶These payoff changes were used before in the literature on experimental study of the propensity to cooperate in symmetric prisoner’s dilemma (see Bonacich 1970).

penalty payments are additive, so that a penalty implies a payoff transfer from a defector to a cooperator. Let \mathcal{H}_i be the set of payoff transfers that are implied by penalties player i can offer to pay. There is no restriction on how much player i can offer to pay for unilaterally defecting, so that $\mathcal{H}_i = [0, \infty)$. A penalty configuration H changes the prisoner's dilemma game in Figure 1 into a game, $\Gamma^p(H)$, in Figure 2.

		Player 2	
		C	D
Player 1	C	R_1, R_2	$P_1 - L_1 + H_2, R_2 + G_2 - H_2$
	D	$R_1 + G_1 - H_1, P_2 - L_2 + H_1$	P_1, P_2

Figure 2: The Subgame $\Gamma^p(H)$ corresponding to Penalty Configuration H .

The timing of moves is that first players decide penalties to offer to pay on a take-it-or-leave-it basis and then play the prisoner's dilemma game upon observing each other's offers. It follows that the players can condition choices of actions in the prisoner's dilemma game on penalties they offer to pay. Denote by Φ_i a mapping that maps each penalty configuration $H \in \mathcal{H}_1 \times \mathcal{H}_2$ into a probability distribution $\Phi_i(H)$ over the action set $\{C, D\}$. Such a mapping specifies how player i 's action in the prisoner's dilemma game depends on penalty configurations. We call it an *action plan* for player i . A strategy for player i is a thus pair with a penalty he will offer to pay and an action plan he will subsequently use to guide his action choice. Denote by (H_i, Φ_i) a generic strategy for player i .

A. A Characterization of Penalties Inducing the Players to Cooperate

Given an action plan Φ_i for player i , we let $\Phi_i(C, H)$ denote the probability assigned to action C conditional on penalty configuration $H \in \mathcal{H}$.

DEFINITION 1: A penalty configuration $H^* = (H_1^*, H_2^*)$ induces the players to cooperate if there are action plans Φ_1^* and Φ_2^* such that (i) the strategy profile $((H_1^*, \Phi_1^*), (H_2^*, \Phi_2^*))$ is a subgame-perfect equilibrium and (ii) $\Phi_1^*(C, H^*) = \Phi_2^*(C, H^*) = 1$.

Suppose $H^* \in \mathcal{H}$ induces the players to cooperate. From Figure 2, player 1 receives payoff R_1 and player 2 receives payoff R_2 in the subgame-perfect equilibrium. Player 1's payoff becomes $G_1 - H_1^*$ if he defects in $\Gamma^p(H^*)$, given that player 2 cooperates. Thus H_1^* must satisfy $H_1^* \geq G_1$. Next, consider $H_2 \in \mathcal{H}_2$ with $H_2 < G_2$. It must be $\Phi_1^*(C, (H_1^*, H_2)) < 1$; otherwise, given player 1's strategy (H_1^*, Φ_1^*) , player 2 receives payoff $T_2 - H_2 > R_2$ by simply offering to pay H_2 instead of H_2^* and by defecting in $\Gamma^p(H_1^*, H_2)$. Notice that given player 1's strategy (H_1^*, Φ_1^*) , by offering to pay $H_2 < G_2$ and by cooperating in $\Gamma^p(H_1^*, H_2)$, player 2 receives expected payoff $\Phi_1^*(H_1^*, H_2)R_2 + [1 - \Phi_1^*(H_1^*, H_2)][S_2 + H_1^*]$. Since $\Phi_1^*(H_1^*, H_2) < 1$ as argued above, H_1^* must also satisfy $H_1^* \leq R_2 - S_2$ for $\Phi_1^*(H_1^*, H_2)R_2 + [1 - \Phi_1^*(H_1^*, H_2)][S_2 + H_1^*] \leq R_2$.

In summary, we have shown $G_1 \leq H_1^* \leq R_2 - S_2$. Similarly, $G_2 \leq H_2^* \leq R_1 - S_1$. We thus have:

LEMMA 1: Suppose $H^* \in \mathcal{H}$ induces the players to cooperate. Then, for $i \neq j$, $G_i \leq H_i^* \leq R_j - S_j$.

When $L_1 < G_2$, H_1^* must satisfy $L_2 \leq H_1^* \leq G_1$. To see this, consider $H_2 \in \mathcal{H}_2$ with $L_1 < H_2 < G_2$. Then from Figure 2, $H_2 > L_1$ implies that the unique optimal action for player 1 in $\Gamma^p(H_1^*, H_2)$ is C , given that player 2 defects. If $H_1^* > G_1$, then C remains to be the unique optimal action in $\Gamma^p(H_1^*, H_2)$ for player 1, given that player 2 cooperates. Thus together, $H_1^* > G_1$ and $H_2 > L_1$ imply that C is strictly dominant for player 1 in $\Gamma^p(H_1^*, H_2)$. Consequently, in this case, since $H_2 < G_2$ and since $\Phi^*(H_1^*, H_2)$ is a Nash equilibrium for $\Gamma^p(H_1^*, H_2)$, we must have $\Phi_1^*(C, (H_1^*, H_2)) = 1$ and $\Phi_2^*(C, (H_1^*, H_2)) = 0$.

This shows that by deviating from (H_2^*, Φ_2^*) to (H_2, Φ_2^*) with $L_1 < H_2 < G_2$, player 2 can guarantee himself a payoff equal to $R_2 + G_2 - H_2 > R_2$, while his payoff would at most be R_2 were he to offer to pay H_2^* . This contradicts the fact that H^* induces the players to cooperate. We conclude $H_1^* \leq G_1$ whenever $L_1 < G_2$.

Now suppose $H_1^* < L_2$. Then this together with $H_2 < G_2$ implies that action D is strictly dominant for player 2 in $\Gamma^p(H_1^*, H_2)$. In this case, since $H_2 > L_1$ and since $\Phi^*(H_1^*, H_2)$ is a Nash equilibrium for $\Gamma^p(H_1^*, H_2)$, it must be $\Phi_1^*(C, (H_1^*, H_2)) = 1$ and $\Phi_2^*(C, (H_1^*, H_2)) = 0$. Thus, again, by deviating from (H_2^*, Φ_2^*) to (H_2, Φ_2^*) with $L_1 < H_2 < G_2$, player 2 can guarantee himself a payoff equal to $R_2 + G_2 - H_2 > R_2$. This shows $H_1^* \geq L_2$ whenever $L_1 < G_2$.

