GAMES AND ECONOMIC BEHAVIOR **20,** 169–176 (1997)
ARTICLE NO. GA970559

Games of Manipulation in Marriage Problems

Tayfun Sönmez*

Department of Economics, University of Michigan, Ann Arbor, Michigan 48109-1220

Received April 8, 1996

We analyze the "equilibrium" outcomes of the *preference revelation games* induced by *Pareto efficient* and *individually rational* solutions in the context of *marriage problems*. We employ a Nash equilibrium refinement which allows deviations by a set of permissible coalitions, and show that the set of equilibrium outcomes coincides with a variant of the core that allows blocking by only permissible coalitions, *Journal of Economic Literature* Classification Numbers: C71, C78, D71, D78.

1. INTRODUCTION

In this paper we deal with manipulation and implementation of solutions in the context of *marriage problems* (Gale and Shapley, 1962). There are two disjoint sets of agents, say the set of men and the set of women. Each man has a preference relation over the set of women and staying single, and each woman has a preference relation over the set of men and staying single. An allocation is a matching of men and women. A matching is *stable* if no agent ends up worse than remaining single (i.e., if it is *individually rational*), and no man–woman pair prefer each other to their mates. The stability criterion has been central to the studies of marriage problems and to the analysis of *two-sided matching problems* in general.[1]

Recently, Alcalde (1996), Ma (1994, 1995), and Shin and Suh (1996) characterized the "equilibria" of the *preference revelation games* induced by

*I thank Jim Schummer, William Thomson, an anonymous associate editor, and an anonymous referee for their extensive comments. All errors are my own responsibility.

[1] See Roth and Sotomayor (1990) for an extensive analysis and Roth (1984, 1991) for applications in the United States and the United Kingdom medical residency markets.

0899-8256/97 $25.00

stable solutions to marriage problems.[2] It turns out that when the chosen equilibrium concept considers only unilateral deviations (i.e., when the Nash equilibrium is employed) the set of equilibrium outcomes coincides with the set of *individually rational* matchings (Alcalde, 1996); when it considers both unilateral deviations as well as deviation by pairs, the set of equilibrium outcomes coincides with the set of *stable* matchings (Ma, 1995); and when it considers deviations by all coalitions, the set of equilibrium outcomes coincides with the *core* (Ma, 1994, Shin and Suh, 1996). In this paper we unify these results and relate them to the early work of Kalai *et al*. (1979).

In many situations agents in particular coalitions (henceforth referred to as *permissible* coalitions) can coordinate their actions, whereas agents in other coalitions cannot. This observation motivates the following refinement of the Nash equilibrium due to Kalai *et al*. (1979): Let G be a set of permissible coalitions. A strategy profile is a *G-proof Nash equilibrium* if it is immune to joint deviations of agents in any permissible coalition. Similarly, they define the *G-core* to be the set of allocations such that no permissible coalition can improve the welfare of all its members by reallocation of its resources. We adopt this setup and characterize the set of *G-proof Nash equilibrium* outcomes of the *preference revelation games* induced by *Pareto efficient* and *individually rational* solutions. We show that the set of equilibrium outcomes coincides with the *G-core*.[3] The results of Alcalde (1996), Ma (1994, 1995), and Shin and Suh (1996) are corollaries to this result.

2. THE MODEL

A *marriage problem* is an ordered triplet (M, W, R) where M and W are two nonempty, finite, and disjoint sets of agents, and $R = (R_i)_{i \in M \cup W}$ is a list of preference relations of the agents. Let P_i denote the strict relation associated with the preference relation R_i for all $i \in M \cup W$. We refer to M as the set of men and W as the set of women. We consider the case where M and W are fixed and hence each problem is defined by a preference profile.

[2] Equilibria of preference revelation games are studied extensively in exchange economies: Hurwicz (1978) characterized the equilibria of the preference revelation games induced by the Walrasian solution in 2-good, 2-person economies. Otani and Sicilian (1982) extended this result to l-good, 2-person economies. Thomson (1984, 1988) characterized the equilibria of the preference revelation games induced by *monotonic* solutions and the *Shapley value* respectively in l-good, n-person economies.

[3] Kalai *et al*. (1979) construct a mechanism for which the set of *G-proof Nash equilibrium* outcomes coincides with the *G-core* in the context of public goods economies.

