Revealed Preference in Game Theory

Ádám Galambos[1]
MEDS Department
Kellogg School of Management
Northwestern University
a-galambos@kellogg.northwestern.edu

October 8, 2004

[1]I am grateful to Professor Marcel K. Richter for many inspiring and stimulating discussions on these topics, as well as many suggestions. I wish to thank Professors Beth Allen, Andrew McLennan and Jan Werner, and participants of the Micro/Finance and Micro/Game theory workshops at the University of Minnesota for their comments. This paper is based on my doctoral dissertation at the University of Minnesota. I gratefully acknowledge the financial support of the NSF through grant SES-0099206 (principal investigator: Professor Jan Werner).

Abstract

I characterize behavior generated by the pure strategy Nash equilibrium solution concept. My results are based on a revealed preference approach and are more general than those in the literature (1; 2). Dropping the unrealistic "complete domain" assumption, I characterize Nash equilibrium behavior. I show that determining whether observed choices were generated by the Nash equilibrium solution or not is an NP-complete problem. In contrast, the analogous problem for one decision maker can be determined in polynomial time.

JEL classifications: C72, D70.

Keywords: Nash equilibrium; Revealed preference; Testable implications; Complexity

1 Testability of Nash behavior

What does Nash equilibrium behavior look like? If we observed a group of agents play different games, could we tell, without knowing their preferences, whether they are playing according to the Nash equilibrium concept? Such questions could be of interest to a regulatory agency, wanting to know if some firms they observe in the market are behaving in a competitive or in a collusive way. A manager might ask the same question about her employees. A mechanism designer might want to test if a certain group of agents behave according to the Nash equilibrium solution, to see if he can realistically assume that the agents will behave that way when faced with his mechanism. But these questions are also of interest in themselves from a theoretical point of view. Our objective is to obtain behavioral characterizations of the Nash equilibrium solution.

The revealed preference literature asks questions of the form: "What conditions characterize choice behavior that is generated by maximization of a preference relation with certain properties?" Starting with (3) and (4), answers to such questions have been obtained in quite general settings, when there is an individual decision maker (5; 6; 7). The analogous revealed preference questions in the multi-person decision making problem were addressed only recently.[1] In many applications of game theory, such as analyzing oligopolistic markets or interaction between employees of a firm, a set of agents interact with each other repeatedly but under changing circumstances. That is, the same set of agents play different games. Due to a change in outside circumstances, their feasible strategies might vary from one occasion to the next, or the outcome function might change. When the Nash equilibrium solution is used to analyze such situations, it would be useful to know if, based on observations of past choices of the players, using that solution concept is realistic. In other words, what conditions on choice behavior correspond to the Nash equilibrium solution concept?

An intriguing application of revealed preference analysis in game theory involves comparing behavior generated by cooperative (e.g. core) and non-cooperative (e.g. Nash equilibrium) solution concepts. What are the observable behavioral differences between

[1]See (8) for a survey. (1) and (2) are most closely related to our work, because they formulate their results for normal form games, as we do. A complementary literature (9; 10) considers similar questions for extensive form games.

cooperative and non-cooperative behavior? When is it justifiable to assume that agents in a certain setting play cooperatively or non-cooperatively? While these questions are not addressed in this paper, we hope that the current results will lead to further research in that direction.

The direct link between the preferences and the choices of an individual decision maker is obscured in the multi-person case by the implicit aggregation of preferences. What information is revealed about *individual* preferences by the *collective* (Nash equilibrium) choices? The analysis to follow will hinge on finding the right answer to this question. Our approach differs from the approach used in much of the received literature on the behavioral characterization of Nash equilibrium in that it is based on a general notion of revelation and revealed preferences. This will be discussed in detail at a more suitable point below.

2 Nash equilibrium rationalizability

Suppose we observe a finite set I of players play different games. Their strategy spaces are subsets of their *universal strategy space* $\mathcal{S}_i$. Let Λ be a finite set of game forms, i.e. a set of Cartesian product subsets of[2] $\mathcal{S}$. We call an element $S = \prod_{i \in I} S_i$ of Λ a *game form*[3]. For each such game form S, we observe the strategy profiles played. We assume that if there are several strategy profiles which players would be willing to choose, then we observe all of these as chosen. Previous authors have assumed that *all* conceivable game forms are observed, i.e. that Λ contains all Cartesian product subsets of $\mathcal{S}$. This "complete domain" assumption is very extreme in the context of revealed preference theory, where one would like to assume as little as possible about the set of observations given, so as to incerase the applicability of the theory. In a companion paper, I show that Nash behavior can be characterized, using revealed preference conditions, under a substantially weakened version of the complete domain assumption. In this paper, I impose no restrictions on the observations, i.e. Λ is an *arbitrary* set of game forms. In addition, I generalize the notion of a choice correspondence.

[2]Let $\mathcal{S} := \prod_{i \in I} \mathcal{S}_i$.

[3]i.e. we assume, as commonly done in the literature, that the outcome space is S and the outcome function is the identity.

Formally, suppose we are given a (possibly empty-valued) choice correspondence $\mathfrak{C} : \Lambda \rightrightarrows \mathcal{S}$ and a (possibly empty-valued) *non-choice* correspondence $\hat{\mathfrak{C}} : \Lambda \rightrightarrows \mathcal{S}$, satisfying, for every $S \in \Lambda$,

$$\mathfrak{C}(S) \subseteq S, \qquad \hat{\mathfrak{C}}(S) \subseteq S, \qquad \mathfrak{C}(S) \cap \hat{\mathfrak{C}}(S) = \emptyset.$$

The strategy profiles in $\mathfrak{C}(S)$ are observed as chosen, while the strategy profiles in $\hat{\mathfrak{C}}(S)$ are observed as *not* chosen. A special case is the one implicit in most discussions: for all $S \in \Lambda$, $\hat{\mathfrak{C}}(S)$ is defined to be the complement of $\mathfrak{C}(S)$ in S, i.e., what is not observed as chosen, is assumed to be not chosen. In this case, the "rationalizability question" asks if observed choices can be rationalized as coincident with the set of pure strategy Nash equilibria. When observations are imperfect, it is more relevant to ask whether observed choices can be rationalized as a subset, or even a superset, of the set of pure strategy Nash equilibria. In the terminology of (11), the first notion might be called *Nash sub-semirationality*, and the second *Nash supra-semirationality*. Using a pair $(\mathfrak{C}, \hat{\mathfrak{C}})$ makes these notions special cases of our notion of rationalizability (in the sense of (1,2)): the first corresponds to assuming $\hat{\mathfrak{C}}(S) = \emptyset$, and the second to assuming $\mathfrak{C}(S) = \emptyset$.

For simplicity, we will pose the rationalizability question for the case of strict preferences. It is straightforward to modify the definitions and the Theorem below to characterize rationalizability by a weak preference relation (see the Remark below).

