JOURNAL OF ECONOMIC THEORY 34, 383–387 (1984)

Misrepresentation and Stability in the Marriage Problem

ALVIN E. ROTH*

Department of Economics, University of Pittsburgh, Pittsburgh, Pennsylvania 15260

Received September 28, 1983; revised March 7, 1984

It has been shown previously that, for two-sided discrete markets of the kind exemplified by the "marriage problem," no strategy-proof procedure for aggregating preferences into stable outcomes exists. Here it is shown that (Nash) equilibrium misrepresentation of preferences nevertheless results in a stable outcome in terms of the true preferences when the aggregation procedure yields the optimal stable outcome in terms of the stated preferences for one side of the market. *Journal of Economic Literature* Classification Numbers: 022, 025, 026.

1. INTRODUCTION

Perhaps the simplest kind of two-sided discrete market is the "marriage problem," which involves two disjoint sets of agents, M and W ("men" and "women" or "firms" and "workers"), with each agent having a preference order defined on agents of the opposite set. To simplify the exposition, it is assumed here that the two sets are of equal size and that, for each agent, being matched with any agent of the opposite set is preferable to being unmatched. An outcome of such a market is a (monogomous) pairing of agents, which can be represented by an invertible function x from M to W.

An outcome x is *unstable* if there exists an agent m_i in M and an agent w_j in W who are not paired with one another (i.e., w_j does not equal $x(m_i)$) and who prefer one another to the partner they are paired with. An outcome x which is not unstable in this way is called *stable*. It is readily verified that the set of stable outcomes equals the core of the cooperative game whose rules are that any two agents from opposite sets may be paired together if they both agree.

* This work has been partially supported by grants from the National Science Foundation and the Office of Naval Research, and by Fellowships from the John Simon Guggenheim Memorial Foundation and the Alfred P. Sloan Foundation.

0022-0531/84 $3.00

642/34/2-13

Gale and Shapley [1] showed that the set of stable outcomes is nonempty for any preferences of the agents. They further showed that, when all agents' preferences are strict, the set of stable outcomes contains an "M-optimal" and a "W-optimal" stable outcome, such that no agent in M receives a more preferred match at any stable outcome than at the M-optimal stable outcome, and no agent in W receives a more preferred match at any stable outcome than at the W-optimal stable outcome. Gale and Shapley's proof was by means of an algorithm that constructed the M-optimal stable outcome. A number of recent papers have generalized these results to models more easily interpretable as labor markets (see, e.g., [2], [5]).

A related line of inquiry focusses on the incentives agents have to reveal their true preferences. Since each agent's preferences are known to him alone, any procedure that determines an outcome as a function of agents' stated preferences induces a noncooperative game in which each agent's strategy set consists of all the possible preferences he might state. In Roth [3] it was shown that no procedure yielding a stable outcome as a function of stated preferences exists for which truthful revelation of preferences is a dominant strategy for all agents.

In view of this, it is reasonable to ask if it is consistent to expect that rational agents can achieve outcomes that are stable with respect to the agents' true preferences, since even when a procedure yielding stable outcomes in terms of stated preferences is employed, some agents will typically have an incentive to misrepresent their preferences. To phrase this question more precisely, we can ask if stable outcomes are *implementable*; i.e., if any procedures exist which yield a stable outcome in terms of agents' true preferences, when agents misrepresent their preferences optimally with respect to one another, so that their (mis) stated preferences constitute a Nash equilibrium.

This question is answered below, in the affirmative. It is shown that procedures which yield the M-optimal or W-optimal stable outcomes when all preferences are strict, and which have consistent tie-breaking procedures to deal with nonstrict preferences, yield stable outcomes in terms of agents' true preferences when equilibrium misrepresentation occurs.

2. Equilibrium Misrepresentation

First consider the case in which all agents have strict preferences. The algorithm proposed by Gale and Shapley [1] is the following.

Step 1. (a) Each m_i in M proposes to his most preferred w_j in W.

(b) Each w_j rejects all but her most preferred m_i in the set proposing to her, who she keeps "engaged."

Step k. (a) Each m_i who has been rejected in the previous step proposes to his most preferred w_j among those who have not yet rejected him (i.e., to whom he has not yet proposed).

(b) Each w_j keeps engaged her most preferred m_i from among her current proposers (including anyone kept engaged from step $k-1$) and rejects the rest.

The algorithm terminates at any step when no m_i is rejected, at which point each m_i is matched with the w_j to whom he is engaged. The resulting outcome is stable, since each m_i who prefers some w_k to his assigned partner has been rejected by w_k in favor of someone she prefers. And it is straightforward to verify that the resulting outcome is the M-optimal stable outcome, since it can be shown by induction that no m_i is ever rejected by any "achievable" w_j; i.e., one with whom he is matched at any stable outcome. Thus the outcome produced by this algorithm matches each m_i with his most preferred achievable partner.

From the point of view of what incentives agents have to reveal their true preferences, two procedures are entirely equivalent if they yield the same outcome when given the same preferences. Thus we can speak of the incentives in any algorithm yielding the M-optimal stable outcome, for example, without specifying the internal steps involved in the algorithm (cf. the equivalent algorithm studied in [4]). The following result was proved in [3].

