Two Sided Matching with Endogenous Preferences

By YAIR ANTLER*

We modify the stable matching problem by allowing agents' preferences to depend on the endogenous actions of agents on the other side of the market. Conventional matching theory results break down in the modified setup. In particular, every game that is induced by a stable matching mechanism (e.g. the Gale-Shapley mechanism) may have equilibria that result in matchings which are not stable w.r.t the agents' endogenous preferences. However, when the Gale-Shapley mechanism is slightly modified, every equilibrium of its induced game results in a pairwise stable matching w.r.t the endogenous preferences as long as they satisfy a natural reciprocity property.
JEL: C72, C78, D82

The conventional mechanism design literature implicitly assumes that there is no link between the preference environment and the institutional setting. In many circumstances, this is a restrictive assumption since agents may be either embarrassed or admired as a result of the preferences, actions and opinions revealed by a mechanism. For example, consider the problem of eliciting agents' opinions: An agent might be embarrassed if it is stated publicly that his opinion differs from those of other agents. This may affect agents' considerations and the properties of the mechanism's result. In recent years, some researchers have relaxed this assumption and have allowed agents' preferences to be linked to the mechanism. Among them are Glazer and Rubinstein (1998), Bierbrauer and Netzer (2012) and Gradwohl (2013). The analysis presented here continues this line of research and focuses on two-sided matching.

We relax the above-mentioned assumption and generalize the one-to-one stable matching problem. In the conventional problem, there is a set of men and a set of women, each of whom has strict preferences defined over agents on the other side of the market. The agents' preferences are common knowledge among themselves but are unknown to a planner. The planner chooses a mechanism in order to match the two sides of the market such that the matching is pairwise stable (hereafter: stable) w.r.t the agents' preferences. A well-known result (see Roth (1984) and Gale and Sotomayor (1985)) is that the set of stable matchings is Nash equilibrium with undominated strategies (hereafter: equilibrium) implemented by the game that is induced by the Gale-Shapley mechanism (hereafter, GSM). We attempt to obtain an analogous result when agents' preferences are endogenous.

* Antler: The Eitan Berglas School of Economics, Tel Aviv University, Tel Aviv 69978, Israel. *E-mail:* yair.an@gmail.com. I am grateful to Rani Spiegler for his valuable guidance. I thank Eddie Dekel, Kfir Eliaz, Jacob Glazer, Ronen Gradwohl, Alvin Roth, Ariel Rubinstein, Tayfun Sonmez, Neil Thakral and an anonymous referee for helpful comments. I owe special thanks to Alex Frug for endless discussions.

To illustrate the model, consider the game induced by the GSM. In this game, agents' messages are stated rankings of agents on the other side of the market, which we interpret as public statements. For example, suppose that $Alice$, Bob and Dan are participating in this game. If $Alice$ submits a ranking in which Bob is ranked above Dan, then this is interpreted as a public statement that she likes Bob better than Dan. The desirability of an agent depends on his/her stated ranking. For example, in the case that Bob submits a ranking in which $Alice$ is his first choice, he might become more attractive in her eyes. Conversely, if he submits a ranking in which she is an unacceptable match (i.e. he states that he prefers being single over matching with her), it could make him less attractive to her.

In the above example, each agent's message space consists of rankings of agents on the other side of the market. Moreover, if agent i assigns a relatively high rank to agent j, this can be interpreted as a statement of j's desirability. Note that the same statement may be interpreted differently given different outcome functions or message spaces. In other words, our interpretation of a message in terms of the desirability of another agent depends on the matching mechanism.

At this point, it is worth emphasizing that the model presented below is one of complete information such that agents' preferences are common knowledge among the agents. The messages sent as part of the mechanism are known to the mechanism's participants and an audience of outside observers who see the messages as "signals" of the agents' true preferences. The existence of outside observers accounts in part for the agents' preferences though they are not modelled explicitly. The idea of outside observers who affect agents' considerations appears in Gradwohl (2013) in the context of privacy concerns. In a different context, an agent may have to justify his choice or preferences to an outside observer (see Spiegler (2002)). For example, suppose that $Alice$ and Bob are two agents taking part in some matching process and that the mechanism matches $Alice$ to Bob. Bob may need to justify his actions or his feelings towards $Alice$ to his parents and friends, which will be harder to do if $Alice$'s action is interpreted as a negative statement about Bob's desirability (for example, if she has ranked him relatively low).