DEFINITION 1 $(\mathfrak{C}, \hat{\mathfrak{C}})$ is *(pure strategy Nash equilibrium) rationalizable* if there exist total, transitive and asymmetric binary relations $(\prec_i)_{i\in I}$ on $\mathcal{S}$ such that for all $S \in \Lambda$,

1. the chosen strategies $\mathfrak{C}(S)$ are Nash equilibria of $((S_i)_{i\in I}, (\prec_i)_{i\in I})$, i.e.

$$s^* \in \mathfrak{C}(S) \Longrightarrow \forall_{i\in I}\, \forall_{s_i\in S_i\setminus\{s_i^*\}}\ (s_i, s_{-i}^*) \prec_i s^*, \tag{1}$$

and

2. the non-chosen strategies $\hat{\mathfrak{C}}(S)$ are not Nash equilibria of $((S_i)_{i\in I}, (\prec_i)_{i\in I})$, i.e.

$$s^* \in \hat{\mathfrak{C}}(S) \Longrightarrow \exists_{i\in I}\, \exists_{s_i\in S_i\setminus\{s_i^*\}}\ s^* \prec_i (s_i, s_{-i}^*). \tag{2}$$

■

To characterize rationalizability, we define a notion of *direct revelation* ((5), following (3)). Partly this is based on observing that certain strategy profiles are chosen (the $\mathfrak{C}$-revelations), and partly it is based on observing that some other strategy profiles are not chosen (the $\hat{\mathfrak{C}}$-revelations).

Example 1 (Direct revelation) Suppose that player 1's strategy space is $\{U, D\}$, and player 2's strategy space is $\{L, M, R\}$. In the game form $S = \{U, D\} \times \{L, M, R\}$ we observe that the framed strategy profile is chosen ($\mathfrak{C}(S) = \{(U, R)\}$), while the underlined strategy profile is not chosen ($\hat{\mathfrak{C}}(S) = \{(D, M)\}$).

1 \ 2	L	M	R
U	$(\mathbf{U}, \mathbf{L})$	$(\mathbf{U}, \mathbf{M})$	$\boxed{(\mathbf{U}, \mathbf{R})}$
D	$(\mathbf{D}, \mathbf{L})$	$\underline{(\mathbf{D}, \mathbf{M})}$	$(\mathbf{D}, \mathbf{R})$

It is then directly revealed (via $\mathfrak{C}$) that (U, R) is preferred to anything else obtainable from it by unilateral deviation, i.e.

$$(U, L) \underset{2}{\lhd} (U, R), \tag{3}$$

$$(U, M) \underset{2}{\lhd} (U, R), \tag{4}$$

$$(D, R) \underset{1}{\lhd} (U, R). \tag{5}$$

It is also directly revealed (via $\hat{\mathfrak{C}}$) that one of the players prefers some strategy profile obtainable for her from (D, M) by unilateral deviation. We write this as

$$[(D, M) \underset{2}{\lhd} \{(D, L), (D, R)\}] \quad \textbf{or} \quad [(D, M) \underset{1}{\lhd} \{(U, M)\}]. \tag{6}$$

Note that each disjunct corresponds to a player, and that the "revealed preference relations" $\underset{2}{\lhd}$ and $\underset{1}{\lhd}$ are one–many relations. Rationalizability (1,2) is clearly equivalent to (3-6).

An equivalent formulation of the same rationalizability question is in terms of payoff functions rather than (revealed) preferences. Let u_1 and u_2 be the (unknown) real-valued payoff functions of players 1 and 2. To simplify notation, let $u_1(U, L)$ denote the payoff to player 1 of the strategy profile (U, L). Then the information revealed by $\mathfrak{C}$ ((3),(4), and (5) above) can be written as a system of inequalities:

$$u_2(U, L) \quad < \quad u_2(U, R), \tag{7}$$

$$u_2(U, M) \quad < \quad u_2(U, R), \tag{8}$$

$$u_1(D, R) \quad < \quad u_1(U, R). \tag{9}$$

To write the information revealed by $\hat{\mathfrak{C}}$ ((6) above) as an inequality, we translate each revealed preference statement into an inequality involving maxima:

$$[u_2(D, M) \quad < \quad u_2(D, L) \vee u_2(D, R)] \quad \textbf{or} \quad [u_1(D, M) \quad < \quad u_1(U, M)]. \tag{10}$$

Rationalizability (1,2) is equivalent to the solvability of the set of inequalities from $\mathfrak{C}$ and $\hat{\mathfrak{C}}$:

$$u_2(U, L) \quad < \quad u_2(U, R), \tag{11}$$

$$u_2(U, M) \quad < \quad u_2(U, R), \tag{12}$$

$$u_1(D, R) \quad < \quad u_1(U, R), \tag{13}$$

$$[u_2(D, M) < u_2(D, L) \vee u_2(D, R)] \quad \textbf{or} \quad [u_1(D, M) < u_1(U, M)]. \tag{14}$$

It will be more convenient to write this set of statements in disjunctive normal form:

$$\boxed{\begin{array}{c} u_2(U, L) < u_2(U, R) \\ u_2(U, M) < u_2(U, R) \\ u_1(D, R) < u_1(U, R) \\ u_2(D, M) < u_2(D, L) \vee u_2(D, R) \end{array}} \quad \textbf{or} \quad \boxed{\begin{array}{c} u_2(U, L) < u_2(U, R) \\ u_2(U, M) < u_2(U, R) \\ u_1(D, R) < u_1(U, R) \\ u_1(D, M) < u_1(U, M) \end{array}}$$

This reduces rationalizability to solving systems of inequalities involving suprema. ■

Now we turn to the general definition of direct revelation. Let $S_i \diamond s'$ denote the set of strategy profiles which player i can obtain in S by unilaterally deviating from s'.

Definition 2

1. **$\mathfrak{C}$-revelations:** a chosen strategy profile is directly revealed preferred by *each* player to *every* strategy profile obtainable from it by her unilateral deviation. Formally,

s'' is *directly revealed preferred to s' by player i* if there exists a game form $S \in \Lambda$ such that $s'' \in \mathfrak{C}(S)$ and $s' \in S$, and strategy profiles s', s'' differ only in player i's strategy. We write this as

$$s' \underset{i}{\lhd} s''. \tag{NE-15}$$

2. **$\hat{\mathfrak{C}}$-revelations:** if a strategy profile is not chosen, it is directly revealed that *some* player prefers *some* strategy profile obtainable from it by her unilateral deviation. Formally,

   if $S \in \Lambda$ and $s' \in \hat{\mathfrak{C}}(S)$, then it is *directly revealed that some player $i \in I$ prefers an element of $S_i \diamond s'$ to s'*. We write this as[4]

$$\bigvee_{i \in I} \; [s' \underset{i}{\lhd} S_i \diamond s']. \tag{nNE-16}$$

■

The $\mathfrak{C}$-revelations are much like revealed preference relations in consumer theory, and the $\hat{\mathfrak{C}}$-revelations are disjunctions, with each disjunct corresponding to a player. Rationalizability (1,2) is clearly equivalent to all $\mathfrak{C}$- and $\hat{\mathfrak{C}}$-revelations holding. It will be more convenient to treat these statements in a logically equivalent form as a disjunction of conjunctions[5] – which we call the *canonical form*.

