Purely Noncooperative Farsighted Stable Set in an n-Player Prisoners' Dilemma

Nakanishi, Noritsugu

Purely Noncooperative Farsighted Stable Set in an n-Player Prisoners' Dilemma

Noritsugu NAKANISHI*

Graduate School of Economics, Kobe University
Rokkodai-cho 2-1, Nada-ku, Kobe 657-8501
JAPAN

July 2007†

JEL Classification: C71, C72, C79

Keywords: prisoners' dilemma, farsighted stability, theory of social situations (TOSS), von Neumann-Morgenstern stable set.

Abstract

We examine an n-player prisoners' dilemma game in which only individual deviations are allowed, while coalitional deviations (even non-binding ones) are not, and every player is assumed to be farsighted enough to understand not only the direct outcome of his own deviation, but also the ultimate outcome resulting from a chain of subsequent deviations by other players. By constructing a purely noncooperative farsighted stable set concretely, we prove its existence and uniqueness; further, we show that it supports the "all-defection" outcome as well as at least one Pareto-efficient outcome, which may or may not be the "all-cooperation" outcome.

*The author is very grateful to Professor Eiichi Miyagawa for his insightful comments. He also acknowledges financial supports from Japan Society for the Promotion of Science (Grant-in-Aid for Scientific Research (C), No. 18530175). Address for correspondence: Noritsugu Nakanishi, Graduate School of Economics, Kobe University, Rokkodai-cho 2-1, Nada-ku, Kobe 657-8501, JAPAN. Tel: +78-803-6837 (Office Direct). Fax: +78-803-7293 (Faculty Office). E-mail: `nakanishi@econ.kobe-u.ac.jp`

†Printed: July 24, 2007 (ver. 1, Nov. 2006; ver. 2, Jan. 2007).

1 Introduction

Individual interests and social desirability often disagree with each other. Such a situation is lucidly illustrated by the prisoners' dilemma (PD) game and, to resolve the discrepancy between individual/social desirability, many approaches have been proposed and examined within the framework of PD games: Repetition of one-shot PD game (see, for example, Fudenberg and Maskin [1986]) and introduction of preplay negotiation among players (see Kalai [1981]) are well-known ones. Other possible approaches can be found in Okada (1993) and Nishihara (1997). The former has considered an institutional arrangement game in which players can establish an enforcement agency in advance of the actual play of the PD game; the latter has introduced sequential moves into the PD game.

All of the above studies have adopted the Nash equilibrium and its variants as the solution concept. There is another line of research that adopts the von Neumann-Morgenstern (vN-M) stable set and some related concepts as the solution concept. Although the vN-M stable set had been introduced by von Neumann and Morgenstern (1953) originally as the solution concept for games in characteristic function form, the theory of social situations developed by Greenberg (1990) has opened the way to applying it (at least, its spirit) to games in other forms.[1] Examples of PD games along this line of research include Arce (1994), Muto (1993)[2], and Nakanishi (2001), each of which has shown that some Pareto-efficient outcomes can be supported by the vN-M stable set for each of the PD games they analyzed.[3]

By Harsanyi (1974) and, later, by Chwe (1994), the notion of the vN-M stable set has been criticized for its lack of farsightedness. In the original definition of the vN-M stable set, a player (or a group of players) contemplating a deviation from an outcome is supposed to consider whether the immediate outcome that is realized just after his deviation is better than the initial outcome (i.e., whether the immediate outcome *directly dominates* the initial outcome). Each player ignores the possibility of subsequent deviations by other players that may occur after his own (first) deviation. In this sense, players are considered to be myopic.

[1] One of the solution concepts in TOSS is called the optimistic stable standard of behavior (OSSB), which is closely related to (in a sense, it is equivalent to) the vN-M stable set. For formal relationship between the vN-M stable set and the OSSB, see Greenberg (1990, Chap. 4).

[2] Actually, Muto (1993) has not adopted the vN-M stable set as the solution concept. He has shown that the set of the conservative Markov perfect equilibrium of the PD game he analyzed can be seen as the vN-M stable set.

[3] Examples of other games along this line of research can be found in Muto and Okada (1996, 1998) and Nakanishi (1999).

