Protective Behavior in Matching Models*

S. BARBERÀ

Departament d'Economia i d'Història Econòmica, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain

AND

B. DUTTA

Indian Statistical Institute, New Delhi—110016 India

Received January 30, 1991

This paper analyzes the use of a version of lexical maximin strategies, called protective behavior, in two-sided matching models. It restricts attention to mechanisms which produce *stable* matchings, that is, matchings which are individually rational and pairwise optimal. The main results of the paper show that truth-telling is the *unique* form of protective behavior in two such mechanisms. The first is the one which selects the student-optimal stable matching in the college-admissions model, while the second is the mechanism which selects the buyer-optimal matching in the Shapley–Shubik assignment model. *Journal of Economic Literature* Classification Numbers: C78, D81.

1. INTRODUCTION

In this paper we test a very general insight about the incentive properties of collective decision-making mechanisms against the specific yet rich and relevant class of two-sided matching models. The insight is the following:

* This paper was written when the second author was visiting the Instituto de Análisis Económico, C.S.I.C., Barcelona. The authors thank the Secretaría de Estado de Universidades e Investigación for financial support through its program for sabbaticals and through C.I.C.Y.T. Grant PB89-0294. The authors gratefully acknowledge the helpful comments of the anonymous referee and Associate Editor.

although not in general a dominant strategy, truth-telling may become a very salient and sensible type of behavior in many revelation games, provided agents are sufficiently risk averse and poorly informed about other players. We shall provide a specific formulation of this general idea through the analysis of what we call protective behavior, a concept we introduced in Barberà and Dutta (1982) and that we extend here to cover a richer class of situations. This general insight is then tested against the performance of mechanisms that are specifically designed to solve two-sided matching problems. In this introduction we shall elaborate on the interest and scope of matching models, on the adequacy of our concept of protective behavior, and on the relevance of our results.

There is by now a vast literature on two-sided matching problems.[1] Matching processes have been modeled as cooperative and noncooperative games. The alternatives faced by society may be finite (as in the marriage problem or the college-admissions model, where the only relevant decision involves matching agents on one side of the market to agents on the other side) or infinite (as in the assignment game, where monetary payments between agents who are matched are also part of the outcome). The domains of preferences under consideration are also different in each case, and in our analysis of strategic behavior we retain the assumptions on preferences that are standard for each model. But the basic questions addressed through these models are common to all.

The existence of stable matchings, the structure of the set of such matchings and the design of algorithms to compute them have been major concerns of the literature. The incentives for agents to act strategically under matching mechanisms have also been scrutinized, with special attention paid to those mechanisms that guarantee stable matchings. No mechanism which always produces stable matchings with respect to stated preferences can make it a dominant strategy for all agents to always reveal their true preferences. However, some of the best known mechanisms will always make truth a dominant strategy for one side of the market. These facts are due to Roth (1982) for the discrete models and to Demange (1982) and Leonard (1983) for the assignment game. The nonexistence of dominant strategy mechanisms opens the door to more detailed analysis of strategic behavior, which will now be dependent on the specific setup and equilibrium concept to be used. For example, Roth (1984) showed that in the one-to-one models, the mechanism that always yields the optimal stable matching for one side of the market has the property that every Nash equilibrium in undominated strategies will yield a matching that is stable with respect to the true preferences. This result, therefore, guarantees that if we can reasonably expect that only strategy n-tuples that are

[1] Roth and Sotomayor (1990) provide an excellent survey and presentation.

Nash equilibria in undominated strategies will be used, then unstable matchings will not occur.

Although implementation results of this type are powerful, they implicitly assume that the preferences of all agents are common knowledge. This is unlikely to be the case in many situations. As Roth and Vande Vate (1989) remark, "one of the difficulties that arises in attempting to apply theoretical studies of equilibrium to empirical studies is that the information required for agents to implement some kinds of equilibrium strategies frequently exceeds the information that agents can reasonably be thought to have." One alternative is to model the market as a game of incomplete information, as in Roth (1989). But an attempt to model actual markets as a game of incomplete information raises conceptual problems of its own because of the strong assumption that agents share a common prior probability distribution.[2]

In this paper we consider a framework which is the polar extreme of the complete information model. Specifically, we analyze situations where agents adopt what we call protective behavior, and we argue that such behavior, which is based on a refinement of the traditional maxmin criterion, would be adequate if agents were extremely risk averse and had no information at all about the other agents' preferences. The concept of protective behavior is based on a binary comparison between strategies. Strategy s protectively dominates strategy s' for agent i under a given social decision rule if: (a) i is guaranteed to stay above a certain utility level $r(s)$ against a larger set of actions from the rest of society by using strategy s than by using strategy s', and (b) s and s' lead to the same outcomes for all actions of others under which i would obtain a utility level below $r(s)$. Protective domination is a transitive but not complete relation. A protective strategy is one that is not protectively dominated by any other. The notion of protective behavior when the set of alternatives is finite was introduced in Barberà and Dutta (1982) and characterized in Barberà and Jackson (1988). In the present paper we extend it to the