Besides romantic relationships, there are other environments in which it is reasonable to assume that messages enter into agents' considerations in this manner. For example, in the labor market a worker who is ranked low by a firm may have a sense of inferiority among his colleagues who will realize that he was ranked low though they do not know the firm's true preferences. The parent of a child who was put on a waiting list before finally being admitted to a school may be upset by this and prefer sending him to another school. In professional American football, if a player declares in public that he does not want to play for a specific team this may affect the draft choice considerations of that team's manager.[1] Although the player's preferences are known to the manager (since teams invest a great deal of resources in scouting and observing players), they are unknown to the team's fans who are only aware of the player's declaration. Thus, the fans

[1]For example, there is the case of Eli Manning who was the most prominent player of the 2004 NFL draft class. Manning publically stated that he would refuse to play for the San Diego Chargers (who held the rights to the first overall draft pick).

may change their opinion of the player as a result of his declaration. It will be costly for the manager to explain to the fans why he would nonetheless like to select this player, which makes the player less attractive to him.

Since agents' preferences depend on the matching mechanism and are therefore endogenous, the notions of stability and equilibrium need to be clarified. Stability is treated as an ex post criterion since a planner would not want to give agents an incentive to challenge a prescribed match. Thus, we are interested in stability w.r.t the agents' preferences after the statements have been made. The standard notion of Nash equilibrium is used. A profile of strategies is an equilibrium if each agent's strategy is a best response (w.r.t the endogenous preferences that are generated by the profile of strategies) to the profile of strategies made by the other agents.

It is first shown that in the absence of restrictions on the structure of preferences, each game that is induced by an individually rational matching mechanism may have equilibria that result in unstable matchings (w.r.t the endogenous preferences that result from these equilibria). Restrictions are imposed on the formation of preferences such that if agent i's action expresses that he desires agent j, then as a result agent i will become more attractive to agent j. A general criterion that makes it possible to compare agents' actions in this context is provided. It turns out that even under these restrictions on the preferences formation, several matching theory results do not hold. In particular, each game that is induced by a stable matching mechanism or a monotonic priority mechanism may have equilibria that result in unstable matchings (w.r.t the agents' endogenous preferences).

We propose a new mechanism that is a variant of the GSM, in which each woman states the name of one man (and is also allowed to state that she prefers to remain unmatched), while each man announces a strict ranking of the women and the option of remaining unmatched. The Gale-Shapley algorithm with men making the proposals is used as an outcome function. A man who was not stated by a particular woman is treated by the algorithm as if that woman had stated that he is an unacceptable match for her. This mechanism makes it a weakly dominant strategy for each man to behave as if each of the women had declared that he is her only choice. Single deviations by women in the mechanism's induced game are aligned with pairwise deviations so that the stable matching problem with endogenous preferences is solved. In other words, each equilibrium of this mechanism's induced matching game results in a stable matching w.r.t the agents' endogenous preferences. Moreover, we show that in the conventional setup, in which agents' preferences are exogenous, this mechanism preserves the classical GSM's properties. It implements the stable correspondence and it is strategy-proof for agents on one side of the market.

The paper proceeds as follows: Section 1 introduces the general framework. Section 2 presents the analysis of the modified stable matching problem when there are no restrictions on the formation of preferences. In section 3, we impose a structure on the preferences formation and analyze the problem given this structure. Section 4 concludes. All proofs are to be found in the appendix.

I. The general framework

There are two disjoint sets $\mathcal{M}$ and $\mathcal{W}$, where $\mathcal{M} = \{m_1, ..., m_l\}$ is the set of men and $\mathcal{W} = \{w_1, ..., w_n\}$ is the set of women. A matching $\mu : \mathcal{M} \cup \mathcal{W} \to \mathcal{M} \cup \mathcal{W}$ is a one-to-one correspondence such that $w = \mu(m)$ if and only if $m = \mu(w)$ and for all m and w either $\mu(w) \in \mathcal{M}$ or $\mu(w) = w$ and either $\mu(m) \in \mathcal{W}$ or $\mu(m) = m$. We use the notation $\mu = \left((m, w), m', ...\right)$ to indicate that in the matching μ man m is married to woman w and man m' is single. Let $\mathcal{N}$ denote the set of matchings.