To summarize, we have shown $L_2 \leq H_1^* \leq G_1$ whenever $L_1 < G_2$. By analogy, $L_1 \leq H_2^* \leq G_2$ whenever $L_2 < G_1$. We thus have:

LEMMA 2: *Suppose $H^* \in \mathcal{H}$ induces the players to cooperate. Then, $L_j \leq H_i^* \leq G_i$ whenever $L_i < G_j$ for $i \neq j$.*

Conditions in Lemmas 1 and 2 turn out to be not only necessary but also sufficient for penalty configuration H^* to induce the players to cooperate. This result is summarized in the following theorem.

THEOREM 1: *A penalty configuration H^* induces the players to cooperate if and only if*

$$G_i \leq H_i^* \leq R_j - S_j, \tag{1}$$

and

$$L_j \leq H_i^* \leq G_i \text{ whenever } L_i < G_j, \tag{2}$$

for $i \neq j$.

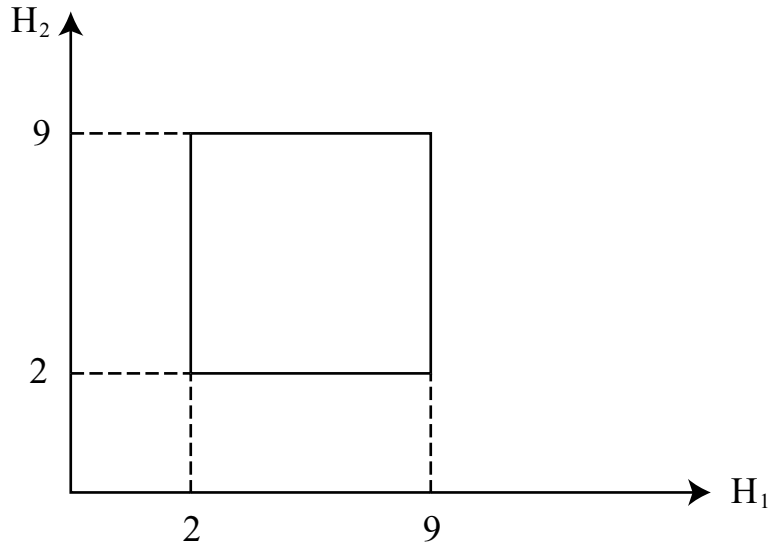


Figure 3: The Set of Penalty Configurations Inducing the Players to Cooperate in a Prisoner's Dilemma Game with $(S_1, P_1, R_1, T_1) = (2, 8, 11, 13)$ and $(S_2, P_2, R_2, T_2) = (0, 7, 9, 11)$.

PROOF: See the Appendix.

Notice that when $L_2 \geq G_1$ and $L_1 \geq G_2$, the set of penalty configurations inducing the players to cooperate is determined completely by condition (1). In that case, the set is rectangular. Figure 3 provides such an example.

The lower bound G_i on H_i^* is needed to deter player i from defecting. On the other hand, the upper bound $R_j - S_j$ on H_i^* is needed for player i to deter player j from offering to pay any penalty smaller than G_j . Player i 's action to defect conditional on such smaller penalties to be paid by player j provide the needed deterrence. However, such deterrence is not credible when $H_i^* > R_j - S_j$, because then player j would rather have player i defect in which case he receives $S_j + H_i^*$ by subsequently cooperating, than have both cooperate in which case he receives $R_j < S_j + H_i^*$. A similar explanation can be given for the necessity of the lower bound G_j and the upper bound $R_i - S_i$ on H_j^* .

REMARK 1: Notice that the compatibility of the upper and the lower bounds imply $T_1 + S_2 \leq R_1 + R_2$ and $T_2 + S_1 \leq R_1 + R_2$. Hence mutual cooperation must be efficient, in the sense that the sum of players' payoffs is the greatest, for there to be a penalty configuration to induce the players to cooperate.

B. A Variant of the Penalty Scheme

Consider a variant of the penalty scheme under which each player can offer to pay a penalty when he defects, regardless of whether the other player cooperates or not. Denote by $\Gamma^{p'}(H)$ the subgame conditional on penalty configuration H . It is shown in Figure 4.

		Player 2	
		C	D
Player 1	C	R_1, R_2	$P_1 - L_1 + H_2, R_2 + G_2 - H_2$
	D	$R_1 + G_1 - H_1, P_2 - L_2 + H_1$	$P_1 - H_1 + H_2, P_2 - H_2 + H_1$

Figure 4: The Subgame $\Gamma^{p'}(H)$ corresponding to Penalty Configuration H .

Under this variant of the penalty scheme necessary and sufficient conditions for a penalty configuration to induces the players to cooperate change to:

THEOREM 1': A penalty configuration H^* induces the players to cooperate under the variant of the penalty mechanism if and only if

$$G_i \leq H_i^* \leq \min\{L_i, R_j - P_j\}, \quad i \neq j. \quad (1')$$

PROOF: See the Appendix.

REMARK 2: Since $S_1 < P_1$ and $S_2 < P_2$, (1') implies $G_1 \leq H_1^* < R_2 - S_2$ and $G_2 \leq H_2^* < R_1 - S_1$. This means that (1') implies (1). However, (1') does not necessarily imply both (1) and (2) nor conversely. We show in Section 4 that (1') imposes stronger conditions than those that are necessary and sufficient for rewards to induce the players to cooperate.

C. A Contractual Interpretation of the Lower Bound G_i and the Upper Bound $R_j - S_j$ in (1)

In the law of contracts, an “expectation remedy” is defined as a payment that places the victim of a breached contract in the position he or she would have been in had the other party performed (Cooter and Ulen 2000, p. 226). Suppose player 1 and player 2 have agreed to cooperate. From Figure 1, if player 1 cooperates and player 2 defects, player 1’s payoff is S_1 . Player 1’s payoff would have been R_1 if player 2 had cooperated. Thus to place player 1 in the position he would have been in had player 2 cooperated, it would require that player 2 pay player 1 the amount $R_1 - S_1$. It follows that the expectation remedy when player 2 is held liable for is $R_1 - S_1$. Similarly, the expectation remedy when player 1 is held liable for is $R_2 - S_2$.

A “disgorgement remedy” is a payment paid to the victim of a breached contract to eliminate the breacher’s profit from wrong doing (Cooter and Ulen 2000, p. 234). Suppose player 1 and player 2 have agreed to cooperate. From Figure 1, player 2 gets G_2 units more from defecting given that player 1 cooperates. Thus to eliminate this gain from wrong doing (defecting), it would require that player 2 pay player 1 the amount equal to G_2 when he defects. It follows that the disgorgement remedy when player 2 is held liable for is G_2 . Similarly, the disgorgement remedy when player 1 is held liable for is G_1 .