In contrast, if a player is farsighted enough to understand not only the immediate outcome but also the ultimate outcome resulting from a chain of deviations, then the player may deviate from an outcome to another that is not better than the initial outcome, anticipating that subsequent deviations by other players bring about an ultimate, stable outcome better than the initial outcome; conversely, a player may refrain from deviating from an outcome even if the immediate outcome is certainly preferable to the initial outcome, anticipating that subsequent deviations result in an ultimate outcome that is worse than the initial outcome. What is important for a farsighted player is whether the ultimate outcome, not the immediate outcome, is better than the initial outcome (i.e., whether the ultimate outcome *indirectly dominates* the initial outcome). To capture the farsightedness of the players just described, Harsanyi (1974) has defined an indirect dominance relation and proposed the notion of the strictly stable set based on his indirect domination; Chwe (1994) has also defined another indirect dominance relation and examined the farsighted stable set based on his indirect domination as well as his new solution concept called the largest consistent set.

Recently, taking account of the criticism raised by Harsanyi and Chwe, Suzuki and Muto (2005) have examined an n-player PD game in which players are farsighted. They have shown that essentially any individually rational and Pareto-efficient outcome itself forms a farsighted stable set as defined by Chwe (1994). In their model, however, players are supposed to be able to communicate freely, form and dissolve coalitions, and make joint deviations. Although these cooperative elements do not mean that binding agreements are possible, they divert the model from the *noncooperative* situation that the original PD game was intended to represent.

In this paper, in line with the original PD game, we modify (or restrict) the PD game analyzed by Suzuki and Muto (2005) so that players can make only individual deviations, but not coalitional deviations, and examine the existence and efficiency of the farsighted stable set in a modified, purely noncooperative setting. By constructing a purely noncooperative farsighted stable set concretely, we prove its existence and uniqueness; further, we show that it supports the "all-defection" outcome as well as at least one Pareto-efficient outcome, which may or may not be the "all-cooperation" outcome.

Diamantoudi (2005) and Kamijo and Muto (2006) have dealt with similar issues in the context of a price-leadership cartel formation model in which players (firms) decide whether to enter or exit from the existing dominant cartel. Incorporating the notion of farsightedness into the price-leadership model by D'Aspremont et al. (1983), Diamantoudi (2005) has shown that there exists a unique, non-empty set of (farsighted) stable cartels, but she has not specified the stable cartels nor shown the efficiency properties of them.

On the other hand, Kamijo and Muto (2006) have modified Diamantoudi's model to allow for coalitional deviations, which are not allowed in our model, and shown essentially the same results by Suzuki and Muto (2005).

2 Definitions and notation

Most of definitions and notation in this paper are borrowed from Suzuki and Muto (2005). Let $N \equiv \{1, 2, \ldots, n\}$ be the set of players. The strategy set of player i is denoted by $X_i \equiv \{C, D\}$, where C stands for "cooperation" and D for "defection." $X \equiv \prod_{i \in N} X_i$ denotes the set of outcomes. For $x \in X$, with a slight abuse of notation, $C(x)$ denotes the set of players who play C in x. We call a player who plays C a "cooperator." The expression $|C(x)|$ means the cardinality of $C(x)$, that is, the number of cooperators in x. For each $h = 0, 1, \ldots, n$, let $V(h)$ be a subset of such outcomes that the number of cooperators in each of them is equal to h, that is, $V(h) \equiv \{x \in X |\ h = |C(x)|\}$. $V(0)$ is a singleton that consists of the outcome of all D; $V(n)$ is also a singleton that consists of the outcome of all C. As we will show soon, $V(h)$ plays an important role in proving the existence of the purely noncooperative farsighted stable set for the n-player prisoners' dilemma game.

The payoff of player i depends not only on his own strategy, but also on the number of cooperators. Then, for $x \in X$, the payoff function $u_i \colon X \to \mathbb{R}$ of player i can be written as $u_i(x) \equiv f(x_i, h)$, where x_i is the action (C or D) taken by player i and h is the number of cooperators in x.[4] We assume that the payoff functions are identical for all players. Following Suzuki and Muto (2005), we assume the following properties of f:[5]

Assumption 1.

(i). $f(D, h) > f(C, h+1)$ *for all* $h = 0, 1, 2, \ldots, n-1$,

(ii). $f(C, n) > f(D, 0)$,

(iii). $f(C, h)$ *and* $f(D, h)$ *are increasing in* h.

Assumption 1-(i) means that each player prefers playing D to playing C regardless of the other players' actions. Assumption 1-(ii) means that, for

[4]It should be noted that the definition of f in this paper is slightly different from f in Suzuki and Muto (2005). They defined f as a function of the action taken by player i and the number h of cooperators in x *other than* player i. In our definition of f, on the other hand, h includes player i himself when it plays C.