[2] This induces Roth and Vande Vate (1989) to concentrate on a class of "plausible and informationally parsimonious" strategies, called *truncated* strategies. As the name suggests, a truncated strategy is simply a truncation of a player's true preference list. Roth and Vande Vate (1989) study marriage markets in which the matchings are arranged through a random process. They show that a player will always have a truncated strategy as a best response to any strategy combination used by other players. Moreover, any stable matching can be achieved as a Nash equilibrium in undominated strategies via truncated strategies. However, though truncated strategies are simpler to use, it is not clear to us that truncated strategies are actually "informationally parsimonious." *All* truncations are not best responses. In order to find out *which* truncation to use, an agent must know the strategy combination being used by the other players. Thus, we are back to the complete information framework after all.

infinite alternative case. Comparisons with a related concept, Moulin's notion of prudent strategy, are found in Section 2.

We prove that there are different matching models under which truthful revelation of preferences is the *unique* protective strategy for all agents. Specifically, this is the case for the mechanism that would always choose the buyers' optimal stable matching in the assignment model and for the one choosing the students' optimal stable matching in the college admissions problem (the latter also covers the men's optimal and women's optimal mechanism for the marriage problem). It is interesting to notice, however, that truth-telling plays a somewhat different role in each of these models: it protectively dominates all other strategies in the college admissions and the marriage problem, while in the assignment game it no longer dominates all others, but still remains the only protectively undominated strategy. We also show that other natural mechanisms, like the one that would always choose the college-optimal matching, do not share the same property.

We want to stress the importance of the fact that under the above-mentioned mechanisms truth-telling is the *unique* protective strategy. Indeed, in these and many other contexts, truth-telling is a maxmin strategy, but so are many other strategies. Because of that, statements about maxmin behavior become extremely inconclusive. This was noted by Thomson (1979) and Dasgupta *et al.* (1979), and it partly justifies the sparse interest for maxmin type behavior in the recent literature on implementation, in spite of the earlier attention it got in seminal papers like Drèze and de la Vallée Poussin (1971). By showing here that truth-telling stands out as the unique protective strategy, we provide an appropriate statement for the intuition that truth-telling may be a plausible form of behavior in situations involving uncertainty.

2. Protective Behavior

In this section, we describe the concept of *protective behavior*. We define this concept for an arbitrary game in normal form. In subsequent sections, we will point out how the concept can be applied to the specific matching models considered in this paper.

Let $I = \{1, 2, \ldots, n\}$ be a set of individuals, and A a set of *outcomes*. The set of outcomes may be finite or infinite. Each individual $i \in I$ has a real-valued utility function $u_i : A \to \mathbb{R}$. Note that individual utilities may be given *ordinal* or *cardinal* meaning, depending upon the specific context. Let $\underline{u} = (u_1, \ldots, u_n)$ denote a vector of utility functions.

For each $i \in I$, let S_i be individual i*'s strategy set*, and $S = \prod S_i$. Let

$g : S \to A$ be an *outcome function*, specifying an outcome for each n-tuple of strategies in S. $G = [I, A, S, g, \underline{u}]$ constitutes a game in normal form.

In order to introduce the concept of protective behavior in G, we need some further notation. Choose any real number $k \in \mathbb{R}$, any $i \in I$ and $s_i \in S_i$. Define the set $c(k, s_i)$ as follows:

$$c(k, s_i) = \{s_{-i} \in S_{-i} \mid u_i(g(s_i, s_{-i})) = k\}.$$

Thus, $c(k, s_i)$ is the set of $(n-1)$-tuples of "complementary" strategies which in conjunction with s_i gives individual i a utility level of k.

DEFINITION 1. For any $i \in I$, and $s_i, s_i' \in S_i$, s_i *protectively dominates* s_i', denoted $s_i\, d(u_i)\, s_i'$, if there exists $k \in \mathbb{R}$ such that

$$\text{(i)}\ \ c(r, s_i) \cap c(r', s_i') = \varnothing \qquad \text{for all } r \leq k \text{ and } r < r',$$
$$\text{(ii)}\ \ c(k, s_i) \subset c(k, s_i').$$

Let $D(u_i) = \{s_i \in S_i \mid \nexists s_i' \in S_i$ such that $s_i'\, d(u_i)\, s_i\}$. $D(u_i)$ is the set of protective strategies of individual i in the game $G = [I, A, S, g, \underline{u}]$.

Suppose s_i and s_i' satisfy conditions (i) and (ii) of Definition 1. Condition (i) guarantees that *up to* the threshold utility level of k, individual i cannot lose by employing the strategy s_i instead of s_i'. For suppose s_{-i} is a complementary strategy profile such that $u_i(g(s_i, s_{-i})) = r \leq k$. Then, condition (i) ensures that $u_i(g(s_i', s_{-i})) \leq r$. Moreover (ii) implies that there are complementary profiles s_{-i} such that $u_i(g(s_i', s_{-i})) = k$ and $u_i(g(s_i, s_{-i})) > k$. Thus, if individuals are extremely risk averse and hence more concerned with avoiding "disasters," the concept of protective behavior is an appealing option.