Example 1 (Continued) Suppose we observe the game form in the previous Example, with (U, R) chosen, and (D, M) not chosen, but now we also observe that (U, L) is not chosen:

$1 \diagdown 2$	L	M	R
U	$\underline{(\mathbf{U}, \mathbf{L})}$	$(\mathbf{U}, \mathbf{M})$	$\boxed{(\mathbf{U}, \mathbf{R})}$
D	$(\mathbf{D}, \mathbf{L})$	$\underline{(\mathbf{D}, \mathbf{M})}$	$(\mathbf{D}, \mathbf{R})$

[5] I.e. in disjunctive normal form.

We have the following revelations, with $S_2 \diamond (D,M)$ denoting $\{(D,L),(D,R)\}$, and $S_1 \diamond (D,M)$ denoting $\{(U,M)\}$, etc.:

$$(U,L) \underset{2}{\lhd} (U,R) \tag{17}$$

$$(U,M) \underset{2}{\lhd} (U,R) \tag{18}$$

$$(D,R) \underset{1}{\lhd} (U,R) \tag{19}$$

$$[(D,M) \underset{2}{\lhd} S_2 \diamond (D,M)] \quad \vee \quad [(D,M) \underset{1}{\lhd} S_1 \diamond (D,M)] \tag{20}$$

$$[(U,L) \underset{2}{\lhd} S_2 \diamond (U,L)] \quad \vee \quad [(U,L) \underset{1}{\lhd} S_1 \diamond (U,L)]. \tag{21}$$

The canonical form is

$$\boxed{\begin{array}{c} (U,L) \underset{2}{\lhd} (U,R) \\ (U,M) \underset{2}{\lhd} (U,R) \\ (D,R) \underset{1}{\lhd} (U,R) \\ (D,M) \underset{2}{\lhd} S_2 \diamond (D,M) \\ (U,L) \underset{2}{\lhd} S_2 \diamond (U,L) \end{array}} \bigvee \boxed{\begin{array}{c} (U,L) \underset{2}{\lhd} (U,R) \\ (U,M) \underset{2}{\lhd} (U,R) \\ (D,R) \underset{1}{\lhd} (U,R) \\ (D,M) \underset{2}{\lhd} S_2 \diamond (D,M) \\ (U,L) \underset{1}{\lhd} S_1 \diamond (U,L) \end{array}} \bigvee \boxed{\begin{array}{c} (U,L) \underset{2}{\lhd} (U,R) \\ (U,M) \underset{2}{\lhd} (U,R) \\ (D,R) \underset{1}{\lhd} (U,R) \\ (D,M) \underset{1}{\lhd} S_1 \diamond (D,M) \\ (U,L) \underset{1}{\lhd} S_1 \diamond (U,L) \end{array}} \bigvee \boxed{\begin{array}{c} (U,L) \underset{2}{\lhd} (U,R) \\ (U,M) \underset{2}{\lhd} (U,R) \\ (D,R) \underset{1}{\lhd} (U,R) \\ (D,M) \underset{1}{\lhd} S_1 \diamond (D,M) \\ (U,L) \underset{2}{\lhd} S_2 \diamond (U,L) \end{array}}$$

The top three lines are the same for each box and correspond to the $\mathfrak{C}$-revelations. The bottom two lines come from the $\hat{\mathfrak{C}}$-revelations and assign a deviating player to each strategy profile that is not chosen. Thus the bottom lines differ across boxes – the first box assigns player 2 to both (D,M) and (U,L), the second box assigns player 2 to (D,M) but player 1 (U,L), etc. Because the canonical form always has this general structure, each of its disjuncts (i.e. the boxes in the example) will be called an *assignment*. ■

If all the assignments in the canonical form were self-contradictory in some sense, there would exist no rationalization. What is the right notion of "self-contradictory" in this setting? To answer this question, we will consider the very simple special case of just one player, i.e. a single decision maker.

Example 2 (One player) Suppose there is only one player (say player 1), i.e. we observe a single decision maker choosing from subsets of a finite set $\mathcal{S}$. The set of these choice sets ("budgets") is denoted by Λ. Suppose further that we observe no alternative as chosen, i.e. for all $S \in \Lambda$ we have $\mathfrak{C}(S) = \emptyset$. Of course, we still observe via $\hat{\mathfrak{C}}$ that some alternatives are not chosen. (This setup can be interpreted as a revealed preference problem with a notion of rationality that is weaker than preference maximization. In particular, we ask if the decision maker is rational in the sense that her observed chosen set for S *contains* the preference-maximal elements of S. This question arises naturally when observations are imperfect, and the *actual* chosen set is known to be only a subset of the *observed* chosen set. This notion was labeled *supra-semirationality* in (11).)

Since we have no observations of any alternative being chosen, the only direct revelations we have are those which state that an alternative not chosen is revealed worse than some other available alternative. Formally, our "data set" looks like

$$\begin{aligned} s_0^1 &\in \hat{\mathfrak{C}}(\{s_0^1, s_1^1, \ldots, s_{k_1}^1\}) && (22)\\ s_0^2 &\in \hat{\mathfrak{C}}(\{s_0^2, s_1^2, \ldots, s_{k_2}^2\}) && (23)\\ &\vdots && (24)\\ s_0^m &\in \hat{\mathfrak{C}}(\{s_0^m, s_1^m, \ldots, s_{k_m}^m\}), && (25) \end{aligned}$$

where $s_r^q \in \mathcal{S}$ for $q = 1, \ldots, m$ and $r = 0, \ldots, k_q$.

Thus the set of direct revelations is

$$\begin{aligned} s_0^1 &\underset{1}{\lhd} \{s_1^1, \ldots, s_{k_1}^1\} && (26)\\ s_0^2 &\underset{1}{\lhd} \{s_1^2, \ldots, s_{k_2}^2\} && (27)\\ &\vdots \\ s_0^m &\underset{1}{\lhd} \{s_1^m, \ldots, s_{k_m}^m\}. && (28) \end{aligned}$$

The interpretation is that the decision maker strictly prefers some element of $\{s_1^1, \ldots, s_{k_1}^1\}$ to s_0^1, and strictly prefers some element of $\{s_1^2, \ldots, s_{k_2}^2\}$ to s_0^2, etc. It is clear that rationalizability (1,2) in this setting is equivalent to (26 – 28).