Theorem 1 shows that when mutual cooperation is most efficient, to induce mutual cooperation, each player must offer to pay a penalty that is bounded below by the corresponding disgorgement remedy and above by the corresponding expectation remedy. When

a player pays more than the actual harm his action to defect inflicts upon the other player, not both players are playing the best they can individually if they still mutually cooperate. The incentive compatibility on the part of individual players' actions, as embodied in the notion of subgame-perfect equilibrium, provides a justification for penalty clauses in situations modeled by the prisoner's dilemma.

3.2. A Reward Scheme

Suppose as with the compensation mechanism of Varian (1994), before playing a prisoner's dilemma game each player can offer to pay a reward on a take-it-or-leave-it basis to the other whenever the other player cooperates. A reward from player i implies a payoff transfer from player i to player j conditional on j cooperating. Let \mathcal{H}_i be the set of payoff transfers that are implied by rewards that player i can offer to pay. As with the penalty scheme, there is no restriction on how much player i can offer to pay the other player for his cooperation, so that $\mathcal{H}_i = [0, \infty)$.

A reward configuration $H \in \mathcal{H} = \mathcal{H}_1 \times \mathcal{H}_2$ alters the prisoner's dilemma game in Figure 1 to $\Gamma^r(H)$ in Figure 5.

		Player 2	
		C	D
Player 1	C	$R_1 - H_1 + H_2, R_2 - H_2 + H_1$	$P_1 - L_1 + H_2, R_2 + G_2 - H_2$
	D	$R_1 + G_1 - H_1, P_2 - L_2 + H_1$	P_1, P_2

Figure 5: The Subgame $\Gamma^r(H)$ corresponding to Reward Configuration H .

The timing of moves is the same as before; first players decide on a take-it-or-leave-it basis what rewards to offer to pay and then play the prisoner's dilemma game upon observing

each other's offers. A strategy of player i must specify a reward that player i will offer to pay and an action he will take in the prisoner's dilemma game upon each reward configuration. We use the same notation (H_i, Φ_i) as before to denote a generic strategy for player $i = 1, 2$.

REMARK 3: *It should be pointed out that under the reward scheme, the higher the reward a player offers to pay the other player the lower payoff the player will receive when the other cooperates. Correspondingly, to induce the other player to cooperate, the reward must be minimal in the sense that it cannot be reduced and still induces the other player to cooperate.*

A. A Characterization of Rewards Inducing the Players to Cooperate

As with the penalty scheme, we consider the possibility to induce the players to cooperate through rewards for cooperation in subgame-perfect equilibrium.

DEFINITION 2: *A reward configuration $H^* = (H_1^*, H_2^*)$ induces the players to cooperate if there are action plans Φ_1^* and Φ_2^* such that (i) the strategy profile $((H_1^*, \Phi_1^*), (H_2^*, \Phi_2^*))$ is a subgame-perfect equilibrium and (ii) $\Phi_1^*(C, H^*) = \Phi_2^*(C, H^*) = 1$.*

Let $H^* \in \mathcal{H}$ be a reward configuration that induces the players to cooperate. Player 1 receives payoff $R_1 - H_1^* + H_2^*$ and player 2 receives payoff $R_2 - H_2^* + H_1^*$ in the subgame-perfect equilibrium. From Figure 5, player 1's payoff becomes $R_1 + G_1 - H_1^*$ if he defects in $\Gamma^r(H^*)$ given that player 2 cooperates. Thus, it must be $H_2^* \geq G_1$. Similarly, $H_1^* \geq G_2$. Next, let $H_2 \in \mathcal{H}_2$ be such that $H_2 < \min\{G_1, L_1\}$. Then action D is strictly dominant for player 1 in $\Gamma^r(H_1^*, H_2)$. Since $\Phi^*(H_1^*, H_2)$ is a Nash equilibrium for $\Gamma^r(H_1^*, H_2)$, $\Phi_1^*(C, (H_1^*, H_2)) = 0$. Consequently, given player 1's strategy (H_1^*, Φ_1^*) , player 2's payoff would be $P_2 - L_2 + H_1^*$ when he cooperates in $\Gamma^r(H_1^*, H_2)$ and P_2 when he defects in $\Gamma^r(H_1^*, H_2)$. We conclude that

H^* must satisfy $R_2 - H_2^* + H_1^* \geq S_2 + H_1^*$ and $R_2 - H_2^* + H_1^* \geq P_2$. This shows $H_2^* \leq R_2 - S_2$ and $H_2^* - H_1^* \leq R_2 - P_2$. Similarly, $H_1^* \leq R_1 - S_1$ and $H_1^* - H_2^* \leq R_1 - P_1$.

We have:

LEMMA 3: *Suppose H^* induces the players to cooperate. Then $G_i \leq H_j^* \leq R_j - S_j$ and $H_j^* - H_i^* \leq R_j - P_j$, for $i \neq j$.*

Conditions in Lemma 3 are not enough to guarantee that H^* induces the players to cooperate. Indeed, when $L_i > G_i$, reward $H_j^* > L_i$ is too large because it makes it strictly dominant for player i to subsequently cooperate. The reward is thus not minimal (see Remark 3). Theorem 2 below establishes necessary and sufficient conditions for a reward configuration to induce the players to cooperate.

THEOREM 2: *A reward configuration $H^* \in \mathcal{H}$ induces the players to cooperate if and only if*

$$G_i \leq H_j^* \leq R_j - S_j, \quad (3)$$

$$H_j^* - H_i^* \leq R_j - P_j, \quad (4)$$

$$H_j^* \leq G_i \text{ whenever } L_i \leq G_i, \quad (5)$$

and

$$H_i^* \leq L_j \text{ and } H_j^* \leq L_i \text{ whenever } L_i > G_i, \quad (6)$$

for $i \neq j$.

PROOF: See the Appendix.