The preference relation R_m of each man $m \in M$ is a binary relation on $W \cup \{m\}$ which is *reflexive* (for all $i \in W \cup \{m\}$ we have $iR_m i$), *transitive* (for all $i, j, k \in W \cup \{m\}$, if $iR_m j$ and $jR_m k$ then $iR_m k$), and *total* (for all $i, j \in W \cup \{m\}$ with $i \neq j$ we either have $iR_m j$ or $jR_m i$ but not both). Such preference relations are referred to as *linear orders* (or strict preferences). Similarly the preference relation R_w of each woman $w \in W$ is a linear order on $M \cup \{w\}$. Let $\mathcal{R}_i$ be the class of all such preference relations for agent $i \in M \cup W$. Let $\mathcal{R} = \prod_{i \in M \cup W} \mathcal{R}_i$. That is, $\mathcal{R}$ is the class of all problems for M and W.

A *matching* μ is a function from $M \cup W$ onto itself such that (i) $\mu(m) \in W \cup \{m\}$ for all $m \in M$, (ii) $\mu(w) \in M \cup \{w\}$ for all $w \in W$, and (iii) $\mu(\mu(i)) = i$ for all $i \in M \cup W$. We refer to $\mu(i)$ as the *mate* of i. We denote the set of all matchings by $\mathcal{M}$. Given a preference relation R_m of a man $m \in M$, initially defined over $W \cup \{m\}$, we extend it to the set of matchings $\mathcal{M}$, in the following natural way: m prefers the matching μ to the matching μ' if and only if he prefers $\mu(m)$ to $\mu'(m)$. We slightly abuse the notation and also use R_m to denote this extension. We do the same for each woman $w \in W$.

A matching μ is *blocked by an agent* $i \in M \cup W$ under R if $iP_i\,\mu(i)$. A matching μ is *individually rational* under R if it is not blocked by any agent under R. We denote the set of individually rational matchings under R by $\mathcal{I}(R)$. A matching μ is *blocked by a man–woman pair* $(m, w) \in M \times W$ under R if $wP_m\,\mu(m)$ and $mP_w\,\mu(w)$. A matching μ is *stable* under R if it is not blocked by any agent or any man–woman pair under R. We denote the set of stable matchings under R by $\mathcal{S}(R)$. A matching μ is *Pareto efficient* under R if there is no matching μ' such that we have $\mu'(i)R_i\,\mu(i)$ for all $i \in M \cup W$ and $\mu'(i)P_i\,\mu(i)$ for some $i \in M \cup W$. We denote the set of Pareto efficient matchings under R by $\mathcal{P}(R)$.

A *matching rule* is a function $\varphi\colon \mathcal{R} \to \mathcal{M}$. A matching rule φ is *Pareto efficient* if $\varphi(R) \in \mathcal{P}(R)$ for all $R \in \mathcal{R}$; it is *individually rational* if $\varphi(R) \in \mathcal{I}(R)$ for all $R \in \mathcal{R}$; and it is *stable* if $\varphi(R) \in \mathcal{S}(R)$ for all $R \in \mathcal{R}$. A *matching correspondence* is a mapping $\psi\colon \mathcal{R} \Rightarrow \mathcal{M}$. Some examples of matching correspondences are the *individually rational correspondence* that selects the set of *individually rational* matchings and the *stable correspondence* that selects the set of *stable* matchings for each problem.

3. MANIPULATION AND IMPLEMENTATION

In many real life markets agents are asked to report their preferences and a particular matching rule is used to match them. Technically speaking, they are confronted with a game where their strategy space is a class of possible preferences and the outcome is determined by the chosen matching rule. It is very natural to study the equilibria of such games where one can consider several equilibrium notions. This problem is already studied by Alcalde (1996) who considers the Nash equilibrium, Ma (1994), and Shin and Suh (1996) who consider the strong Nash equilibrium, and Ma (1995) who considers the rematching-proof equilibrium as their underlying equilibrium concepts. They characterize the equilibria of the games induced by *stable* matching rules.

We extend these papers in two directions. First, we do not restrict ourselves to any of these equilibrium notions. In situations where agents cannot coordinate their strategies the natural equilibrium notion is the *Nash equilibrium*. In situations where all agents can coordinate their strategies, the natural equilibrium notion is the *strong Nash equilibrium*. In most real life applications, however, agents in some groups can coordinate their actions while agents in others cannot. Hence, following Kalai *et al*. (1979), we consider a class of equilibrium notions where the two polar cases are the Nash equilibrium and the strong Nash equilibrium. The second direction of extension is that we employ a wider class of matching rules, namely the class of *Pareto efficient* and *individually rational* matching rules.