The concept of protective behavior is closely related to Moulin's (1981) notion of *prudent behavior*. We define below prudent strategies in a game G where the set of strategies for any player is finite.

DEFINITION 2. For any $i \in I$, a strategy s_i *prudently dominates* s_i' iff there exists $k \in \mathbb{R}$ such that

$$\text{(i)}\ \ |c(r, s_i)| \leq |c(r, s_i')| \qquad \text{for all } r \leq k,$$
$$\text{(ii)}\ \ |c(k, s_i)| < |c(k, s_i')|.$$

(We use the notation $|c|$ to denote the cardinality of set c.)

As we have remarked in Barberà and Dutta (1982), the concept of prudent behavior implicitly assumes that an individual considers all complementary profiles to be equally likely, whereas protective behavior does not require individuals to have any subjective probability distribution about other individuals' strategies. We point out at this stage that in any

game G where agents have a finite set of strategies, individual i's set of protective strategies $D(u_i)$ will be a superset of i's set of prudent strategies. Since our purpose is to highlight situations in which truthtelling is the *unique* protective strategy in various matching models, it follows that analogous statements are true when protective behavior is replaced by prudent behavior.

DEFINITION 3. For any $i \in I$, strategies $s_i, s_i' \in S_i$ are *equivalent* iff for all $s_{-i} \in S_{-i}$, $g(s_i, s_{-i}) = g(s_i', s_{-i})$.

It follows that if $s_i \in D(u_i)$, then so must any strategy s_i' which is equivalent to s_i. The reader should keep in mind that our subsequent results which show that truth-telling is the *unique* protective strategy are modulo equivalence.

3. PROTECTIVE BEHAVIOR IN THE MARRIAGE AND COLLEGE ADMISSIONS MODEL

In this section, we will analyze protective behavior in some simple two-sided matching models without money.

Since utility is ordinal in these models, we will use individual *preference orderings* instead of utility functions in this section.

We will first describe the simplest two-sided matching model in the literature—the marriage market of Gale and Shapley (1962). There are two finite, disjoint sets M and W; $M = \{m_1, \ldots, m_n\}$ is the set of men, and $W = \{w_1, w_2, \ldots, w_k\}$ is the set of women. Each man has preferences over women. Similarly, each woman has preferences over men. These preferences may be such that man m would prefer to remain single rather than be married to a woman w, say, whom he particularly dislikes. A woman w is *acceptable* to man m if he prefers her to remaining single. Similarly, man m is acceptable to w if she prefers him to remaining single. We will assume that everyone's preference is *strict*. The preference of each man m will be represented by an ordered list of preferences, $P(m)$, on the set $W \cup \{m\}$. That is m's preference may be of the form:

$$P(m) = w_1, w_3, \ldots, w_k$$

indicating that his first choice is to be married to w_1, his second choice is to be married to w_3 and he would much rather remain single than be married to anyone else. Similarly, each woman w has an ordered list of preferences $P(w)$, on the set $M \cup \{w\}$.

Let $\mathbf{P} = \{P(m_1), \ldots, P(m_n), P(w_1), \ldots, P(w_k)\}$ denote the set of

preferences, one for each man and woman. We will often use the notation P_{-w} (or P_{-m}) to represent the preferences of agents other than w (or m).

An outcome in this market is a set of marriages. Of course, some individuals may not find a partner.

DEFINITION 4. A matching μ is a one-to-one correspondence from the set $M \cup W$ onto itself such that for each $m \in M$ and $w \in W$, $\mu(m) = w$ iff $\mu(w) = m$ and if $\mu(m) \notin W$, then $\mu(m) = m$, and similarly if $\mu(w) \notin M$, then $\mu(w) = w$.

A matching μ is individually rational if each agent is acceptable to his or her pair. For a given matching μ, a pair (m, w) forms a blocking pair if $\mu(m) \neq w$ and if they both prefer each other to their mates at μ. A matching is *stable* if it is individually rational and there are no blocking pairs.[3]

Gale and Shapley (1962) proved that for any set of preferences P, the set of stable matchings is nonempty. Moreover, when all agents have strict preferences, there always exists a stable matching that all men (resp. women) will unanimously prefer to any other stable matching. This is called the men's optimal (resp. women's optimal) stable matching. Gale and Shapley also constructed an explicit algorithm to locate the optimal stable matching for either side of the market.

We now describe a simple model of many-to-one matchings—the "college admissions" model. This model is meant to represent situations where one side of the market consists of institutions and the other side of individuals. The institutions (colleges, firms) may be matched to several individuals (students, workers), but each individual is matched to only one institution. More formally, first define for any set X, an *unordered family of elements* of X to be a collection of elements, not necessarily distinct, in which the order is immaterial. Hence, a given element may appear more than once in an unordered family. Let C be the set of institutions (colleges), and S the set of individuals (students). As in the marriage model, each college c has preferences $P(c)$ over the set $S \cup \{c\}$, while each student s has preferences $P(s)$ over the set $C \cup \{s\}$.