When $L_1 > G_1$ and $L_2 > G_2$, the set of reward configurations inducing the players to cooperate is determined by conditions (3), (4), and (6) only. In that case, the set is

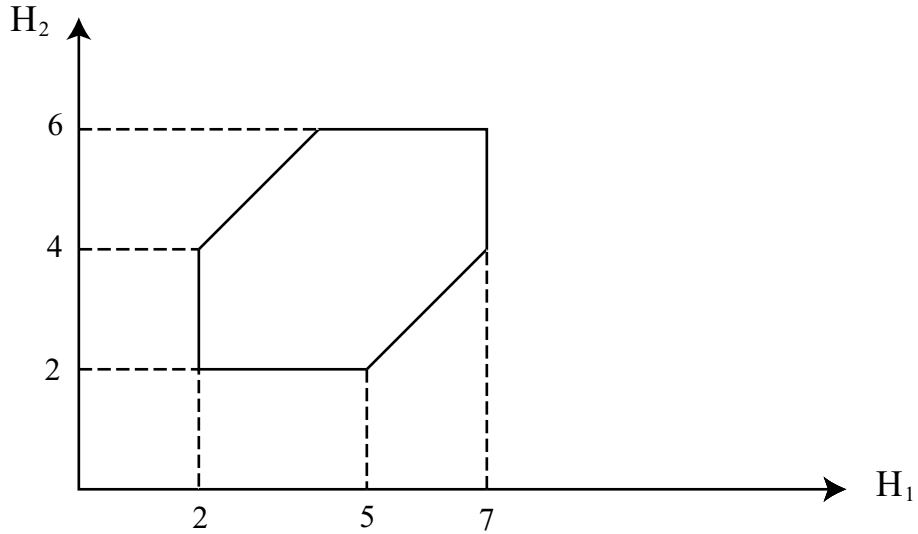


Figure 6: The Set of Reward Configurations Inducing the Players to Cooperate in the Prisoner's Dilemma Game in Figure 3.

sometimes an hexagon. Figure 6 provides such an example.

REMARK 4: *Condition (3) implies $T_1 + S_2 \leq R_1 + R_2$ and $T_2 + S_1 \leq R_1 + R_2$. This means that mutual cooperation must be efficient for there to be a reward configuration to induce the players to cooperate.*

When $H_1^* < G_2$, the reward is not large enough to make it desirable for player 2 to cooperate given that player 1 cooperates. On the other hand, when $H_1^* > R_1 - S_1$, the reward is so large that player 1 would rather have player 2 defect in which case he receives $P_1 - L_1 + H_2^*$ by subsequently cooperating, than have both of them subsequently cooperate in which case he receives $R_1 - H_1^* + H_2^* < P_1 - L_1 + H_2^*$. When $H_1^* - H_2^* > R_1 - P_1$, H_1^* is so large relative to H_2^* that player 1 prefers mutual defection from which he receives P_1 to mutual cooperation from which he receives $R_1 - H_1^* + H_2^* < P_1$. This explains the necessity of (3) and (4) for the case with $i = 2$ and $j = 1$. The case with $i = 1$ and $j = 2$ can be explained analogously.

When $H_2^* > \max\{L_1, G_1\}$, action C is strictly dominant for player 1 in $\Gamma^r(H_1, H_2^*)$, for all $H_1 \in \mathcal{H}_1$. In that case, H_2^* is not minimal (see Remark 3). Hence, $H_2^* \leq \max\{L_1, G_1\}$. By analogy, $H_1^* \leq \max\{L_2, G_2\}$. This implies $L_j < G_j$ whenever $H_i^* > L_j$ for $i \neq j$. Thus with $L_i > G_i$ and $H_i^* > L_j$, we would have both $L_i > G_i$ and $L_j < G_j$ were the players to be induced by configuration H^* . Under these conditions, however, no reward configuration can induce the players to cooperate (see Lemma 4 in the Appendix). This explains the necessity of (5) and (6).

B. Rewards for Cooperation as Coasian Payments

Rewards for cooperation we consider compensate the players for taking the initially less desirable action. They are Coasian payments involved in the Coase Theorem. Theorem 2 characterizes Coasian payments determined through a simple negotiation process that are necessary and sufficient for inducing the players to cooperate in the prisoner's dilemma. An implication of the characterization result is that when a negotiation process is prefixed, the conclusion that costless negotiation leads to an efficient outcome no matter how the law assigns responsibility for damages is sometimes in valid due to strategic behavior.

4. A Partition of Prisoner's Dilemma Games

This section is concerned with a partition of the prisoner's dilemma games according to whether both penalties and rewards can induce the players to cooperate; penalties but not rewards can induce the players to cooperate; rewards but not penalties can induce the players to cooperate; and neither penalties nor rewards can induce the players to cooperate.⁷

⁷Jackson and Wilkie (2002) also considers strategy dependent payoff transfers between the players. Payoff transfers in their paper are not restricted to be achievable through penalties for defection only or through rewards for cooperation only. We consider penalties for defection and rewards for cooperation as means of transferring payoffs between the players because of their relations to liquidated damages and Coasian payments. While Jackson and Wilkie focus on what feasible payoff allocations can be achieved in subgame-perfect equilibrium, we focus on what penalty and reward configurations that induce players to play a particular strategy profile; namely, mutual cooperation.

THEOREM 3: (i) *Necessary and sufficient conditions for penalties to induce the players to cooperate are*

$$(G_1 - L_2)(G_2 - L_1) \geq 0, G_i \leq R_j - S_j, i \neq j; \quad (7)$$

(ii) *necessary and sufficient conditions for penalties under the variant of the penalty scheme to induce the players to cooperate are*

$$G_i \leq \min\{L_i, R_j - P_j\}, i \neq j; \quad (8)$$

and (iii) *necessary and sufficient conditions for rewards to induce the players to cooperate are*

$$(G_1 - L_1)(G_2 - L_2) \geq 0, G_i \leq R_j - S_j, i \neq j. \quad (9)$$

PROOF: See the Appendix.

In words, (7) states that mutual cooperation is efficient and that players' gains from unilaterally defecting either are all no less than or are all no greater than their opponents' losses from unilaterally cooperating. Next, (8) states that players' gains from unilaterally defecting are all no greater than their own losses from unilaterally cooperating and their opponents' gains from mutual cooperation. Finally, (9) states that mutual cooperation is efficient and that players' gains from unilaterally defecting either are all no less than or are all no greater than their own losses from unilaterally cooperating.

Conditions (7) and (9) together imply the following comparison result: *First*, both penalties and rewards can induce the players to cooperate when mutual cooperation is efficient, $(G_1 - L_2)(G_2 - L_1) \geq 0$, and $(G_1 - L_1)(G_2 - L_2) \geq 0$. *Second*, penalties but not rewards can induce the players to cooperate when mutual cooperation is efficient, $(G_1 - L_2)(G_2 - L_1) \geq 0$, and $(G_1 - L_1)(G_2 - L_2) < 0$. *Third*, rewards but not penalties can

induce the players to cooperate when mutual cooperation is efficient, $(G_1 - L_2)(G_2 - L_1) < 0$, and $(G_1 - L_1)(G_2 - L_2) \geq 0$. *Fourth*, neither penalties nor rewards can induce the players to cooperate either when mutual cooperation is inefficient or when $(G_1 - L_2)(G_2 - L_1) < 0$, and $(G_1 - L_1)(G_2 - L_2) < 0$. The above four sets of conditions are mutually exclusive and jointly exhaustive. Hence, they characterize a partition of prisoner's dilemma games.