[5]Although the expressions of the properties are different from those in Suzuki and Muto (2005) because of the difference in the way of defining f, all of these properties in our paper are equivalent to those of Assumption 2.1 in Suzuki and Muto (2005).

each player, *full cooperation* is better than *no cooperation*. Assumption 1-(iii) means that, given a player's action, an increase in the number of cooperators is preferable to the player. Assumptions 1-(i) and 1-(iii) together imply that $f(D,h) > f(C,h)$ as far as both sides of the inequality are defined well. Further, to simplify the exposition, we assume the following:

Assumption 2. *$f(C,h) \neq f(D,k)$ for any $h = 1,\dots,n$ and $k = 0,\dots,n-1$.*

Although it is logically possible to have $f(C,h) = f(D,k)$ for some pair of h and k, this is merely a coincidence. Even if $f(C,h) = f(D,k)$ holds for some pair of h and k, this equality can be easily broken by a very small perturbation of fundamental parameters of the model. In this paper, we shall ignore these degenerate cases.

Let $G = (N, \{X_i\}_{i\in N}, \{u_i\}_{i\in N})$ be an n-player prisoners' dilemma game. For any outcomes $x, y \in X$, we say that outcome y can be *induced* from x through player i if $x_j = y_j$ for all $j \neq i$. In other words, player i can change an outcome x to another y by switching his own action (from C to D, or vice versa). We write this inducement relation as $x \xrightarrow{i} y$. Further, we say that y *indirectly dominates* x if there exist a sequence of outcomes $x^0, x^1, \dots, x^p$ with $x^0 = x$ and $x^p = y$ and a corresponding set of players $\{i_1, i_2, \dots, i_p\}$ such that $x^{r-1} \xrightarrow{i_r} x^r$ and $u_{i_r}(x^{r-1}) < u_{i_r}(y)$ for all $r = 1, 2, \dots, p$. When y indirectly dominates x, we write $x \ll y$. The pair of the set of outcomes and the indirect dominance relation, $(X, \ll)$, is called the abstract system associated with the n-player prisoners dilemma game G. It should be noted that, unlike Suzuki and Muto (2005), no coalitional moves are allowed in our model. In this sense, our model can be said to be purely noncooperative.

A subset K of X is said to be a *purely noncooperative farsighted stable set* for the n-player prisoners' dilemma game G if it satisfies the following two conditions: (a) for any $x, y \in K$, neither $x \ll y$ nor $y \ll x$ holds; (b) for any $x \in X \setminus K$, there exists $y \in K$ such that $x \ll y$. Conditions (a) and (b) are called *internal stability* and *external stability*, respectively.

3 Theorem

Before proceeding, we need to introduce some additional definitions. Let us define nonnegative integers h_t $(t = 0, 1, \dots)$ recursively as follows:

$$h_0 \equiv 0, \tag{1}$$

$$h_t \equiv \min\{h \mid f(C,h) > f(D,h_{t-1})\}, \ t = 1, 2, \dots. \tag{2}$$

Because both $f(C,h)$ and $f(D,h)$ are increasing in h and because $f(C,h) < f(D,h)$, we can easily verify that $h_{t-1} < h_t$ and that there exists a finite

integer T at which the recursive procedure stops. Let $H \equiv \{h_0, h_1, \ldots, h_T\}$ be the set of the above-defined integers. Note that H is always non-empty.

We are now in a position to state our theorem:

Theorem 1. *There exists a unique, purely noncooperative farsighted stable set $K^* \subset X$ for the n-player prisoners' dilemma game G, which is given by*

$$K^* = \bigcup_{h_t \in H} V(h_t). \tag{3}$$

Before proving the theorem, we need to show a lemma, which (partially) characterizes the indirect domination $\ll$ defined on X.

Lemma 1. *In a sequence of outcomes that realizes $x \ll y$, no player in the sequence switches his action from D to C.*

Proof. Let $x = x^0, x^1, x^2, \ldots, x^p = y$ be a sequence of outcomes that realizes $x \ll y$. Suppose, in negation, that there exist players who switch their actions from D to C and player i_k in step k is the last one of such players in the sequence. Note that $x^{k-1} \xrightarrow{i_k} x^k$. By the definition of indirect domination, we have

$$u_{i_k}(x^{k-1}) < u_{i_k}(y). \tag{4}$$

Since player i_k plays D in x^{k-1}, the above inequality and Assumption 1 together imply that the number of cooperators in y is strictly larger than that in x^{k-1}. Since no player switches to C after step k, the only possibility is that step k is the last step in the sequence, that is, $y = x^k$ (or $k = p$). But then, player i_k is making himself worse off in the last step and, hence, inequality (4) does not hold—a contradiction. □

Lemma 1 shows that if $x \ll y$ is realized by a sequence of outcomes, then the number of cooperators monotonically decreases one by one, from $|C(x)|$ to $|C(y)|$, along the sequence. Consequently, we obtain an immediate corollary of Lemma 1: $x \ll y$ implies $|C(x)| > |C(y)|$. Now we give a proof of the theorem.