DEFINITION 5. A matching μ is a function from the set $C \cup S$ into the set of unordered families of elements of $C \cup S$ such that:

1. $|\mu(s)| = 1$ for every student s and $\mu(s) = s$ if $\mu(s) \notin C$.
2. $|\mu(c)| = q_c$ for every college c, and if the number of students in $\mu(c)$, say r, is less than q_c, then $\mu(c)$ contains $q_c - r$ copies of c.
3. $\mu(s) = c$ if $s \in \mu(c)$.

Condition 1 says that a student can be matched to at most one college.

[3] In this model, the set of stable matchings coincides with the core of the game.

Condition 2 states that each college has a quota q_c, so that it can enroll at most q_c students although it can also keep some positions unfilled. Condition 3 requires that if a student is matched to a college, then this college is matched to the student.

To complete the description of the model, we have to describe the preferences of agents over different outcomes. In the marriage market, preferences could be described very easily because agents' preferences over alternative matchings coincided with their preferences over their own assignments at the two matchings. In this model, we can still say the same thing about the students since at each matching, a student is either unmatched or matched to a *single* college. But colleges having a quota greater than 1 must be able to compare *groups* of students on the basis of their preference over single students.[4] Following Roth (1985), we will assume that colleges are endowed with preferences $\hat{P}(c)$ over groups of students satisfying the following condition of responsiveness.

DEFINITION 6. The preference relation $\hat{P}(c)$ over sets of students is *responsive* to the preferences $P(c)$ over individual students if, whenever

$$\mu'(c) = (\mu(c) \cup \{\delta\}) \backslash \{\sigma\} \qquad \text{for } \sigma \in \mu(c) \text{ and } \delta \notin \mu(c),$$

then $\mu(c)\hat{P}(c)\mu'(c) \leftrightarrow \sigma \hat{P}(c)\delta$.

Thus, when $\mu(c)$ and $\mu'(c)$ only differ because one is obtained from the other by exchanging a student σ for another student δ, the college should rank these two sets according to their ranking of the two students that make the difference.

Note that the assumption of responsiveness is a rather weak requirement because restrictions are imposed on how colleges compare some groups of students only. For instance, if one group consists of the first and fourth most preferred students, and the other group contains only the second and third most preferred students, then the condition does not apply.

In addition to responsiveness, we assume that for each college c, $\hat{P}(c)$ is complete and transitive.

A matching is individually rational if no student is matched to an unacceptable college, and no college is matched to any unacceptable student. A college c and student s form a blocking pair to μ if they are not matched to one another at μ, but would prefer to be matched to one another than to (one of) their present assignments. A matching is *stable* if it is individually rational and it is not blocked by any student–college pair. Clearly, the notions of individual rationality, blocking pairs and stable

[4] There is a sizable literature on the general problem of extending preferences over sets to the power set. See, for instance, Barberà and Pattanaik (1984) and Kannai and Peleg (1984).

matchings in the college admissions model are straightforward generalizations of the corresponding concepts in the marriage market.

We now briefly comment on how the definition of protective behavior given in Section 2 can be applied in the marriage market and college admissions models. Note first that the definition can be applied almost straight-away in the case of the marriage model. Here, the strategy set of any agent m(say) is simply the set of possible preference orderings over $W \cup \{m\}$, while the outcome set is the set of possible matchings for m. Suppose m's preference ordering is:

$$P^*(m) = w_1, w_2, \ldots, w_k, m.$$

Then, in comparing two alternative strategies $P(m)$ and $P'(m)$ under say the M-optimal stable matching μ_M, m first compares the set of complementary profiles which together with $P(m)$ leaves him unmatched to the set of complementary profiles which leaves him unmatched when he uses $P'(m)$. If these two sets are identical, then the next round of comparisons is in terms of the sets of complementary profiles which result in w_k, and so on.

In the college admissions model, the above interpretation remains valid for the students. However, the situation is more complicated on the college side of the market. The set of strategies for colleges will still be preference orderings over the set of students. But, for any college c with quota q_c greater than 1, the set of outcomes will be sets of students with cardinality not exceeding q_c. Hence, a college will use $\hat{P}(c)$ to evaluate alternative strategies.

We will now show that in the college admissions model, the mechanism which always chooses the student optimal matching has the property that truthful revelation of preferences is the unique protective strategy for all agents. Since the marriage model is a special case of the college admissions model, our result obviously goes through for the marriage market. Also, note that in the marriage market, both men and women have a symmetric role. Hence, truthful revelation of preferences will be the unique protective strategy in both the M-optimal as well as the W-optimal stable matchings.

The following notation will be useful in the sequel. Given any subset $Y \subset A$ and a preference ordering P over A, we say that Y is *bottom* for P if, $\forall y \in Y$, $\forall x \in A \backslash Y$, xPy. Two preference orderings P and P' *agree* on Y if $\forall x, y \in Y$, xPy iff $xP'y$.

For any college c or student s, we will denote complementary profiles by P_{-c} or P_{-s}.

Given any preference ordering P, and any integer r, *the rth worst alternative in* A is $a_r(P) = \{x \in A \mid \exists$ exactly $(r-1)$ alternatives $y \in A : xPy\}$.

We now state the main result in this section.