REMARK 5: *Notice that since $R_1 - P_1 < R_1 - S_1$ and $R_2 - P_2 < R_2 - S_2$, (9) is satisfied whenever (8) is satisfied. It follows that rewards can induce the players to cooperate whenever penalties under the variant of the penalty scheme can.*

5. Summary

In this paper we considered the possibility for inducing the players to cooperate in prisoner's dilemma through self-stipulated penalties for defection or rewards for cooperation. We completely characterized penalties and rewards that induce the players to cooperate. Our characterization results imply that for penalties to induce the players to cooperate it is necessary and sufficient that mutual cooperation is efficient and that players' gains from unilaterally defecting either are all no less than or are all no greater than their opponents' losses from unilaterally cooperating. On the other hand, for rewards to induce the players to cooperate it is necessary and sufficient that mutual cooperation is efficient and that players' gains from unilaterally defecting either are all no less than or are all no greater than their own losses from unilaterally cooperating.

Our characterization of penalties inducing the players to cooperate implies that when a player offers to pay a penalty larger than the corresponding expectation damages, mutual cooperation is not incentive compatible on the part of individual players. The lack of incentive compatibility provides a justification for penalty clauses in situations modeled by the prisoner's dilemma. Our characterization of rewards inducing the players to cooperate

implies that with a given negotiation process, the result that costless negotiation leads to an efficient outcome no matter how the law assigns responsibility for damages as concluded in the Coase Theorem is sometimes in valid due to strategic behavior.

Finally, our characterizations of penalties and rewards that induce the players to cooperate make it possible to experimentally study how the likelihood of mutual cooperation depends on the regions of penalties in (1)-(2) and rewards in (3)-(6). For example, would mutual cooperation be more likely the larger these regions are? These questions are experimentally investigated in Charness, Frechette, and Qin (2005).

Appendix: PROOFS

Let $U_1((H_1, \Phi_1), (H_2, \Phi_2))$ and $U_2((H_1, \Phi_1), (H_2, \Phi_2))$ denote the payoffs for player 1 and player 2, respectively, at strategy profile $((H_1, \Phi_1), (H_2, \Phi_2))$. They are the payoffs at action pair $(\Phi_1(H), \Phi_2(H))$ in $\Gamma^p(H)$ under the penalty scheme or in $\Gamma^r(H)$ under the reward scheme.

PROOF OF THEOREM 1: The necessity of (1) and (2) follows directly from Lemma 1 and Lemma 2. Thus it only remains to prove the sufficiency of these conditions.

Let $H^* \in \mathcal{H}$ be a penalty configuration satisfying conditions (1)-(2). For $H_2 \in \mathcal{H}_2$, let $\Phi_1^*(H_1^*, H_2)$ and $\Phi_2^*(H_1^*, H_2)$ be defined by

$$\Phi_1^*(C, (H_1^*, H_2)) = \begin{cases} 1 & \text{if } H_2 \geq G_2, \\ 0 & \text{if } L_1 \leq H_2 < G_2, \\ 0 & \text{if } H_1^* \leq L_2, H_2 < \min\{L_1, G_2\}, \\ \frac{H_1^* - L_2}{G_2 - H_2 + H_1^* - L_2} & \text{if } H_1^* > L_2, H_2 < \min\{L_1, G_2\}, \end{cases} \quad (10)$$

and

$$\Phi_2^*(C, (H_1^*, H_2)) = \begin{cases} 1 & \text{if } H_2 \geq G_2, \\ 1 & \text{if } L_1 \leq H_2 < G_2, \\ 0 & \text{if } H_1^* \leq L_2, H_2 < \min\{L_1, G_2\}, \\ \frac{L_1 - H_2}{H_1^* - G_1 + L_1 - H_2} & \text{if } H_1^* > L_2, H_2 < \min\{L_1, G_2\}. \end{cases} \quad (11)$$

For $H_1 \in \mathcal{H}_1$, let $\Phi_1^*(H_1, H_2^*)$ and $\Phi_2^*(H_1, H_2^*)$ be defined analogously. Finally, for $H \in \mathcal{H}$ with $H_1 \neq H_1^*$ and $H_2 \neq H_2^*$, let $(\Phi_1^*(H), \Phi_2^*(H))$ be any Nash equilibrium for $\Gamma^p(H)$. By (1), (10), and (11), $\Phi_1^*(C, H^*) = \Phi_2^*(C, H^*) = 1$.

Consider $H_2 \in \mathcal{H}_2$. Suppose first $H_2 \geq G_2$. By (1), $H_1^* \geq G_1$. It follows that (C, C) is a Nash equilibrium for $\Gamma^p(H_1^*, H_2)$ in this case. Suppose now $L_1 \leq H_2 < G_2$. In this case, $L_1 < G_2$. Hence, by (1) and (2), $L_2 \leq H_1^*$ and $H_1^* = G_1$. Consequently, (D, C) is a Nash equilibrium for $\Gamma^p(H_1^*, H_2)$. Suppose finally $H_2 < \min\{L_1, G_2\}$. In this case, if $H_1^* \leq L_2$, then $P_1 - L_1 + H_2 < P_1$ and $P_2 - L_2 + H_1^* \leq P_2$ implying that (D, D) is a Nash equilibrium for $\Gamma^p(H_1^*, H_2)$. If $H_1^* > L_2$, then (10) and (11) imply that given player 2's strategy (H_2, Φ_2^*) , action C and action D yield the same payoff to player 1 in $\Gamma^p(H_1^*, H_2)$. Similarly, given player 1's strategy (H_1^*, Φ_1^*) , action C and action D yield the same payoff to player 2 in $\Gamma^p(H_1^*, H_2)$. Thus, $\Phi^*(H_1^*, H_2)$ is a Nash equilibrium for $\Gamma^p(H_1^*, H_2)$.

In summary, we have shown that for any $H_2 \in \mathcal{H}_2$, $\Phi^*(H_1^*, H_2)$ as in (10) and (11) is a Nash equilibrium for $\Gamma^p(H_1^*, H_2)$. By analogy, for any $H_1 \in \mathcal{H}_1$, $\Phi^*(H_1, H_2^*)$ is a Nash equilibrium for $\Gamma^p(H_1, H_2^*)$. Thus, to complete the proof of the sufficiency, it only remains to show that players do not have any incentive to unilaterally change their offers.