If every alternative appearing in one of the sets on the right hand side above also appeared on the left hand side on some other line, it is clear that there would exist no rationalization. If there were one, the alternatives that appear in (26),(27),(28) would contain a preference cycle: s_0^1 would be worse than some alternative on the right in (26), which in turn would appear on the left on another line and so would be worse than another alternative, which in turn would appear on the left on another line, and so would be worse than .... With a finite number of alternatives, this would result in a preference cycle. For this reason, if a set of revelations has the property that all alternatives appearing on the right also appear on the left, we say that it is an *implicit cycle*. Thus it is a necessary condition for rationalizability that the set of revelations contain no implicit cycle. In separate work on supra-semirationalizability we show that this condition is also sufficient for rationalizability. This characterizes rationalizability for the setting of this example.

Remark 1 In the equivalent formulation using a payoff function u, the relations (26),(27), and (28) make a system of inequalities:

$$u(s_0^1) < u(s_1^1) \vee \cdots \vee u(s_{k_1}^1) \tag{29}$$

$$u(s_0^2) < u(s_1^2) \vee \cdots \vee u(s_{k_2}^2) \tag{30}$$

$$\vdots$$

$$u(s_0^m) < u(s_1^m) \vee \cdots \vee u(s_{k_m}^m), \tag{31}$$

where "$\vee$" denotes "supremum." Thus rationalizability (1,2) is equivalent to the solvability of (29 – 31). If all alternatives that appear on the right in (29 – 31) also appear on the left, i.e.

$$\bigcup_{j=1}^{m} \{s_1^j, \ldots, s_{k_j}^j\} \subseteq \{s_0^1, s_0^2, \ldots, s_0^m\}, \tag{32}$$

it follows that

$$\bigvee_{j=1}^{m} u(s_1^j) \vee \cdots \vee u(s_{k_j}^j) < u(s_0^1) \vee u(s_0^2) \vee \cdots \vee u(s_0^m). \tag{33}$$

On the other hand, from (29),(30), and (31) it follows that

$$u(s_0^1) \vee u(s_0^2) \vee \cdots \vee u(s_0^m) < \bigvee_{j=1}^{m} u(s_1^j) \vee \cdots \vee u(s_{k_j}^j), \tag{34}$$

which is a contradiction, proving that implicit cycles imply non-rationalizability.

■

The insights of the one player example help us understand the general Nash rationalizability problem. The notion of "self-contradictory" appropriate for this setting is:

<u>DEFINITION 3</u> Let $i \in I$ and $S^1, \ldots, S^m \in \Lambda$, and $s^j \in S^j$ for $j = 1, \ldots, m$. A set of statements of the form

$$s^1 \underset{i}{\lhd} S^1_i \diamond s^1 \tag{35}$$

$$s^2 \underset{i}{\lhd} S^2_i \diamond s^2 \tag{36}$$

$$\vdots$$

$$s^m \underset{i}{\lhd} S^m_i \diamond s^m \tag{37}$$

is an implicit cycle if all strategy profiles contained in the sets on the right hand side above also appear on the left hand side, i.e.

$$\bigcup_{j=1}^{m} s^j \diamond S^j_i \quad \subseteq \quad \{s^1, s^2, \ldots, s^m\}. \tag{38}$$

■

<u>MAIN THEOREM</u> A pair $(\mathfrak{C}, \hat{\mathfrak{C}})$ is not (pure strategy Nash equilibrium) rationalizable if, and only if, all assignments in the canonical form contain an implicit cycle.

<u>**Remark 2**</u> The characterization of (pure-strategy Nash equilibrium) rationalizability allowing weak preferences (total, transitive, reflexive binary relations) can be obtained by a slight modification of the strict preference case. The definition of $\mathfrak{C}$-revelation must reflect that weak preference is allowed, so it is denoted by new symbols $\underset{i}{\unlhd}$. The definition of an implicit cycle remains the same, except we must add that the set of relations forming an implicit cycle must be minimal with respect to the defining property of an implicit cycle and must contain at least one $\underset{i}{\lhd}$.

<u>**Proof**</u>

Necessity: If $(\prec_i)_{i \in I}$ Nash-rationalizes $(\mathfrak{C}, \hat{\mathfrak{C}})$, then (NE-15) and (nNE-16) must hold for $(\prec_i)_{i \in I}$, i.e. all direct revelations are simultaneously satisfied by $(\prec_i)_{i \in I}$. The same set of statements will still be satisfied when written in the logically equivalent disjunctive normal form, so one of the assignments will be satisfied, and thus will have no implicit cycles.

Sufficiency: Suppose that there exists an assignment in the canonical form that has no implicit cycles. By the characterization of supra-semirationality, there exist preference relations $(\prec_i)_{i \in I}$ that satisfy all revealed preference statements in that assignment. These preferences then satisfy all direct revelations written in canonical form, and so satisfy all direct revelations as in (1,2), rationalizing $(\mathfrak{C}, \hat{\mathfrak{C}})$. *Q.E.D.*

3 Complexity

In this section we compare the computational complexity of two *problems*: Nash rationalizability (NR), i.e. determining whether the behavior of finitely many players is consistent with Nash equilibrium, and supra-semirationalizability (SSR), i.e. determining whether the observed choices of a single decision maker are supra-semirationalizable.

3.1 Nash rationalizability

Our characterization of Nash rationalizability involved rewriting a statement in conjunctive normal form (CNF) as one in disjunctive normal form (DNF). Algorithmically, this is a very complex operation: it most likely cannot be done in polynomial time.[6] Is it possible that there is some other way to characterize Nash rationalizability that is algorithmically not so complex? The following result answers this question in the negative.

<u>Theorem 2</u> The (pure strategy Nash equilibrum) rationalizability problem is NP-complete.

In fact, we prove a stronger statement: The (pure strategy Nash equilibrum) rationalizability problem is NP-complete *even if* we assume that for every game form $S \in \Lambda$ we have $\hat{\mathfrak{C}}(S) = S \backslash \mathfrak{C}(S)$, and that players have at most 2 strategies. The proof (in Appendix A) is based on a standard technique in the theory of computational complexity: "polynomially reducing" a problem that is known to be NP-complete to the given problem.[7]

[6] Showing that it can be done in polynomial time would amount to a proof that $P = NP$ (see (12) p. 72, exercise 2.9). Whether $P = NP$ is one of the major unsolved problems of mathematics, and it is widely conjectured that $P \neq NP$ (see (13)).

[7] Specifically, we use 3SAT, a version of the satisfiability problem that was shown to be NP-complete in (14).