THEOREM 1. *Under the mechanism which always yields the student optimal matching μ_S, truthtelling is the unique protective strategy for all agents.*

The proof of the theorem is preceded by a few lemmas. In this lemmas, choose a specific college c, a preference ordering $P(c)$ and some $\hat{P}(c)$ which is responsive to $P(c)$. Let S^u denote the set of unacceptable students for college c according to $P(c)$. The lemmas below specify properties of the set of protective strategies under the mechanism which chooses the matching μ_S.

LEMMA 1. *If $P(c)$ and $P'(c)$ agree on $S\backslash S^u$ and S^u is bottom for $P'(c)$, then $P(c)$ and $P'(c)$ are equivalent strategies.*

In words: the order of unacceptable students in the colleges' preferences does not matter.

Proof. Since μ_S is individually rational, $P(c)$ and $P'(c)$ both guarantee that c is not matched to any unacceptable student. Since the only difference between $P(c)$ and $P'(c)$ is in the ranking of unacceptable students, the result follows.

LEMMA 2. *Let S' be the q_c-top ranked alternatives according to $P(c)$. If $P'(c)$ and $P(c)$ agree on $(S \cup \{c\})\backslash S'$, and $(S \cup \{c\})\backslash S'$ is bottom for $P'(c)$, then $P'(c)$ and $P(c)$ are equivalent strategies.*

In words: the order of students which are all top and within a colleges's quota does not matter.

Proof. If $P'(c)$ satisfies the hypothesis of the lemma, then the only difference between $P'(c)$ and $P(c)$ is in the ranking of the alternatives in S'. It is easy to check that the two orderings must give the same outcome for all complementary profiles.

LEMMA 3. *If for some $s \in S^u$, $sP'(c)c$, then $P(c)d(P(c))P'(c)$.*

In words: all strategies which treat an unacceptable student as acceptable are protectively dominated by the truth.

Proof. Suppose c expresses the preference ordering $P(c)$. Since μ_S is individually rational, c cannot then be matched to any unacceptable student. Now consider the following P^*_{-c}.

1. $\forall s' \neq s$, c is unacceptable to s'.
2. c is the only acceptable college for s.

Clearly, $\mu_S(P'(c), P^*_{-c}, c) = \{s, c, \ldots, c\}$; i.e., c is matched to s, with $q_c - 1$ positions remaining unfilled. On the other hand, $\mu_S(P'(c), P^*_{-c}, c) = \{c, \ldots, c\}$; i.e., c does not fill up any position. Since s is unacceptable

to c, we must have $P(c)d(P(c))P'(c)$. This completes the proof of the lemma.

LEMMA 4. *If for some* $s \in S \backslash S''$, $cP'(c)s$, *then* $P(c)d(P(c))P'(c)$.

In words: all strategies which treat an acceptable student as unacceptable are protectively dominated by truth.

Proof. The proof is similar to that of Lemma 3, and is therefore omitted.

We need some additional notation before we can state the next lemma. Fix a complementary profile P_{-c} for college c. Suppose $\mu_S(P(c), P_{-c}, c) = \{s_j, \ldots, s_k\} = \hat{S}$, where s_k is the worst option for c according to $P(c)$ in the set $\hat{S}$. (We do not rule out the possibility that $s_k = c$; i.e., c may not fill up all positions at $\mu_S(P(c), P_{-c})$.) Let $S' = \{s \in S \mid s_k P(c) s\}$. S' is obviously a bottom set for $P(c)$. Let $P'(c)$ be any preference ordering such that S' is a bottom set for $P'(c)$.

Let $\mathbf{P} = (P(c), P_{-c})$, $\mathbf{P}' = (P'(c), P_{-c})$.

LEMMA 5. *Suppose* $\mu_S(\mathbf{P})$ *is individually rational for* $\mathbf{P}'$. *Then,*

$$\mu_S(\mathbf{P}) \neq \mu_S(\mathbf{P}') \Rightarrow \mu_S(\mathbf{P}, c)\hat{P}(c)\mu_S(\mathbf{P}', c).$$

Proof. First, we show that $\mu_S(\mathbf{P})$ is a stable matching with respect to the preference profile $\mathbf{P}'$. For, suppose that $\mu_S(\mathbf{P})$ is not stable with respect to $\mathbf{P}'$. Then the blocking pair must contain c. Suppose (c, s) is a blocking pair. Then $sP'(c)s_k$. But $\{s' \in S \mid s_k P(c) s'\}$ is bottom for $P'(c)$. So $sP'(c)s_k \rightarrow sP(c)s_k$. But, then (c, s) will also constitute a blocking pair to $\mu_S(\mathbf{P})$. So $\mu_S(\mathbf{P})$ is a stable matching under profile $\mathbf{P}'$.