We need to introduce more notation and definitions to present our results. Fix $\mathsf{G} \subseteq 2^{M \cup W} \setminus \{\emptyset\}$. Here G is the set of coalitions within which all agents can coordinate their actions. We assume that $\{i\} \in \mathsf{G}$ for all $i \in M \cup W$.

A matching $\mu \in \mathsf{M}$ is in the G-*core* of the problem R if there is no coalition $G \in \mathsf{G}$ and $\mu' \in \mathsf{M}$ such that $\mu'(i) \in G$ and $\mu'(i) P_i \mu(i)$ for all $i \in G$. Note that when $\mathsf{G} = 2^{M \cup W} \setminus \{\emptyset\}$, this definition reduces to *core*; when $\mathsf{G} = \{G \subset M \cup W : |G| = 1\}$ it reduces to *individual rationality*; and when $\mathsf{G} = \{G \subseteq M \cup W : |G| \leq 2, |G \cap M| \leq 1, |G \cap W| \leq 1\}$ it reduces to *stability*. We denote the matching correspondence that selects the G-core allocations for each problem by C^{G}.

A *mechanism* is a pair $\Gamma = (S, f) = (\prod_{i \in M \cup W} S_i, f)$ where S_i is agent i's *strategy space* and f: $S \to \mathsf{M}$ is an *outcome function*. Note that the pair (Γ, R) defines a game. In this paper we restrict our attention to a very natural class of mechanisms where $S_i = \mathsf{R}_i$ for all $i \in M \cup W$. Under this restriction any outcome function is a matching rule. Such mechanisms are often referred to as *direct mechanisms* and the resulting games are often referred to as *preference revelation games*.

Next we define a class of Nash equilibrium refinements. For all $G \in \mathsf{G}$, for all $s \in S$, let s_{-G} be the strategy tuple that is obtained from s by removing s_i for all $i \in G$ and let $S_G = \prod_{i \in G} S_i$. A strategy-tuple $s \in S$ is a G-*proof Nash equilibrium* of the game (S, f, R) if for all $G \in \mathsf{G}$, and for all $s'_G \in S_G$ there exists an agent $i \in G$ such that $f(s) R_i f(s_{-G}, s'_G)$. Note that when $\mathsf{G} = \{G \subset M \cup W : |G| = 1\}$ this definition reduces to the Nash equilibrium and when $\mathsf{G} = 2^{M \cup W} \setminus \{\emptyset\}$ it reduces to the strong Nash equilibrium.[4] We denote the set of $\mathcal{G}$-proof Nash equilibria of the game (S, f, R) by $N^{\mathcal{G}}(S, f, R)$ and the set of all equilibrium outcomes by $f[N^{\mathcal{G}}(S, f, R)]$. A mechanism $\mathcal{G} = (S, f)$ *implements the matching rule* φ *in* $\mathcal{G}$*-proof Nash equilibria*[5] if $f[N^{\mathcal{G}}(S, f, R)] = \varphi(R)$ for all $R \in \mathcal{R}$.

Now we are ready to present our main result:

THEOREM. *For any Pareto efficient and individually rational matching rule* φ *the direct mechanism* $\Gamma = (\mathcal{R}, \varphi)$ *implements the* $\mathcal{G}$*-core in* $\mathcal{G}$*-proof Nash equilibria.*

Proof. Let $\varphi: \mathcal{R} \to \mathcal{M}$ be *Pareto efficient* and *individually rational*. Let $R \in \mathcal{R}$. We prove the theorem via two claims.

Claim 1. $\mathcal{C}^{\mathcal{G}}(R) \subseteq \varphi[N^{\mathcal{G}}(\mathcal{R}, \varphi, R)]$.

Proof of Claim 1. Let $\mu \in \mathcal{C}^{\mathcal{G}}(R)$. We need to show that $\mu \in \varphi[N^{\mathcal{G}}(\mathcal{R}, \varphi, R)]$.

Let $R' \in \mathcal{R}$ be such that

1. $\forall m \in M, \forall w \in W \setminus \mu(m) \qquad \mu(m)R'_m m P'_m w,$
2. $\forall w \in W, \forall m \in M \setminus \mu(w) \qquad \mu(w)R'_w w P'_w m.$

Under R' all men rank their mates under μ at the top of their preferences and rank any other woman worse than staying single. The same holds for all women. This together with the preferences being strict imply that $\mathcal{P}(R') \cap \mathcal{I}(R') = \{\mu\}$. But we have $\varphi(R') \in \mathcal{P}(R') \cap \mathcal{I}(R')$ and therefore $\varphi(R') = \mu$. Suppose $R' \notin N^{\mathcal{G}}(\mathcal{R}, \varphi, R)$. Then there exists $G \in \mathcal{G}$, $R''_G \in \mathcal{R}_G$, and $\nu \in \mathcal{P}(R'_{-G}, R''_G) \cap \mathcal{I}(R'_{-G}, R''_G)$ such that $\nu(i) P_i \mu(i)$ for all $i \in G$. We need to consider two cases.