Proof of Theorem 1. [**Internal stability**] Take any x, y in K^*. Without loss of generality, we can assume $|C(x)| = h_t \leq h_s = |C(y)|$. By Lemma 1, $x \ll y$ is not possible. If $h_t = h_s$, then $y \ll x$ is not possible, either. Then, to prove the internal stability, it remains to show that $y \ll x$ is not possible when $h_t < h_s$.

Suppose, in negation, that $y \ll x$ and $h_t < h_s$. Let $y = x^0, x^1, \ldots, x^p = x$ be a sequence of outcomes that realizes $y \ll x$ and let player i be the first

player in the sequence. Since player i plays C in y and D in x, the definition of H and Assumption 1 together imply

$$u_i(x) = f(D, h_t) < f(C, h_{t+1}) < \cdots < f(D, h_{s-1}) < f(C, h_s) = u_i(y), \quad (5)$$

a contradiction. Hence, $y \ll x$ can not be true. The internal stability of K^* obtains.

[**External stability**] Take an arbitrary outcome $x \in X \setminus K^*$. Let us define an integer $h_m \equiv \max\{h \in H |\ h < |C(x)|\}$ and set $k = |C(x)| - h_m$. Consider a set of k players in $C(x)$, $I \equiv \{i_1, i_2, \ldots, i_k\} \subset C(x)$, and a sequence of outcomes $x^0, x^1, \ldots, x^k$ such that $x^{r-1} \xrightarrow{i_r} x^r$ with $x^0 = x$ and $x^r \in V(|C(x)|-r)$ for all $r = 1, 2, \ldots, k$. In the sequence, each moving player in I switches his action from C to D. Clearly, we have $x^k \in V(h_m) \subset K^*$. Now we show that x^k indirectly dominates x.

Note that both $i_r \in C(x^{r-1})$ and $i_r \notin C(y)$ hold for all $r = 1, 2, \ldots, k$. Therefore, we have both $u_{i_r}(x^{r-1}) = f(C, |C(x)| - r + 1)$ and $u_{i_r}(y) = f(D, h_m)$ for all $i_r \in I$. By the definition of H and by Assumptions 1 and 2, we have

$$u_{i_r}(x^k) = f(D, h_m) > f(C, |C(x)| - r + 1) = u_{i_r}(x^{r-1}), \quad \forall i_r \in I. \quad (6)$$

Hence, $x^k \in K^*$ indirectly dominates x. The external stability of K^* obtains.

[**Uniqueness**] Let K be a purely noncooperative farsighted stable set for the n-player PD game. To prove the uniqueness, it suffices to show that $K = K^*$.

First, we show that the "all-defection" outcome $x^D \equiv (D, D, \ldots, D)$ is included in K. Because $0 = |C(x^D)| \leq |C(y)|$ for all $y \in X$, then, by Lemma 1, no outcome can indirectly dominate x^D.[6] By the external stability of K, x^D must be included in K. By definition, $V(h_0) = V(0) = \{x^D\}$ and, then, we have $V(h_0) \subset K$.

Take an arbitrary outcome $x \in X$ with $h_0 < |C(x)| < h_1$ and set $k = |C(x)| - h_0$. Consider a k-step sequence $x = x^0, x^1, \ldots, x^k$, realized by k players in $C(x)$, in which each moving player switches his action from C to D. By definition, $x^k \in V(h_0) \subset K$. (Actually, $x^k = x^D$ in this case.) By the definition of H, we have $u_{i_r}(x^{r-1}) = f(C, |C(x)|-r+1) < f(D, h_0) = u_{i_r}(x^k)$ for all $r = 1, 2, \ldots, k$. That is, x is indirectly dominated by an outcome x^k in K. Hence, $x \notin K$.

[6] This implies x^D is included in the "purely noncooperative farsighted core" for the abstract system $(X, \ll)$ associated with G, which is a set of outcomes in X that are not indirectly dominated. In fact, the singleton set $\{x^D\}$ is the purely noncooperative farsighted core for $(X, \ll)$.

In turn, take an arbitrary outcome $y \in V(h_1)$. If $y \ll z$ for some $z \in K$, then $z \in V(h_0)$ must hold. Suppose there exists a sequence of outcomes that realizes $y \ll z$ and let player i be the first player in the sequence. By Lemma 1 and by the definition of H, we have $u_i(y) = f(C, h_1) > f(D, h_0) = u_i(z)$, a contradiction. No outcome in K can indirectly dominate y. The external stability of K requires $y \in K$. Accordingly, we have $V(h_1) \subset K$.