Suppose $\mu_S(\mathbf{P}) \neq \mu_S(\mathbf{P}')$. Since μ_S is the student-optimal stable matching, $\forall s \in S$, either $\mu_S(\mathbf{P}', s)P(s)\mu_S(\mathbf{P}, s)$ or $\mu_S(\mathbf{P}', s) = \mu_S(\mathbf{P}, s)$. Consider the set of students $S^1 = \{s \in S \mid \mu_S(\mathbf{P}', s) = c \neq \mu_S(\mathbf{P}, s)\}$. This set must be nonempty, and $\forall s \in S^1$, $cP(s)\mu_S(\mathbf{P}, s)$. If for any $s \in S^1$, $sP(c)s_k$ then (c, s) would block $\mu_S(\mathbf{P})$. So $\forall s \in S^1$, $s_k P(c) s$. So for any $\hat{P}(c)$ which is responsive to $P(c)$ we have $\mu_S(\mathbf{P}, c)\hat{P}(c)\mu_S(\mathbf{P}', c)$. This completes the proof of the lemma.

We can now proceed to prove the theorem.

Proof of Theorem 1. Suppose $P'(c)$ is some preference ordering for college c. From Lemmas 3 and 4, it is clear that if the set of unacceptable students according to $P'(c)$ is not S'', then $P(c)d(P(c))P'(c)$. Also, from Lemma 1, the order in which the unacceptable students are ranked does not matter so long as S'' is a bottom set. So, we can assume w.l.o.g. that $S'' = \varnothing$.

Since truth-telling is a dominant strategy for all students under μ_S, it suffices to show that truthtelling is the unique protective strategy for all colleges.

Let $P'(c)$ be some preference ordering which is not equivalent to $P(c)$ and such that

1. For some integer r, $a_k(P(c)) = a_k(P'(c)) \forall k < r$.
2. $a_r(P(c)) \neq a_r(P'(c))$.

We want to show that $P(c)\, d(P(c))P'(c)$. Take any complementary profile P_{-c}. Suppose $\mu_S(P(c), P_{-c}, c) \cap \{a_1(P(c)), \ldots, a_r(P(c))\} \neq \emptyset$. Since $\{a_1(P(c)), \ldots, a_{r-1}(P(c))\}$ is also a bottom set for $P'(c)$, $\mu_S(P(c), P_{-c})$ must be individually rational for $P'(c)$. From Lemma 5, either $\mu_S(P(c), P_{-c}, c) = \mu_S(P'(c), P_{-c}, c)$ or $\mu_S(P(c), P_{-c}, c)\hat{P}(c)\mu_S(P'(c), P_{-c}, c)$. Hence, in order to show that $P(c)d(P(c))P'(c)$, it is sufficient to show that there is some P_{-c} such that $\mu_S(P(c), P_{-c}, c) = \mu_S(P'(c), P_{-c}, c) \backslash \{a_r(P(c))\} \cup \{a_j(P(c))\}$ for some $j > r$.

Note that since $P(c)$ and $P'(c)$ are not equivalent strategies, $|S| - (r-1) > q_c$. Let $a_r(P'(c)) = s_j$ and $a_r(P(c)) = s_r$, and S^* be some set of $(q_c - 1)$ students such that $\forall s \in S^*\, sP(c)s_r$. Consider a complementary profile P_{-c} such that

1. $\forall s \in S^* \cup \{s_j, s_r\}$, $cP(s)sP(s)c' \forall c' \in C \backslash \{c\}$.
2. $\forall s \notin (S^* \cup \{s_j, s_r\})$, c is not acceptable to s.

The reader can check that $\mu_S(P(c), P_{-c}, c) = \{s_j\} \cup S^*$ while $\mu_S(P'(c), P_{-c}, c) = \{s_r\} \cup S^*$. Since $s_j P(c) S_r$, this shows that $P(c)d(P(c))P'(c)$. This completes the proof of Theorem 1.

Notice that in the marriage model, both sets of agents can be identified with the students in the college admissions model. As a result we can state:

THEOREM 2. *In the marriage market, truthtelling is the unique protective strategy in the M-optimal and W-optimal stable matchings.*

We show below that a corresponding result is not true for the college-optimal stable matching. We remark in passing that the example below demonstrates that Lemma 5 does not hold for the college-optimal matching.

EXAMPLE 1.[5] Let $S = \{s_1, s_2, s_3, s_4\}$, $C = \{c_1, c_2, c_3, c_4\}$ with quotas $q_1 = 2$, $q_2 = 1$, $q_3 = 1$. Let college c_1 have the preference ordering

[5] This example is adapted from Roth (1985).

$$P(c_1) = s_1, s_2, s_3, s_4.$$

Consider the complementary profile P_{-c_1}

- $P(s_1) = c_3, c_1, c_2.$
- $P(s_2) = c_2, c_1, c_3.$
- $P(s_3) = c_1, c_3, c_2.$
- $P(s_4) = c_1, c_2, c_3.$
- $P(c_2) = s_1, s_2, s_3, s_4.$
- $P(c_3) = s_3, s_1, s_2, s_4.$

Then, $\mu_r(P(c_1), P_{-c_1}, c_1) = \{s_3, s_4\}$. Let $P'(c_1) = s_2, s_4, s_1, s_3$.