Case 1. For all $i \in G$ we have $\nu(i) \in G$.

Then we have $\mu \notin \mathcal{C}^{\mathcal{G}}(R)$ leading to the contradiction we are looking for.

Case 2. There exists an agent $i \in G$ with $\nu(i) \in (M \cup W) \setminus G$.

Without loss of generality suppose $i \in M$. Note that $\nu(i) P_i \mu(i) R_i i$ and therefore $\nu(i) \in W$. Let $\nu(i) = w$. Recall that $w \notin G$ and we have $\nu \in \mathcal{I}(R'_{-G}, R''_G)$. Therefore $\nu(w) \in \{\mu(w), w\}$. However, $\nu(i) P_i \mu(i)$ implies that $\nu(i) \neq \mu(i)$ and therefore $\nu(w) \neq \mu(w)$. Moreover, $\nu(w) = i \neq w$, leading to the contradiction we are looking for.

[4]A strategy-tuple $s \in S$ is a *strong Nash equilibrium* of the game (S, f, R) if for all $G \subseteq M \cup W$ and for all $s'_G \in S_G$ there exists an agent $i \in G$ such that $f(s) R_i f(s_{-G}, s'_G)$.

[5]Suh (1996) identifies the necessary and sufficient conditions for a solution φ to be implementable in $\mathcal{G}$-proof Nash equilibria.

Hence we have $R' \in N^{\mathcal{G}}(\mathcal{R}, \varphi, R)$. This, together with $\varphi(R') = \mu$, completes the proof of Claim 1.

Claim 2. $\varphi[N^{\mathcal{G}}(\mathcal{R}, \varphi, R)] \subseteq \mathcal{C}^{\mathcal{G}}(R)$.

Proof of Claim 2. Let $R' \in N^{\mathcal{G}}(\mathcal{R}, \varphi, R)$ with $\varphi(R') = \mu$. We need to show that $\mu \in \mathcal{C}^{\mathcal{G}}(R)$. Suppose not. Then there exists $G \in \mathcal{G}$ and $\nu \in \mathcal{M}$ such that $\nu(i) \in G$ and $\nu(i) P_i \mu(i)$ for all $i \in G$. Let $R''_G \in \mathcal{R}_G$ be such that

1. $\forall m \in M \cap G, \forall w \in W \setminus \nu(m) \qquad \nu(m) R''_m m P''_m w,$
2. $\forall w \in W \cap G, \forall m \in M \setminus \nu(w) \qquad \nu(w) R''_w w P''_w m.$

We have $\eta(i) = \nu(i)$ for all $i \in G$, and for all $\eta \in \mathcal{P}(R'_{-G}, R''_G) \cap \mathcal{I}(R'_{-G}, R''_G)$ as the preferences are strict and $\varphi(R'_{-G}, R''_G) \in \mathcal{P}(R'_{-G}, R''_G) \cap \mathcal{I}(R'_{-G}, R''_G)$. Therefore $\varphi_i(R'_{-G}, R''_G) = \nu(i)$ for all $i \in G$ which implies $\varphi_i(R'_{-G}, R''_G) P_i \varphi_i(R')$ for all $i \in G$ contradicting $R' \in N^{\mathcal{G}}(\mathcal{R}, \varphi, R)$. Hence $\mu \in \mathcal{C}^{\mathcal{G}}(R)$, completing the proof of Claim 2. Q.E.D.

We obtain the results of Alcalde (1996), Ma (1994, 1995), and Shin and Suh (1996) as corollaries to our theorem.

COROLLARY 1 (Alcalde, 1996).[6] *For any stable matching rule φ the direct mechanism $\Gamma = (\mathcal{R}, \varphi)$ implements the individually rational correspondence in Nash equilibria.*[7]

Proof. Let $\mathcal{G} = \{G \subset M \cup W : |G| = 1\}$. Then the $\mathcal{G}$-core is equal to the set of *individually rational* matchings, and the notion of the $\mathcal{G}$-proof Nash equilibrium reduces to the Nash equilibrium. Moreover, any matching rule that is *stable* is both *Pareto efficient and individually rational*. These observations together with the theorem complete the proof. Q.E.D.