To this end, consider $H_2 \in \mathcal{H}_2$. By (10), (11), and by Figure 2,

$$U_2((H_1^*, \Phi_1^*), (H_2, \Phi_2^*)) = \begin{cases} R_2 & \text{if } H_2 \geq G_2, \\ P_2 - L_2 + H_1^* & \text{if } L_1 \leq H_2 < G_2, \\ P_2 & \text{if } H_1^* \leq L_2, H_2 < \min\{L_1, G_2\}, \\ R'_2 & \text{if } H_1^* > L_2, H_2 < \min\{L_1, G_2\}, \end{cases} \quad (12)$$

where $R'_2 = \Phi_1^*(C, (H_1^*, H_2))R_2 + \Phi_1^*(D, (H_1^*, H_2))(S_2 + H_1^*)$. By (1), $H_1^* \leq R_2 - S_2$ which implies $R'_2 \leq R_2$. Consequently, by (12), $U_2((H_1^*, \Phi_1^*), (H_2, \Phi_2^*)) \leq R_2$. This shows that player 2 does not have any incentive to change his offer. Similarly, player 1 does not have any incentive to change his offer. \blacksquare

PROOF OF THEOREM 1': Let H^* be a penalty configuration inducing the players to cooperate. By Definition 1, there are action plans Φ_1^* and Φ_2^* such that $\Phi_1^*(C, H^*) = \Phi_2^*(C, H^*) = 1$ and the strategy profile $((H_1^*, \Phi_1^*), (H_2^*, \Phi_2^*))$ is a subgame-perfect equilibrium. From Figure 4, player 1 receives payoff R_1 and player 2 receives payoff R_2 in this subgame-perfect equilibrium. Player 1's payoff becomes $T_1 - H_1^*$ if he defects in $\Gamma^{p'}(H^*)$, given that player 2 cooperates. Thus it must be $H_1^* \geq G_1$.

Suppose $H_1^* > L_1$. Consider penalty $H_2 = 0$. From Figure 4, action D is strictly dominant for player 2 in $\Gamma^{p'}(H_1^*, 0)$. This together with $H_1^* > L_1$ implies that the unique Nash equilibrium for $\Gamma^{p'}(H_1^*, 0)$ is (C, D) . Consequently, given player 1's strategy (H_1^*, Φ_1^*) , player 2 receives payoff $T_2 > R_2$ by deviating from (H_2^*, Φ_2^*) to $(0, \Phi_2^*)$. This contradicts the fact that H^* induces the players to cooperate. Hence $H_1^* \leq L_1$ and $\Phi_1^*(C, (H_1^*, 0)) < 1$. Now observe that given player 1's strategy (H_1^*, Φ_1^*) , by offering to pay $H_2 = 0$ and by defecting in $\Gamma^{p'}(H_1^*, 0)$, player 2's payoff would become $\Phi_1^*(C, (H_1^*, 0))T_2 + \Phi_1^*(D, (H_1^*, 0))[P_2 + H_1^*]$. Since $\Phi_1^*(C, (H_1^*, 0)) < 1$, it must be $P_2 + H_1^* \leq R_2$ or equivalently $H_1^* \leq R_2 - P_2$ for $\Phi_1^*(C, (H_1^*, 0))T_2 + \Phi_1^*(D, (H_1^*, 0))[P_2 + H_1^*] \leq R_2$.

In summary, we have shown that H_1^* must satisfy $G_1 \leq H_1^* \leq \min\{L_1, R_2 - P_2\}$. By analogy, H_2^* must satisfy $G_2 \leq H_2^* \leq \min\{L_2, R_1 - P_1\}$.

Conversely, let H^* be a penalty configuration satisfying (1'). We show that H^* induces the players to cooperate. To this end, for $H_j \in \mathcal{H}_j$, let $\Phi_i^*(H_i^*, H_j)$ and $\Phi_j^*(H_i^*, H_j)$ be defined by

$$\Phi_i^*(C, (H_i^*, H_j)) = \Phi_j^*(C, (H_i^*, H_j)) = \begin{cases} 1 & \text{if } H_j \geq G_j, \\ 0 & \text{if } H_j < G_j. \end{cases} \quad (13)$$

For $H \in \mathcal{H}$ with $H_1 \neq H_1^*$ and $H_2 \neq H_2^*$, let $\Phi^*(H)$ be any Nash equilibrium for the subgame $\Gamma^{p'}(H)$. Notice that, since $H_1^* \geq G_1$ and $H_2^* \geq G_2$, (13) implies $\Phi_1^*(C, H^*) = \Phi_2^*(C, H^*) = 1$. Notice also that the conditions on H^* imply $G_1 \leq H_1^* \leq L_1$ and $G_2 \leq H_2^* \leq L_2$. Thus for $H_2 \in \mathcal{H}_2$, (C, C) is a Nash equilibrium for $\Gamma^{p'}(H_1^*, H_2)$ when $H_2 \geq G_2$; (D, D) is a Nash equilibrium for $\Gamma^{p'}(H_1^*, H_2)$ when $H_2 < G_2$. This shows that the action pair $\Phi^*(H_1^*, H_2)$ as in (13) is a Nash equilibrium for $\Gamma^{p'}(H_1^*, H_2)$, for all $H_2 \in \mathcal{H}_2$. By analogy, for all $H_1 \in \mathcal{H}_1$, the action pair $\Phi^*(H_1, H_2^*)$ is a Nash equilibrium for $\Gamma^{p'}(H_1, H_2^*)$. Thus to show that H^* induces the players to cooperate, it only remains to prove that the players do not have any incentive to unilaterally change penalties that constitute H^* .

Consider $H_2 \in \mathcal{H}_2$. By (13) and Figure 4, player 2's payoff at $((H_1^*, \Phi_1^*), (H_2, \Phi_2^*))$ is R_2 when $H_2 \geq G_2$ and his payoff is $P_2 - H_2 + H_1^*$ when $H_2 < G_2$. Since $H_1^* \leq R_2 - P_2$ and since player 2's payoff at strategy profile $((H_1^*, \Phi_1^*), (H_2^*, \Phi_2^*))$ is R_2 , player 2 has no incentive to unilaterally deviate from H_2^* . By analogy, player 1 has no incentive to unilaterally deviate from H_1^* . ■

To prove Theorem 2, we first prove the following two lemmas:

LEMMA 4: *There does not exist any reward configuration inducing the players to cooperate when $G_i < L_i$ and $G_j > L_j$ for $i \neq j$.*