Clearly, $\mu_r(P'(c_1), P_{-c_1}, c_1) = \{s_2, s_4\}$ and $\{s_2, s_4\} \hat{P}(c_1) \{s_3, s_4\}$. Moreover, if $c_1 \in \mu_c(P'(c_1), \bar{P}_{-c_1}, c_1)$ (that is if college c_1 does not fill its quota q_{c_1}) for some $\bar{P}_{-c_1}$, then it can be checked that $c_1 \in \mu_c(P(c_1), \bar{P}_{-c_1}, c_1)$. This implies that $P(c_1)$ does not protectively dominate $P'(c_1)$.

4. Protective Behavior in the Assignment Game

In this section, we will analyze protective behavior in a matching model with money. More specifically, we consider the *assignment* game arising between a set P of m potential buyers of objects owned by the set Q of n potential sellers. Each seller owns and each buyer demands exactly one indivisible object. Seller j has a reservation price of r_j for the object owned by him, while buyer i's valuation of seller j's object is α_{ij}. If buyer i buys from seller j at price p_j, then i's utility is $u_i = \alpha_{ij} - p_j$, while j's utility is $v_j = p_j - r_j$. Of course, mutually beneficial trade is possible only if $\alpha_{ij} > r_j$.

Let α be the (mxn) matrix of α_{ij}'s and r be the n-vector of sellers' reservation prices. A *feasible assignment* for (P, Q, α, r) is a matrix $x = (x_{ij})$ satisfying (a) $\Sigma_{i \in P} x_{ij} \leq 1$, (b) $\Sigma_{j \in Q} x_{ij} \leq 1$, (c) $\forall i, j\ x_{ij} \in \{0, 1\}$.

The obvious interpretation is that if $x_{ij} = 1$, then buyer i is "matched" to seller j. Condition (b) imposes the restriction that i is matched to at most one seller, while condition (a) states that each seller can sell to at most one buyer.

This is a model of one-to-one matching between agents in P and Q. An *outcome* in this game is a triple $(u, v; x)$ where $u \in \mathcal{O}^m$, $v \in \mathcal{O}^n$ and x is an assignment. Note that in contrast to the ordinal matching models considered in Section 3, an individual agent's utility level is no longer restricted to discrete values.

A triple $(u, v; x)$ is *feasible* if x is a feasible assignment such that

$$\sum_{i\in P} u_i + \sum_{j\in Q} v_j = \sum_{\substack{i\in P\\ j\in Q}} (\alpha_{ij} - r_j)x_{ij}.$$

DEFINITION 7. A feasible outcome $(u, v; x)$ is *stable* if $\forall i \in P$, $\forall j \in Q$.

(i) $u_i \geq 0, v_j \geq 0$
(ii) $u_j + v_j \geq \max(0, \alpha_{ij} - r_j)$.

Shapley and Shubik (1972) proved that there is a P-optimal stable payoff $(\bar{u}, \underline{v})$ with the property that for any stable payoff (u, v), $\bar{u} \geq u$ and $\underline{v} \leq v$. Similarly, there is a Q-optimal stable payoff $(\underline{u}, \bar{v})$ with symmetrical properties. Moreover, Demange (1982) and Leonard (1983) proved that in the "direct revelation" game in which buyer i announces a vector α_i, and seller j announces r_j, truth-telling is the dominant strategy for buyers if the outcome function always selects the P-optimal payoff. Of course, revelation of the true reservation price is not in general the dominant strategy for sellers in this game.

Let μ_p denote the outcome function which selects the P-optimal stable payoff. We show below that revelation of the true reservation price is the unique protective strategy for sellers under μ_p.

Let $p_j(\alpha, r)$ denote the price obtained by seller j under μ_p when (α, r) are the announced vectors of buyers' valuations and sellers' reservation prices. By convention, we set $p_j(\alpha, r) = r_j^*$ if seller j does not sell his object at (α, r), where r_j^* is seller j's *true* reservation price.

THEOREM 3. *Under μ_p, truthtelling is the unique protective strategy for all agents in $P \cup Q$.*

Proof. Clearly, we only need to prove that truthtelling is the unique protective strategy for all sellers.

Choose any seller j, and let his true reservation price be r_j^*.

Step 1. Suppose j announces a reservation price of $r_j < r_j^*$. Then, r_j^* dominates r_j.

To see this, choose any (α, r_{-j}) Note that if $p_j(\alpha, r_{-j}, r_j) \geq r_j^*$ then $p_j(\alpha, r_{-j}, r_j^*) = p_j(\alpha, r_{-j}, r_j)$. Moreover, if $p_j(\alpha, r_{-j}, r_j) \in (r_j, r_j^*)$, then seller j is unmatched by μ_p if he announces r_j^*.

Hence, r_j^* dominates r_j.

Step 2. Suppose j announces a reservation price of $r_j > r_j^*$. Let $r_j = r_j^* + \varepsilon$. Then, let $r_j' = r_j^* + (\varepsilon/2)$. We show that r_j' protectively dominates r_j.

Consider any (α, r_{-j}) such that $p_j(\alpha, r_{-j}, r_j') = r_j^*$. Clearly, if j is unmatched at (α, r_{-j}, r_j'), then j is also unmatched at (α, r_{-j}, r_j) since $r_j > r_j'$. Hence, $p_j(\alpha, r_{-j}, r_j) = r_j^*$. Now consider (α, r_{-j}) such that:

(i) $\alpha_{ij} = r_j'$, $\alpha_{ik} < r_k$ for all $k \neq j$
(ii) $\alpha_{il} < r_j'$ for all $i \neq l$
(iii) $\alpha_{ik} = r_k$ for all $i \neq l$, for all $k \neq j$.