3.2 Supra-semirationalizability

In contrast to the above result, the supra-semirationalizability (**SSR**) problem (involving one decision maker) is polynomial. Let $\mathcal{S}$ denote the set of alternatives, and suppose we observe for each S^i, $i = 1, \dots k$ that $\mathfrak{C}(S^i)$ is chosen. Recall from the single player example in section 2 that this means that the *actual* chosen set is a subset of $\mathfrak{C}(S^i)$. Define an instance of SSR as a set of pairs of sets

$$\{(\mathfrak{C}(S^1), S^1\backslash\mathfrak{C}(S^1)), (\mathfrak{C}(S^2), S^2\backslash\mathfrak{C}(S^2)), \dots, (\mathfrak{C}(S^k), S^k\backslash\mathfrak{C}(S^k))\}. \tag{39}$$

Such a list is a *yes*-instance if there exists a preference relation on $\mathcal{S}$ such that for each S^i, the set $\mathfrak{C}(S^i)$ contains the preference-maximal elements. Otherwise it is a *no*-instance.

<u>THEOREM 3</u> The supra-semirationalizability problem can be decided in polynomial time.

<u>Proof</u>

The following algorithm determines in polynomial time whether an instance of SSR is a *yes*-instance or a *no*-instance. By "polynomial time" we mean, intuitively, that the number of steps in the algorithm is polynomial in the length of the input string ((39) above, with the finite sets $\mathfrak{C}(S^i)$ and $S^i\backslash\mathfrak{C}(S^i)$ written out element by element).

Algorithm:

1. Let $I := \{1, \dots, k\}$. Let $Q = \emptyset$.

2. Let $q = 1$ and $r = 1$.

3. Scan the sets $S^i\backslash\mathfrak{C}(S^i)$, $i \in I$, to check if the qth element of $\mathfrak{C}(S^r)$ (denote it by x^*) appears in any of them.

   3.1. If it does:

   3.1.1. If $\mathfrak{C}(S^r)$ has q elements and r is the highest index in I, STOP.

   3.1.2. If $\mathfrak{C}(S^r)$ has q elements, let $q = 1$ and increase r by 1. Go to step 3.

   3.1.3. Increase q by 1 and go to step 3.

   3.2. If x^* does not appear in any $S^i\backslash\mathfrak{C}(S^i)$, add it to Q and set $x \prec x^*$ for all $x \in \mathcal{S}\backslash Q$.

3.3. Scan the sets $\mathfrak{C}(S^i)$, $i \in I$ to check if x^* appears in any of them, and let $I' := \{i \in I : x^* \notin \mathfrak{C}(S^i)\}$ (note that by the definition of x^* in step 3, $I' \subsetneqq I$). If $I' = \emptyset$, set $I = \emptyset$ and STOP. (Note that the observations with labels in $I \backslash I'$ are now supra-semirationalized by $\prec$.)

3.4. Relabel the pairs $(\mathfrak{C}(S^i), S^i \backslash \mathfrak{C}(S^i))$, $i \in I'$, with the labels $1, \ldots, |I'|$. Let $I := \{1, \ldots, |I'|\}$. Go to step 2.

At every iteration the algorithm either returns to step 2 or 3 or it stops. The algorithm stops after at most $\sum_{i=1}^{k} |\mathfrak{C}(S_i)|$ iterations. If $I \neq \emptyset$ when the algorithm stops, the input choice correspondence is a *no-instance*. In this case $\{(\mathfrak{C}(S^i), S^i \backslash \mathfrak{C}(S^i))|\ i \in I\}$ is an implicit cycle, and by section 2 the input choice correspondence is not rationalizable. If $I = \emptyset$ when the algorithm stops, the input choice correspondence is a *yes-instance*. In this case $\prec$ is a partial order on $\mathcal{S}$ that supra-semirationalizes the input choice correspondence.[8] It is clear that each step is polynomial in the length of the input, and so is the number of iterations. *Q.E.D.*

4 Conclusion

We characterized behavior generated by the pure strategy Nash equilibrium concept for normal form games, first for "closed domains," and then without placing any restrictions on observations. In particular, we did not impose the unrealistic "complete domain" assumption, which was central to results in the received literature. We obtained our characterizations using a revealed preference approach. This suggests that further extensions to solution concepts other than Nash eqilibrium might be possible with revealed preference techniques.

We showed that, from a computational perspective, Nash rationalizability in its full generality is a very complex problem. This can be a significant issue for applications. Note, however, that our result means only that the running time of a "Nash rationalizability algorithm" is (most likely[9]) not polynomially bounded as the number of players increases.

[8]This can be seen by noticing that indicies are ommitted from I in step 3.3 only if the corresponding observation is rationalized by the partial order defined so far.

[9]See footnote 6 on p. 11.

Our polynomial-time algorithm for the one player case in section 3.2 suggests that one might hope to find efficient algorithms for determining Nash rationalizability with a fixed number of players.

Another interesting question for future research is the role of beliefs in multi-agent decision making. Since the literature so far has addressed only the Nash equilibrium solution concept, the role of beliefs has been hidden by the implicit assumption that agents' beliefs correspond exactly to the actions taken. If one were to study behavior generated by other solution concepts that are not "Nash-like," such as Pearce-Bernheim rationalizability, the prominent role of beliefs would become apparent.

Yet another interesting aspect of this problem is the relationship between the analyst or observer and the decision making process. In rationalizability for individual choice problems, it seems clear that the observer and the decision making process are entirely separate. That is, the analyst is outside the decision making problem, observing the behavior of the decision maker. In collective decision making situations, it is conceivable that the analyst is himself one of the decision makers. For example, a player in a game, not knowing the preferences of the other players, might attempt to draw conclusions concerning the plausibility of certain possible outcomes, based on some previous experiences of games played by the same agents. Analyzing situations of this kind might lead to interesting applications.

Appendix

Here we prove Theorem 2 of section 3.1, which states that the Nash rationalizability problem (NR) is NP-complete. Our proof involves two additional problems: Nash rationalizability with each player having at most two strategies in each observed game form (NR2), and the classic problem of determining the satisfiability of a Boolean formula in conjunctive normal form with three disjuncts in each conjunct (3SAT).

Proof [Theorem 2]

We will prove the theorem using *polynomial-time reduction*, a standard technique in the theory of computational complexity. We will show that the 3SAT problem, known to be NP-complete (see (14) and (13)), polynomially transforms into the Nash rationalizability problem for two-strategy games (henceforth denoted by NR2), which is a special case of the Nash rationalizability problem (henceforth denoted by NR). That is, we will construct an algorithm that runs in polynomial time, and, given any instance of 3SAT, produces an instance of NR2 with the property that the NR2 instance is rationalizable if and only if the 3SAT instance is satisfiable. This will imply that if there exists a polynomial-time algorithm for deciding NR2, then any instance of 3SAT can be decided in polynomial time by first polynomially transforming it into an instance of NR2 and then deciding that in polynomial time. Since 3SAT is NP-complete, this argument will establish that NR2 is NP-complete.