Repeatedly applying the same argument, we can show that any outcome $z \in X$ such that $h_{t-1} < |C(z)| < h_t$ $(t = 1, 2, \ldots, T)$ or $h_T < |C(z)|$ can not be included in K and that $V(h_t) \subset K$ for all $h_t \in H$. Hence, $K = K^*$. □

4 Remarks

Let s^* and $\bar{s}$ be integers such that $f(C, s^* - 1) < f(D, 0) \leq f(C, s^*)$ and $f(D, \bar{s} - 1) \leq f(C, n) < f(D, \bar{s})$, respectively. Suzuki and Muto (2005) have shown that an outcome x other than x^D is *individually rational* if and only if $|C(x)| \geq s^*$ (of course, x^D itself is individually rational) and that an outcome x is *Pareto-efficient* if and only if $|C(x)| \geq \bar{s}$.

Under our Assumption 2, we have $s^* = h_1$. Since $s^* = h_1 < h_2 < \cdots < h_T$, we have $|C(x)| \geq s^*$ for any $x \in K^*$ other than x^D. Accordingly, any outcome in the purely noncooperative farsighted stable set is individually rational. In other words, the set K^* as a whole is individually rational.

If $h_T < \bar{s}$, then $f(D, h_T) < f(C, n)$ under Assumption 2; this, however, contradicts the definition of h_T. Therefore, we have $h_T \geq \bar{s}$. In turn, this implies that $|C(x)| \geq \bar{s}$ for any $x \in V(h_T) \subset K^*$. Hence, the purely noncooperative farsighted stable set includes Pareto-efficient outcomes. The "all-cooperation" outcome $x^C \equiv (C, C, \ldots, C)$, which is Pareto-efficient, can be included in K^* if and only if $h_T = n$.

References

[1] Arce MDG (1994 Stability criteria for social norms with applications to the prisoner's dilemma. *Journal of Conflict Resolution* 38 (4): 749–765.

[2] Chwe M S-Y (1994) Farsighted coalitional stability. *Journal of Economic Theory* 63: 299–325.

[3] D'Aspremont C, Jacquemin A, Gabszewicz J J, and Weymark J A (1983) On the stability of collusive price leadership. *Canadian Journal of Economics* 16 (1): 17–25.

[4] Diamantoudi E (2005) Stable cartels revisited. *Economic Theory* 26: 907–921.

[5] Fudenberg D and Maskin E (1986) The folk theorem in repeated games with discounting or with incomplete information. *Econometrica* 54: 533–556.

[6] Greenberg J (1990) *The Theory of Social Situations: An Alternative Game-Theoretic Approach*. Cambridge University Press.

[7] Harsanyi J (1974) An equilibrium-point interpretation of stable sets and a proposed alternative definition. *Management Science* 20: 1472–1495.

[8] Kalai E (1981) Preplay negotiations and the prisoner's dilemma. *Mathematical Social Sciences* 1: 375–376.

[9] Kamijo Y and Muto S (2006) Farsighted stability of collusive price leadership. (unpublished manuscript)

[10] Muto S (1993) Alternating-move preplays and vN-M stable sets in two person strategic form games. Discussion paper No. 9371, CentER for Economic Research.

[11] Muto S and Okada D (1996) Von Neumann-Morgenstern stable sets in a price-setting duopoly. *Keizai-to-Keizaigaku*— (*Economy and Economics*) 81 (Tokyo Metropolitan University): 1–14.

[12] Muto S and Okada D (1998) Von Neumann-Morgenstern stable sets in Cournot competition. *Keizai-to-Keizaigaku* (*Economy and Economics*) 85 (Tokyo Metropolitan University): 37–57.

[13] Nakanishi N (1999) Reexamination of the international export quota games through the theory of social situations. *Games and Economic Behavior* 27: 132–152.

[14] Nakanishi N (2001) On the Existence and Efficiency of the von Neumann-Morgenstern Stable Set in a n-Player Prisoners' Dilemma. *International Journal of Game Theory* 30 (2): 291–307.

[15] Nishihara K (1997) A resolution of N-person prisoners' dilemma. *Economic Theory* 10: 531–540.

[16] Okada A (1993) The possibility of cooperation in an n-person prisoners' dilemma with institutional arrangements. *Public Choice* 77: 629–656.

[17] Suzuki A and Muto S (2005) Farsighted Stability in an n-Person Prisoner's Dilemma. *International Journal of Game Theory* 33: 431–445.

[18] von Neumann J and Morgenstern O (1953) *Theory of Games and Economic Behavior* (3rd ed.). Princeton University Press.