Clearly, $p_j(\alpha, r_{-j}, r_j') = r_j'$ while $p_j(\alpha, r_{-j}, r_j) = r_j^*$. Noting that for all (α, r_{-j}), $p_j(\alpha, r_{-j}, r_j') \geq r_j^*$ and $p_j(\alpha, r_{-j}, r_j) \geq r_j^*$, the above arguments show that r_j' protectively dominates r_j.

Step 3. r_j^* is protectively undominated.

In view of Step 1, we only need to show that if $r_j \succ r_j^*$, then r_j does not protectively dominate r_j^*.

We leave it to the reader to construct (α, r_{-j}) such that

(i) $p_j(\alpha, r_{-j}, r_j^*) \in (r_j^*, r_j)$ and (ii) $p_j(\alpha, r_{-j}, r_j) = r_j^*$.

Clearly, the existence of such a vector (α, r_{-j}) demonstrates that r_j^* is protectively undominated.

This concludes the Proof of Theorem 3.

Remark 1. Theorems 1 and 3 both show that truthtelling is the unique protective strategy in the respective matching models. Moreover, under the student optimal matching in the college admissions model, truth-telling protectively dominates all other (nonequivalent) strategies. However, truthtelling does not protectively dominate all other strategies for sellers under μ_p.

Remark 2. Theorem 3 raises the question whether truthtelling is the unique protective strategy in the mechanism which selects the Q-optimal stable payoff. Truth-telling is the dominant strategy for all sellers in this mechanism. Unfortunately, buyers do not have any protective strategy at all. To see this, consider the special case where there is only one seller. Then, in the seller-optimal matching the object is sold at the price announced by the highest bidder (or not sold at all if all bids are below the seller's reservation price). Hence, truthtelling can *never* yield any buyer a positive utility. So, truthtelling is dominated by any strategy which announces a lower bid. But, if α_i is the true valuation of bidder i, then a bid of $\overline{\alpha}_i = (\alpha_i - \varepsilon)$, is protectively dominated by the bid $(\alpha_i - \varepsilon/2)$.

References

Barberà, S., and Dutta, B. (1982). "Implementability via Protective Equilibria," *J. Math. Econ.* **10**, 49–65.

Barberà, S., and Jackson, M. (1988). "Maximin, Leximin and the Protective Criterion," *J. Econ. Theory* **46**, 34–44.

Barberà, S., and Pattanaik, P. K. (1984). "Extending an Order to the Power Set: Some Remarks on Kannai and Peleg's Approach," *J. Econ. Theory* **32**, 185–191.

DASGUPTA, P., HAMMOND, P., AND MASKIN, E. (1979). "The Implementation of Social Choice Rules: Some General Results on Incentive Compatibility," *Rev. Econ. Stud.* **46**, 185–216.

DEMANGE, G. (1982). "Strategyprogress in the Assignment Market Game," mimeo.

DRÈZE, J., AND DE LA VALLÉE POUSSIN, D. (1971). "A Tatonnement Process for Public Goods," *Rev. Econ. Stud.* **38**, 133–150.

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

KANNAI, Y., AND PELEG, B. (1984). "A Note on the Extension of an Order on a Set to the Power Set," *J. Econ. Theory* **32**, 192–196.

LEONARD, H. B. (1983). "Elicitation of Honest Preferences for the Assignment of Individuals to Positions," *J. Polit. Econ.* **91**, 461–479.

MOULIN, H. (1981). "Prudence versus Sophistication in Voting Strategy," *J. Econ. Theory* **23**, pp. 398–412.

ROTH, A. (1982). "The Economics of Matching: Stability and Incentives," *Math. Oper. Res.* **7**, 617–628.

ROTH, A. (1984). "Misrepresentation and Stability in the Marriage Problem," *J. Econ. Theory* **34**, 383–387.

ROTH, A. (1985). "The College Admissions Problem Is Not Equivalent to the Marriage Problem," *J. Econ. Theory* **36**, 277–288.

ROTH, A. (1989). "Two-Sided Matching with Incomplete Information about Other's preferences," *Games Econ. Behav.* **2**, 191–209.

ROTH, A. AND SOTOMAYOR, M. (1990). "Two-Sided Matching: A Study in Game-Theoretic Modelling and Analysis," London/New York: Cambridge University Press.

ROTH, A. AND VANDE VATE, J. H. (1989). "Incentives in Two-Sided Matching with Random Stable Mechanisms," mimeo.

SHAPLEY, L. S., AND SHUBIK, M. (1972). "The Assignment Game I: The Core," *Int. J. Game Theory* **1**, 111–130.

THOMSON, W. (1979). "Maximin Strategies and Elicitation of Preferences," in *Aggregation and Revelation of Preferences*, (J. J. Laffont, Ed.), pp. 245–268. Amsterdam: North-Holland.