NR2: The Nash rationalizability problem for two-strategy games can be described as follows. Let

$$\mathbf{P} = \{1, 2, 3, \dots\} \cup \{1', 2', 3', \dots\} \cup \{1'', 2'', 3'', \dots\} \tag{40}$$

be the set of potential players. In all subforms below, the number of players with non-trivial strategy sets will be at most 3. The strategy set of player $i \in \mathbf{P}$ is $S_i : \{a_i, b_i, c_i\}$, and for any finite subset $I \subset \mathbf{P}$, we write $S_I := \prod_{i \in I} S_i$. An *instance of* NR2 consists of a finite set I of players, and a choice function on a set of subforms of S_I. A subform will be described by the strategy sets of the players involved. A typical instance of NR2 is

$$[(s_{p_1^1}, s_{p_2^1}, s_{p_3^1}), o_1],\ [(s_{p_1^2}, s_{p_2^2}, s_{p_3^2}), o_2],\ \dots,\ [(s_{p_1^k}, s_{p_2^k}, s_{p_3^k}), o_k] \tag{41}$$

with $s_{p_j^i} \subseteq S_{p_j^i}$ and $o_i \in s_{p_1^i} \times s_{p_2^i} \times s_{p_3^i}$ for all $i \in \{1, \dots, k\}$. The player set

$$I = \bigcup_{i=1}^{k} \{p_1^i, p_2^i, p_3^i\} \tag{42}$$

is implicit in the description of an instance of NR2 (41). The players in I who are not involved in a subform are assumed to play their a strategy. The intended interpretation is that in the subform $s_{p_1^i} \times s_{p_2^i} \times s_{p_3^i}$ the strategy profile o_i is the strategic outcome observed. We assume that for every subform $s_{p_1^i} \times s_{p_2^i} \times s_{p_3^i}$ the *only* choice is o_i.[10] An instance of NR2 is a *yes-instance* if the corresponding choice function is (pure strategy Nash equilibrium) rationalizable, and it is a *no-instance* if it is not. A polynomial-time algorithm for NR2 is a polynomial-time algorithm that returns, for any given instance of NR2, a *yes* if and only

[10] In the notation of section 2, we assume that for every subform S we have $\hat{\mathfrak{C}}(S) = S \backslash \mathfrak{C}(S)$.

if it is a yes-instance. Below we will show that if there exists a polynomial-time algorithm for NR2, then there exists a polynomial-time algorithm for 3SAT, which proves that NR2 is NP-complete.[11]

3SAT: Suppose that $X = \{x_1, x_2, \ldots, x_m\}$ is a set of Boolean variables and $\bar{X} = \{\bar{x}_1, \ldots, \bar{x}_m\}$ is the set of their negations. For any truth assignment $T : X \to \{\mathfrak{t}, \mathfrak{f}\}$, we define for $\bar{x} \in \bar{X}$ the extension of T by $T(\bar{x}) = \mathfrak{t}$ if, and only if $T(x) = \mathfrak{f}$. The set $X^* := X \cup \bar{X}$ is the set of *literals*. A subset C of X^* is a *clause*. Suppose a set $\{C_1, \ldots, C_k\}$ of clauses is given. A truth assignment $T : X \to \{\mathfrak{t}, \mathfrak{f}\}$ *satisfies* $\{C_1, \ldots, C_k\}$ if for every clause C_i there exists $x \in C_i$ with $T(x) = \mathfrak{t}$. A set of clauses is *satisfiable* if there exists a truth assignment that satisfies it. We can now state 3SAT: *Given an arbitrary finite set of clauses with exactly three elements in every clause, does there exist a satisfying truth assignment?* 3SAT is known to be NP-complete (see (13)).

3SAT → NR2: We now define the polynomial-time transformation mentioned at the beginning of the proof. That is, we define a polynomial-time algorithm that takes any instance of 3SAT as its input, and produces an instance of NR2 that is rationalizable if and only if the input 3SAT instance is satisfiable. Suppose we are given an arbitrary instance of 3SAT:

$$V = \left\{\{v_1^1, v_1^2, v_1^3\}, \{v_2^1, v_2^2, v_2^3\}, \cdots, \{v_l^1, v_l^2, v_l^3\}\right\}, \tag{43}$$

where $v_j^i \in X^*$. Suppose w.l.o.g. that the set of variables that appear in V is $\{x_1, \ldots, x_k\}$. We will construct an instance of NR2 with player set

$$\{1, 1', 1'', 2, 2', 2'' \ldots, k, k', k''\}. \tag{44}$$

First we construct a set of games for every variable that is negated in V. That is, suppose $\{v_j^1, v_j^2, \bar{x}_h\} \in V$. Then Γ_h consists of the following subform–outcome pairs (see figure 1):

1. $[(\{a_h\}, \{a_{h'}, b_{h'}\}, \{a_{h''}, b_{h''}\},), a_h a_{h'} b_{h''}]$ (see fig. 2)

2. $[(\{a_h, b_h\}, \{a_{h'}, b_{h'}\}, \{b_{h''}\},), a_h a_{h'} b_{h''}]$ (see fig. 3)

3. $[(\{b_h\}, \{a_{h'}, b_{h'}\}, \{a_{h''}, b_{h''}\},), b_h b_{h'} a_{h''}]$ (see fig. 4)

[11] It is clear that NR2 is in the class NP: given an instance of NR2 and preference relations for every player, it can be checked in polynomial time whether the preferences Nash rationalize the given choice function.

4. $[(\{a_h, b_h\}, \{a_{h'}, b_{h'}\}, \{a_{h''}\},), b_h b_{h'} a_{h''}]$ (see fig. 5)

5. $[(\{a_h, b_h\}, \{a_{h'}\}, \{a_{h''}, b_{h''}\},), a_h a_{h'} b_{h''}]$ (see fig. 6)

6. $[(\{a_h, b_h\}, \{b_{h'}\}, \{a_{h''}, b_{h''}\},), b_h b_{h'} a_{h''}]$ (see fig. 2)

Now we transform the problem V into an instance of NR2 as follows.

1. In the following steps, ignore every clause that includes a variable x and its negation $\bar{x}$ (these clauses are satisfied by any truth assignment).

2. Replace every clause of the form $\{x_e, x_f, x_g\}$ with

$$[(\{a_e, b_e\}, \{a_f, b_f\}, \{a_g, b_g\}), b_e b_f b_g]. \tag{45}$$

3. Replace every clause of the form $\{x_e, x_f, \bar{x}_g\}$ with

$$[(\{a_e, b_e\}, \{a_f, b_f\}, \{a_{g'}, b_{g'}\}), b_e b_f b_{g'}] \tag{46}$$

   and Γ_g.

4. Replace every clause of the form $\{x_e, \bar{x}_f, \bar{x}_g\}$ with

$$[(\{a_e, b_e\}, \{a_{f'}, b_{f'}\}, \{a_{g'}, b_{g'}\}), b_e b_{f'} b_{g'}] \tag{47}$$

   and Γ_g and Γ_f.