PROOF: For $k = 1, 2$, define $f_k : \mathfrak{R} \rightarrow \mathfrak{R}$ by $f_k(x) = (x - P_k)/(R_k - P_k)$. Then, $f_k(P_k) = 0$, $f_k(R_k) = 1$,

$$f_k(S_k) = 1 - \frac{R_k - S_k}{R_k - P_k},$$

and

$$f_k(T_k) = 1 + \frac{G_k}{R_k - P_k}.$$

The pair (f_1, f_2) normalizes the payoffs for prisoner's dilemma into those as considered in Ziss(1997). Note also $L_k < G_k$ if and only if $f_k(P_k) - f_k(S_k) < f_k(T_k) - f_k(R_k)$. The rest of the proof can be completed by applying Proposition 1 of Ziss (1997). ■

LEMMA 5: *Suppose reward configuration H^* induces the players to cooperate. Then, $H_j^* \leq \max\{G_i, L_i\}$, for $i \neq j$.*

PROOF: Suppose $H_j^* > \max\{G_i, L_i\}$. Then from Figure 5, action C is strictly dominant for player i conditional on player j offering to pay H_j^* , regardless of the reward player i offers to pay. Thus given H_i^* , H_j^* can be reduced and still induces player i to cooperate. This shows that H_j^* is not minimal (see Remark 3). ■

PROOF OF THEOREM 2: Suppose that $H^* \in \mathcal{H}$ induces the players to cooperate. Then, the necessity of (3)-(4) follows from Lemma 3. The necessity of (5) follows from Lemma 5. To show the necessity of (6), notice that by Lemma 3 and Lemma 5, $H_i^* > L_j$ implies $L_j < G_j$. Hence, by Lemma 4 and Lemma 5, H^* cannot induce the players to cooperate if either $H_i^* > L_j$ and $L_i > G_i$ or if $H_j^* > L_i$ and $L_i > G_i$.

To prove the sufficiency, suppose that $H^* \in \mathcal{H}$ satisfies (3)-(6). For $H_2 \in \mathcal{H}_2$, define $\Phi_1^*(H_1^*, H_2)$ and $\Phi_2^*(H_1^*, H_2)$ by

$$\Phi_1^*(C, (H_1^*, H_2)) = \begin{cases} 1 & \text{if } H_2 \geq H_2^*, \\ 0 & \text{if } G_1 \leq H_2 < H_2^*, \\ 0 & \text{if } H_2 < G_1, L_2 \leq G_2, \\ 0 & \text{if } H_2 < G_1, L_2 > G_2, \end{cases} \quad (14)$$

and

$$\Phi_2^*(C, (H_1^*, H_2)) = \begin{cases} 1 & \text{if } H_2 \geq H_2^*, \\ 0 & \text{if } G_1 \leq H_2 < H_2^*, \\ 1 & \text{if } H_2 < G_1, L_2 \leq G_2, \\ 0 & \text{if } H_2 < G_1, L_2 > G_2. \end{cases} \quad (15)$$

For $H_1 \in \mathcal{H}_1$, let $\Phi_1^*(H_1, H_2^*)$ and $\Phi_2^*(H_1, H_2^*)$ be defined analogously. Finally, for $H \in \mathcal{H}$ with $H_1 \neq H_1^*$ and $H_2 \neq H_2^*$, let $\Phi^*(H)$ be any Nash equilibrium for $\Gamma^r(H)$. By (14) and (15), $\Phi_1^*(C, H^*) = \Phi_2^*(C, H^*) = 1$.

Consider $H_2 \in \mathcal{H}_2$. Suppose first $H_2 \geq H_2^*$. By (3), $H_1^* \geq G_2$ and $H_2^* \geq G_1$. Thus $H_2 \geq H_2^*$ implies that (C, C) is a Nash equilibrium for $\Gamma^r(H_1^*, H_2)$.

Suppose now $G_1 \leq H_2 < H_2^*$. In this case, since $H_2^* > G_1$, (5) implies $L_1 > G_1$. Hence, by (6), $H_1^* \leq L_2$ and $H_2^* \leq L_1$. Thus $H_2 < H_2^*$ implies that (D, D) is a Nash equilibrium for $\Gamma^r(H_1^*, H_2)$.

Suppose finally $H_2 < G_1$. In this case, if $L_2 \leq G_2$, then (3) and (5) imply $H_1^* = G_2$. It follows $H_1^* \geq L_2$. Together, $H_1^* \geq L_2$ and $H_2 < G_1$ imply that (D, C) is a Nash equilibrium for $\Gamma^r(H_1^*, H_2)$. If $L_2 > G_2$, then (6) implies $H_1^* \leq L_2$ and $H_2^* \leq L_1$. Since $H_2 < G_1 \leq H_2^*$, we have $H_2 < L_1$. With $H_1^* \leq L_2$ and $H_2 < L_1$, (D, D) is a Nash equilibrium for $\Gamma^r(H_1^*, H_2)$.

In summary, we have shown that, for any $H_2 \in \mathcal{H}_2$, $\Phi^*(H_1^*, H_2)$ as in (14) and (15) is a Nash equilibrium for $\Gamma^r(H_1^*, H_2)$. By analogy, for any $H_1 \in \mathcal{H}_1$, $\Phi^*(H_1, H_2^*)$ is a Nash equilibrium for $\Gamma^r(H_1, H_2^*)$. Thus, to complete the proof, it only remains to show that

players do not have any incentive to unilaterally change their offers To this end, consider $H_2 \in \mathcal{H}_2$. By (14), (15), and Figure 5,

$$U_2((H_1^*, \Phi_1^*), (H_2, \Phi_2^*)) = \begin{cases} R_2 + H_1^* - H_2 & \text{if } H_2 \geq H_2^*, \\ P_2 & \text{if } G_1 \leq H_2 < H_2^*, \\ P_2 - L_2 + H_1^* & \text{if } H_2 < G_1, L_2 \leq G_2, \\ P_2 & \text{if } H_2 < G_1, L_2 > G_2. \end{cases} \quad (16)$$

By (3) and (4), (16) implies $U_2((H_1^*, \Phi_1^*), (H_2, \Phi_2^*)) \leq R_2 + H_1^* - H_2^*$. This concludes that player 2 has no incentive to unilaterally change his offer. Similarly, player 1 has no incentive to unilaterally change his offer. ■

PROOF OF THEOREM 3: The necessity of (7) follows directly from (1) and (2). For the sufficiency, let $H_1^* = G_1$ and $H_2^* = G_2$. Then, with (7), $H^* = (H_1^*, H_2^*)$ satisfies (1) and (2). By Theorem 1, H^* induces the players to cooperate.

The necessity of (8) follows directly from (1'). For the sufficiency, set $H_1^* = G_1$ and $H_2^* = G_2$. It then follows from (8) that $H^* = (H_1^*, H_2^*)$ satisfies (1'). By Theorem 1', H^* induces the players to cooperate.