COROLLARY 2 (Ma, 1994; Shin and Shu, 1996). *For any stable matching rule φ the direct mechanism $\Gamma = (\mathcal{R}, \varphi)$ implements the stable correspondence in strong Nash equilibria.*

Proof. Let $\mathcal{G} = 2^{M \cup W} \setminus \{\emptyset\}$. Then the $\mathcal{G}$-core is equal to the *core*, and the notion of the $\mathcal{G}$-proof Nash equilibrium reduces to the strong Nash equilibrium. Moreover, the *core* is equal to the set of *stable* matchings (Roth and Sotomayor, 1990, Theorem 3.3), and any matching rule that is *stable* is both *Pareto efficient* and *individually rational*. These observations together with the theorem complete the proof. Q.E.D.

[6] See also Roth (1985).

[7] See Kara and Sönmez (1996) for an analysis of matching rules that are Nash implementable (not necessarily via direct mechanisms) in the context of marriage problems.

Ma (1995) introduces the following equilibrium notion: A preference profile is a *rematching-proof equilibrium* of a preference revelation game if it is a Nash equilibrium and it also is immune to joint deviations by any man–woman pair.

COROLLARY 3 (Ma, 1995). *For any stable matching rule φ the direct mechanism $\Gamma = (\mathcal{R}, \varphi)$ implements the stable correspondence in rematching-proof equilibria.*

Proof. Let $\mathcal{G} = \{G \subseteq M \cup W : |G| \leq 2, |G \cap M| \leq 1, |G \cap W| \leq 1\}$. Then the $\mathcal{G}$-core is equal to the set of *stable* matchings, and the notion of the $\mathcal{G}$-proof Nash equilibrium reduces to the rematching-proof equilibrium. Moreover, any matching rule that is *stable* is both *Pareto efficient* and *individually rational*. These observations together with the theorem complete the proof. Q.E.D.

REFERENCES

Alcalde, J. (1996). "Implementation of Stable Solutions to the Marriage Problem," *J. Econ. Theory* **69**, 240–254.

Gale, D., and Shapley, L. (1962). "College Admissions and the Stability of Marriage," *Amer. Math. Monthly* **69**, 9–15.

Hurwicz, L. (1978). "On the Interaction between Information and Incentives in Organizations," *in Communication and Control in Society* (K. Krippendorff, Ed.), pp. 123–147, New York: Scientific Publishers.

Kalai, E., Postlewaite, A., and Roberts, J. (1979). "A Group Incentive Compatible Mechanism Yielding Core Allocations," *J. Econ. Theory* **20**, 13–22.

Kara, T., and Sönmez, T. (1996). "Nash Implementation of Matching Rules," *J. Econ. Theory* **68**, 425–439.

Ma, J. (1994). "Manipulation and Stability in a College Admissions Problem," Rutgers University mimeo.

Ma, J. (1995). "Stable Matchings and Rematching-Proof Equilibria in a Two-Sided Matching Market," *J. Econ. Theory* **66**, 352–369.

Otani, Y., and Sicilian, J. (1982). "Equilibrium of Walras Preference Games," *J. Econ. Theory*, **27**, 47–68.

Roth, A. (1984). "The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory," *J. Polit. Econ.* **92**, 991–1016.

Roth, A. (1985). "The College Admissions Problem is not Equivalent to the Marriage Problem," *J. Econ. Theory*, **36**, 277–288.

Roth, A. (1991). "A Natural Experiment in the Organization of Entry Level Labor Markets: Regional Markets for New Physicians and Surgeons in the U.K.," *Amer. Econ. Rev.* **81**, 415–440.

Roth, A., and Sotomayor, M. (1990). *Two-Sided Matching: A Study in Game Theoretic Modeling and Analysis*. London/New York: Cambridge Univ. Press.

Shin, S., and Suh, S-C. (1996). ‘‘A Mechanism Implementing the Stable Rule in Marriage Problems,’’ *Econ. Lett.* **51**, 185–189.

Suh, S-C. (1996). ‘‘Implementation with Coalition Formation: A Complete Characterization,’’ *J. Math. Econ.* **26**, 409–428.

Thomson, W. (1984). ‘‘The Manipulability of Resource Allocation Mechanisms,’’ *Rev. Econ. Stud.* **51**, 447–460.

Thomson, W. (1988). ‘‘The Manipulability of the Shapley Value,’’ *Int. J. Game Theory* **17**, 101–127.