5. Replace every clause of the form $\{\bar{x}_e, \bar{x}_f, \bar{x}_g\}$ with

$$[(\{a_{e'}, b_{e'}\}, \{a_{f'}, b_{f'}\}, \{a_{g'}, b_{g'}\}), b_{e'} b_{f'} b_{g'}] \tag{48}$$

   and Γ_g, Γ_f and Γ_e.

The resulting instance of NR2 will be denoted by NR_V.

In the worst case, all variables that appear in V are distinct and are negated, which gives $l \cdot 19$ subform–outcome pairs, i.e. the input size is increased by a multiplicative factor. The transformation involves only replacing each clause by at most 19 subform–outcome pairs, as described above, and so it runs in polynomial time (in fact in linear time).

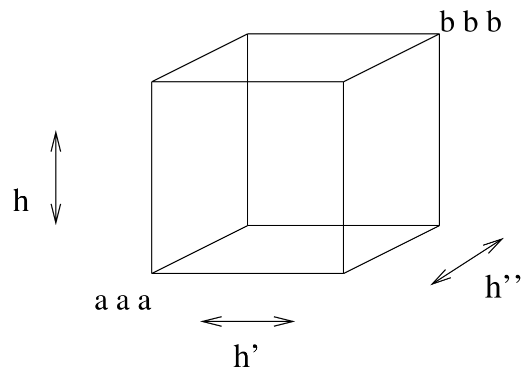

Figure 1: The subforms in Γ_h involve players h, h', h''

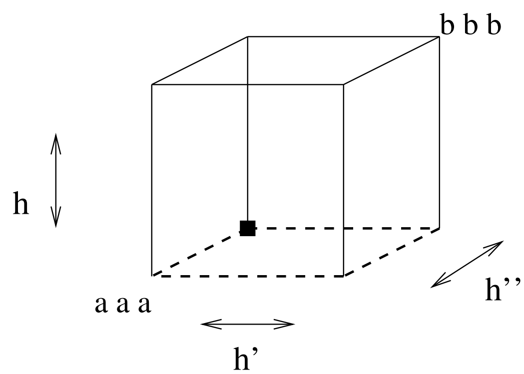

Figure 2: The first subform in Γ_h

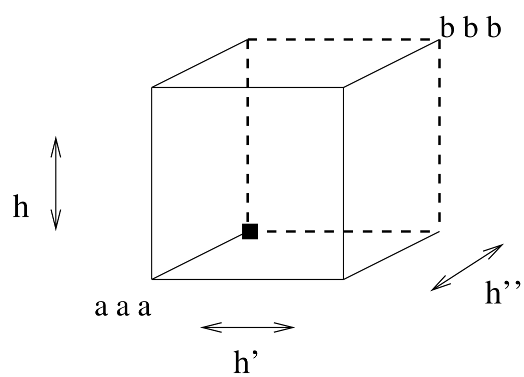

Figure 3: The second subform in Γ_h

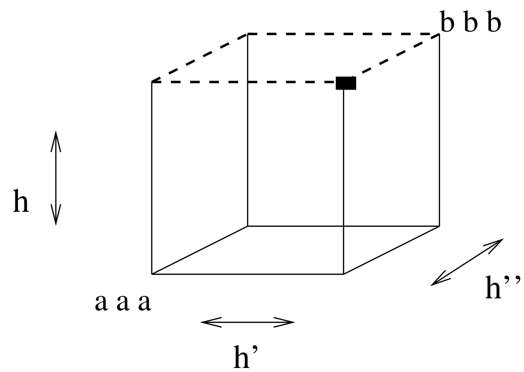

Figure 4: The third subform in Γ_h

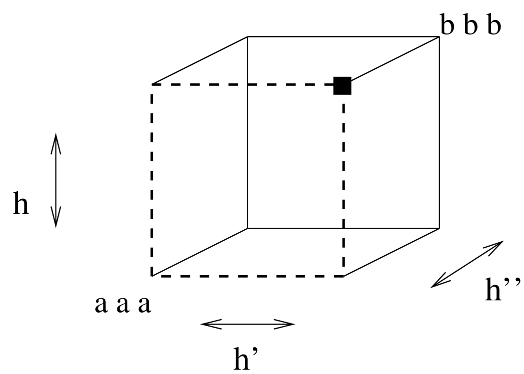

Figure 5: The fourth subform in Γ_h

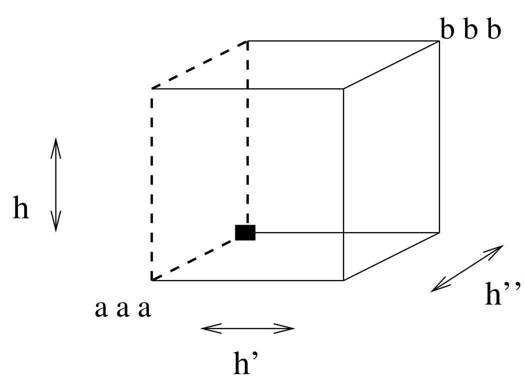

Figure 6: The fifth subform in Γ_h

Now we must show that the polynomial transformation $V \mapsto NR_V$ constructed above has the property mentioned at the beginning of the proof: V is satisfiable if and only if NR_V is Nash rationalizable. First, suppose NR_V is Nash rationalizable. Denote the rationalizing preference relations by $(\prec_i)_{i\in\{1,1',1'',\ldots,k,k',k''\}}$. Define, for each variable x_i with $i \in \{1,2,\ldots,k\}$ (recall that these are exactly the variables that appear in V) a truth assignment:

$$T_{\prec}(x_i) = \mathfrak{t} \iff a_1 a_{1'} a_{1''} \cdots a_{k''} \prec_i a_{-i} b_i. \tag{49}$$

Consider a clause of the form $\{x_e, x_f, x_g\}$. Since NR_V contains the pair

$$[(\{a_e, b_e\}, \{a_f, b_f\}, \{a_g, b_g\}), b_e b_f b_g], \tag{50}$$

and since $a_e a_f a_g$ is not a Nash equilibrium in this subform, it must be that

$$[a_e a_f a_g \prec_e b_e a_f a_g] \vee [a_e a_f a_g \prec_f a_e b_f a_g] \vee [a_e a_f a_g \prec_g a_e a_f b_g]. \tag{51}$$

Under $T_{\prec}$ this means that $\{x_e, x_f, x_g\}$ is satisfied.