PROPOSITION. *In a matching procedure which always yields the M-optimal stable outcome of the marriage problem, truthful revelation is a dominant strategy for all agents in M.*

The symmetric result obviously holds for procedures yielding the W-optimal stable outcome.

When preferences are not strict, a stable outcome can be achieved by employing the above algorithm together with any arbitrary (but fixed) tie-breaking procedure for use when an agent expresses indifference between two alternatives. The initial step of such an algorithm would be to convert any stated preference relation R, which might contain some indifferences, into a strict preference P that is identical to R except when R indicates indifference, in which case P resolves the indifference in some fixed way (e.g. by making the strict preference conform to alphabetical order). The algorithm would then proceed as before, using for each agent the strict preference derived in this way from the agent's stated preference. Such an algorithm will be referred to as an M-optimal stable procedure with tie-breaking, since it produces an M-optimal stable outcome when preferences are strict. When preferences are not strict, the outcome of such a procedure will continue to be stable, but no M-optimal stable outcome need exist.

Note that the proposition holds not only when all preferences are strict, but also applies to M-optimal stable procedures with tie breaking as well. To see this, consider an agent m_i whose true preference is R, which is converted into the strict preference P by the algorithm. Then the fact that it would be a dominant strategy for an agent whose true preference was P to correctly state P implies that it is a dominant strategy for m_i to correctly state R as his preferences.

Thus in M-optimal stable procedures with tie-breaking, it is a best-reply for each m_i in M to state his true preferences, no matter what preferences may be stated by other agents. This is not also the case in these procedures for agents w_j in W. However we can state the following.

THEOREM. *In the noncooperative preference-revelation game which arises when an M-optimal stable procedure with tie-breaking is employed in the marriage problem, every Nash equilibrium at which no (weakly) dominated strategy is employed yields a stable outcome with respect to the true preferences.*

The requirement that no weakly dominated strategies are employed serves to exclude from consideration any Nash equilibrium at which some m_i might do no worse by stating some incorrect preference than by stating his correct preference. Since each m_i can do no better than to state his true preference, this is not a very substantive exclusion.

Proof. From the point of view of the incentives to the agents, any two M-optimal stable procedures with tie-breaking are equivalent if they break ties in the same way. We may therefore consider, without loss of generality, a procedure which employs the algorithm described above, once preferences have been transformed into strict preferences. For a $2n$-tuple of stated preferences $R = (R(m_1),..., R(m_n), R(w_1),..., R(w_n))$, let $G(R)$ denote a specific realization of this procedure, and denote its outcome by $g(R)$.

Let R be the $2n$-tuple of stated preferences of all agents, and let $x = g(R)$ be the resulting outcome. Let R be a Nash equilibrium containing no weakly dominated strategies, and suppose, contrary to the theorem, that x is unstable with respect to the agents' true preferences. Then there is a pair of agents m_i and w_j such that m_i prefers w_j to $x(m_i)$, and w_j prefers m_i to $x^{-1}(w_j)$. But $R(m_i)$ is agent m_i's true preference, so m_i proposed to w_j at some step in $G(R)$, was rejected, and went on to propose to $x(m_i)$.

Let $R'(w_j)$ be the preference relation which ranks m_i strictly preferred to all other choices, and which agrees with $R(w_j)$ on every comparison that does not include m_i. Let R' be the $2n$-tuple of preferences in which w_j's stated preference is $R'(w_j)$, but which equals R in all components except that corresponding to w_j, and let $y = g(R')$ be the outcome of the algorithm when preferences R' are stated. Then $y^{-1}(w_j) = m_i$, since m_i proposes to w_j at the

same step of $G(R')$ as in $G(R)$, but is not subsequently rejected in $G(R')$. So stating $R'(w_j)$ gives w_j a more preferred outcome than stating $R(w_j)$, given that other agents' stated preferences are given by R. This contradicts the assumption that R is a Nash equilibrium, and proves the theorem.

The theorem shows that no inconsistency is involved in supposing that rational agents can achieve stable outcomes in terms of their true preferences, even though these true preferences will not in general be revealed.

A number of related results and a bibliography of related work is contained in [4]. In [6] the incentives of agents to reveal their preferences are studied in more complex kinds of two-sided markets, such as those explored in [4] and [5], in which agents may be matched with more than one agent on the other side of the market.

REFERENCES

1. D. GALE AND L. SHAPLEY, College admissions and the stability of marriage, *Amer. Math. Monthly* **69** (1962), 9–15.
2. A. S. KELSO, JR., AND V. P. CRAWFORD, Job matching, coalition formation, and gross substitutes, *Econometrica* **50** (1982), 1483–1504.
3. A. E. ROTH, The economics of matching: Stability and incentives, *Math. Oper. Res.* 7 (1982), 617–628.
4. A. E. ROTH, The evolution of the labor market for medical interns: A case study in game theory, *Journal of Political Economy*, forthcoming.
5. A. E. ROTH, Stability and polarization of interests in job matching, *Econometrica* 52 (1984), 47–57.
6. A. E. ROTH, Incentives in the college admissions problem and related two-sided markets (1984), mimeo.