The necessity of (9) are implied by (3)-(6). To show the sufficiency, observe first that (9) implies either $L_1 \leq G_1$ and $L_2 \leq G_2$ or $L_1 \geq G_1$ and $L_2 \geq G_2$. Now consider configuration

$H^* = (H_1^*, H_2^*)$, where

$$H^* = \begin{cases} (G_2, G_1) & \text{if } L_1 \leq G_1, L_2 \leq G_2, \\ (\max\{G_2, G_1 - (R_2 - P_2)\}, G_1) & \text{if } L_1 = G_1, L_2 > G_2, \\ (\max\{G_2, G_1 - (R_2 - P_2)\}, \min\{H_1^* + R_2 - P_2, L_1\}) & \text{if } L_1 > G_1, L_2 \geq G_2. \end{cases} \quad (17)$$

Assume first $L_1 \leq G_1$ and $L_2 \leq G_2$. In this case, (6) is irrelevant. By (17), $H_1^* - H_2^* \leq R_1 - P_1$ if and only if $G_2 \leq T_1 - P_1$. Since $G_2 \leq R_1 - S_1$ by (9) and since $L_1 \leq G_1$ implies $R_1 - S_1 \leq T_1 - P_1$, we have $G_2 \leq T_1 - P_1$. Hence $H_1^* - H_2^* \leq R_1 - P_1$. Similarly, $H_2^* - H_1^* \leq R_2 - P_2$. This shows that H^* satisfies (4). By (9) and (17), H^* also satisfies (3) and (5). It follows from Theorem 2 that H^* induces the players to cooperate.

Assume now $L_1 = G_1$ and $L_2 > G_2$. In this case, $G_1 - (R_2 - P_2) \leq L_2$ if and only if $G_1 \leq R_2 - S_2$. By (9), this latter inequality is satisfied. Now the inequalities $G_1 - (R_2 - P_2) \leq L_2$ and $G_2 < L_2$ together with (17) imply $H_1^* \leq L_2$. On the other hand, $G_1 = L_1 < R_1 - S_1$. Hence, since $G_1 - (R_2 - P_2) < G_1$, (9) and (17) imply $G_2 \leq H_1^* \leq R_1 - S_1$, $G_1 \leq H_2^* \leq R_2 - S_2$, and $H_2^* \leq L_1$. This shows H^* satisfies (3), (5), and (6). Since $H_1^* \geq G_1 - (R_2 - P_2)$ and $H_2^* = G_1$, we have $H_2^* - H_1^* \leq R_2 - P_2$. On the other hand, $H_1^* - H_2^* \leq R_1 - P_1$ if and only if $G_2 \leq T_1 - P_1$. Since $G_2 \leq R_1 - S_1$ by (9) and since $L_1 = G_1$ implies $R_1 - S_1 = T_1 - P_1$, we have $H_1^* - H_2^* \leq R_1 - P_1$. This shows H^* also satisfies (4). Hence, by Theorem 2, H^* induces the players to cooperate.

Assume finally, $L_1 > G_1$ and $L_2 \geq G_2$. In this case, $G_1 < L_1 < R_1 - S_1$. Thus by (9) and (17), $G_2 \leq H_1^* \leq R_1 - S_1$, $H_1^* \leq L_2$, and $H_2^* \leq L_1$. Notice that $L_2 = G_2$ implies $T_2 - P_2 = R_2 - S_2$. Hence, $G_2 \geq G_1 - (R_2 - P_2)$ implying $H_1^* = G_2$. Since $L_1 > G_1$, we have $H_2^* = G_1$ if $H_1^* = G_1 - (R_2 - P_2)$; $H_2^* = \min\{T_2 - P_2, L_1\}$ if $H_1^* = G_2$. On the other hand, $H_1^* = G_2$ if and only if $T_2 - P_2 \geq G_1$. Thus since $L_1 > G_1$, we have $H_2^* \geq G_1$. Next, $L_2 \geq G_2$ implies $R_2 - S_2 \geq T_2 - P_2$. This together with since $R_2 - S_2 \geq G_1$ from (9) implies

$H_2^* \leq R_2 - S_2$. In summary, we have shown that H^* satisfies (3), (5), and (6). Finally, $H_1^* - H_2^* = -(R_2 - P_2)$ if $H_1^* = G_1 - (R_2 - P_2)$; $H_1^* - H_2^* = \max\{P_2 - R_2, G_2 - L_1\}$ if $H_1^* = G_2$. Since $R_2 > P_2$ and since, by (9), $G_2 \leq R_1 - S_1$, we have $H_1^* - H_2^* \leq R_1 - P_1$. Furthermore, since $H_2^* \leq H_1^* + R_2 - P_2$, we have $H_2^* - H_1^* \leq R_2 - P_2$. This shows H^* also satisfies (4). Hence by Theorem 2, H^* induces the players to cooperate. ■

References

- [1] Andreoni, James and Varian, Hal. "Pre-Play Contracting in the Prisoner's Dilemma." *Proceedings of National Academy of Sciences*. August/September 1999, 96, pp. 10933-10938.
- [2] Bonacich, Phillip. "Putting the Prisoner's Dilemma back into Prisoner's Dilemma." *Conflict Resolution*, 14 (1970), 379-387.
- [3] Charness, Gary, Fréchette, Guillaume, and Qin, Cheng-Zhong. "Endogenous Transfers in the Prisoner's Dilemma Game: An Experimental Test Of Cooperation And Coordination". UCSB 2005 Working Paper.
- [4] Coase, Ronald H. "The Problem of Social Cost." *The Journal of Law and Economics*, 3 (1960), 1-44.
- [5] Cooter, Robert and Ulen, Thomas. *Law and Economics*. Addison-Wesley, Third Edition, 2000. Cambridge: Cambridge University Press, 1990, 90-143.
- [6] Jackson, Matthew O. and Wilkie, Simon. "Endogenous Games and Mechanisms: Side Payments Among Players." mimeo: Caltech, 2001.
- [7] Posner, Richard A. *Economic Analysis of Law*. Boston: Little, Brown and Company, Fourth Edition, 1992.
- [8] Varian, Hal R. "A Solution to the Problem of Externalities When Agents Are Well-Informed." *American Economic Review*, December 1994, 84(5), pp. 1278-93.
- [9] Williamson, Oliver E. "Credible Commitments: Using Hostages to Support Exchange." *American Economic Review*, September 1983 73(4), pp. 519-540.

- [10] Ziss, Steffen. "A Solution to the Problem of Externalities When Agents Are Well-Informed: Comment." *American Economic Review*, March 1997, 87(1), pp. 231-235.