Now consider a clause of the form $\{x_e, x_f, \bar{x}_g\}$. It is easy to see that if $(\prec_i)_{i\in\{1,1',1'',\ldots,k,k',k''\}}$ rationalize NR_V, then it follows from the construction of Γ_g that either $a_g a_{g'} a_{g''} \prec_g b_g a_{g'} a_{g''}$ or $a_g a_{g'} a_{g''} \prec_{g'} a_g b_{g'} a_{g''}$, *but not both.*[12] If $\neg[a_g a_{g'} a_{g''} \prec_g b_g a_{g'} a_{g''}]$, then by definition $T_{\prec}(x_g) = \mathfrak{f}$, so $\{x_e, x_f, \bar{x}_g\}$ is satisfied. If, on the other hand, $\neg[a_g a_{g'} a_{g''} \prec_{g'} a_g b_{g'} a_{g''}]$, then since $[(\{a_e, b_e\}, \{a_f, b_f\}, \{a_{g'}, b_{g'}\}), b_e b_f b_{g'}]$ is in NR_V, the fact that $a_e a_f a_{g'}$ is not a Nash equilibrium implies that either $a_e a_f a_{g'} \prec_e b_e a_f a_{g'}$ or $a_e a_f a_{g'} \prec_f a_e b_f a_{g'}$, which implies that either $T_{\prec}(x_e) = \mathfrak{t}$ or $T_{\prec}(x_f) = \mathfrak{t}$, and so $\{x_e, x_f, \bar{x}_g\}$ is satisfied.

The situation for clauses of the type $\{x_e, \bar{x}_f, \bar{x}_g\}$ and $\{\bar{x}_e, \bar{x}_f, \bar{x}_g\}$ is analogous, and these clauses will also be satisfied by $T_{\prec}$. Thus the truth assignment $T_{\prec}$ satisfies V.

To prove the converse, suppose that V is satisfied by a truth assignment T. For all $i \in \{1,2,\ldots,k\}$, let

$$a_1 a_{1'} a_{1''} \cdots a_{k''} \prec_i a_{-i} b_i \iff T(x_i) = \mathfrak{t}. \tag{52}$$

By construction (see footnote 12, p.20), the subform–outcome pairs in Γ_i force

$$a_1 a_{1'} a_{1''} \cdots a_{k''} \prec_{i'} a_{-i'} b_{i'} \iff T(x_i) = \mathfrak{f}. \tag{53}$$

[12] In fact, Γ_g is constructed so that it is rationalizable if and only if the thick "edge cycle" in figure 8 is oriented in one direction or the other. Each of the two points not in the cycle, $b_g b_{g'} a_{g''}$ and $a_g a_{g'} b_{g''}$, must be preferred to everything differing from it in only one player's strategy.

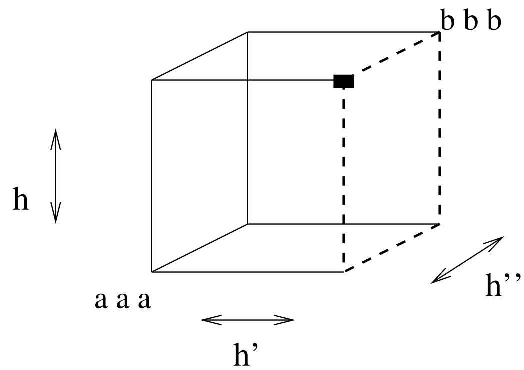

Figure 7: The sixth subform in Γ_h

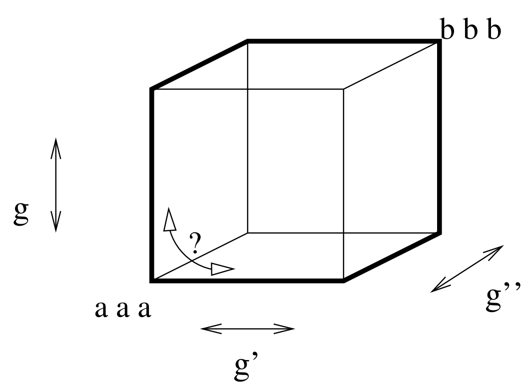

Figure 8: The “edge cycle” must be oriented for rationalizability

Since V is satisfied by T, by construction of NR_V we know that $a_1 a_{1'} a_{1''} \cdots a_{k''}$ is not a Nash equilibrium in the three-player games and the two-player games in NR_V that include it. Now for any subform in NR_V, define preferences to be consistent with the given outcomes, i.e. make the outcome *strictly* preferred to any outcome that is reachable via a one-player deviation. It is easy to see that in the subforms of Γ_h, for $h \in \{1, 2, \ldots, k\}$, this completely defines preferences without violating transitivity — the orientation of the "cycle" (as in fig. 8) is determined by the truth value of x_h. And in the three-player game forms in NR_V this defines all the relevant preferences, and the rest can be defined in an arbitrary (but consistent) way, for instance orienting every (previously undefined) edge in the graph representation in figure 1 towards the chosen outcome *bbb*.

We have shown that our polynomial transformation produces a Nash rationalizable instance of NR2 if and only if the input 3SAT instance is satisfiable. Thus if an algorithm could decide any instance of NR2 in polynomial time, then any instance V of 3SAT could be be decided in polynomial time by first using our algorithm to produce NR_V in polynomial time, and then deciding NR_V in polynomial time. Since 3SAT is NP-complete, this proves that NR2 is NP-complete. *Q.E.D.*

References

[1] E. Yanovskaya, Revealed preference in non-cooperative games, Mathematical Methods in the Social Sciences 24 (13), in Russian.

[2] Y. Sprumont, On the testable implications of collective choice theories, Journal of Economic Theory 93 (2000) 205–232.

[3] P. A. Samuelson, A note on the pure theory of consumer's behaviour, Economica 5 (1938) 61–71.

[4] H. S. Houthakker, Revealed preference and the utility function, Economica NS 17 (66) (1950) 159–174.

[5] M. K. Richter, Revealed preference theory, Econometrica 34 (1966) 635–645.

[6] M. K. Richter, Rational choice, in: J. S. Chipman, L. Hurwicz, M. K. Richter, H. Son-

nenschein (Eds.), Preferences, Utility, and Demand, Harcourt Brace Jovanovich, 1971, pp. 635–645.

[7] M. K. Richter, Rational choice and polynomial measurement theory, Journal of Mathematical Psychology 12 (1975) 99–113.

[8] A. Carvajal, I. Ray, S. Snyder, Equilibrium behavior in markets and games: testable restrictions and identification, Journal of Mathematical Economics 40 (1-2) (2004) 1–40.

[9] I. Ray, L. Zhou, Game theory via revealed preferences, Games and Economic Behavior 37 (2001) 415–424.

[10] I. Ray, S. Snyder, Observable implications of Nash and subgame-perfect behavior in extensive games, Tech. Rep. 2, Department of Economics, Brown University (2003).

[11] R. L. Matzkin, M. K. Richter, Testing strictly concave rationality, Journal of Economic Theory 53 (1991) 287–303.

[12] D.-Z. Du, K.-I. Ko, Theory of Computational Complexity, Discrete Mathematics and Optimization, John Wiley & Sons, Inc., 2000.

[13] M. R. Garey, D. S. Johnson, Computers and Intractability, W.H. Freeman and Company, 1979.

[14] S. A. Cook, The complexity of theorem-proving procedures, in: Proceedings of the third annual ACM symposium on Theory of computing, Association for Computing Machinery, 1971, pp. 151–158.