A matching mechanism $\langle \mathcal{Q}, g\rangle$ includes a message space $\mathcal{Q}_i$ for each agent $i \in \mathcal{M} \cup \mathcal{W}$ and an outcome function $g : \mathcal{Q} \to \mathcal{N}$, where $\mathcal{Q} := \times_{i \in \mathcal{M}\cup\mathcal{W}} \mathcal{Q}_i$. Let $Q_i \in \mathcal{Q}_i$ denote the message of agent $i \in \mathcal{M} \cup \mathcal{W}$ and let $\mathbf{Q} := (Q_i)_{i\in\mathcal{M}\cup\mathcal{W}}$, $\mathbf{Q}_{\mathcal{M}} := (Q_i)_{i\in\mathcal{M}}$, $\mathbf{Q}_{\mathcal{W}} := (Q_i)_{i\in\mathcal{W}}$, $\mathbf{Q}_{-i} := \left(Q_j\right)_{j\in\mathcal{M}\cup\mathcal{W}/\{i\}}$, $\mathcal{Q}_{\mathcal{M}} := \times_{i\in\mathcal{M}} \mathcal{Q}_i$, $\mathcal{Q}_{\mathcal{W}} := \times_{i\in\mathcal{W}} \mathcal{Q}_i$, and $\mathcal{Q}_{-i} := \times_{j\in\mathcal{M}\cup\mathcal{W}/\{i\}} \mathcal{Q}_j$. For each $i \in \mathcal{M} \cup \mathcal{W}$, denote agent i's spouse, given a matching mechanism $\langle \mathcal{Q}, g\rangle$ and a profile of messages $\mathbf{Q}$, by $g_i(\mathbf{Q})$. For each $m \in \mathcal{M}$ ($w \in \mathcal{W}$), define $\mathcal{L}_m$ ($\mathcal{L}_w$) to be the set of strict linear orderings of $\mathcal{W} \cup \{m\}$ ($\mathcal{M} \cup \{w\}$). We refer to a matching mechanism in which $\mathcal{Q}_i = \mathcal{L}_i$ for each $i \in \mathcal{M} \cup \mathcal{W}$ as a ranking mechanism and we say that agent i states that agent j is unacceptable to him/her if $i\, Q_i\, j$.

The novelty of the model is that agents' preferences now depend on the matching mechanism. Given a matching mechanism $\langle \mathcal{Q}, g\rangle$, each man's strict preferences $P_m^{\mathcal{Q},g} : \mathcal{Q}_{\mathcal{W}} \to \mathcal{L}_m$ depend on the profile of messages stated by the women and each woman's strict preferences $P_w^{\mathcal{Q},g} : \mathcal{Q}_{\mathcal{M}} \to \mathcal{L}_w$ depend on the profile of messages stated by the men. The matching mechanism remains fixed throughout the analysis and therefore we can omit the superscript $\mathcal{Q}, g$ from the preferences description. The following notation will be useful in what follows: When $\mathbf{Q} = (\mathbf{Q}_{\mathcal{W}}, \mathbf{Q}_{\mathcal{M}})$, we will often write $P_w(\mathbf{Q})$ instead of $P_w(\mathbf{Q}_{\mathcal{M}})$ and $P_m(\mathbf{Q})$ instead of $P_m(\mathbf{Q}_{\mathcal{W}})$. Let $\mathbf{P}(\mathbf{Q}) := (P_i(\mathbf{Q}))_{i\in\mathcal{M}\cup\mathcal{W}}$.

The preferences are defined w.r.t a matching mechanism and not w.r.t a mechanism's message space (e.g. rankings of agents on the other side of the market). This is for two reasons: First, different matching mechanisms may have different message spaces. Second, different outcome functions induce different interpretations of similar messages so that the message space is not sufficient for the formation of preferences. For example, consider two ranking mechanisms $\langle \mathcal{Q}, g\rangle$ and $\langle \mathcal{Q}, g'\rangle$ such that g (g') matches only couples who stated each other as their first (last) choice. A stated ranking in which $i \in \mathcal{M} \cup \mathcal{W}$ is the first choice is intuitively interpreted as a positive statement about i under g but not under g'. Following is the definition of stability in this model:

DEFINITION 1: *Given a mechanism $\langle \mathcal{Q}, g\rangle$ and a profile of messages $\mathbf{Q} \in \mathcal{Q}$, a matching $g(\mathbf{Q})$ is blocked by individual i if $i\, P_i(\mathbf{Q})\, g_i(\mathbf{Q})$. A matching $g(\mathbf{Q})$ is blocked by the pair (m, w) if $m \neq g_w(\mathbf{Q})$, $w\, P_m(\mathbf{Q})\, g_m(\mathbf{Q})$ and $m\, P_w(\mathbf{Q})\, g_w(\mathbf{Q})$. A matching is stable if it is not blocked by any individual or pair of agents.*

We use Nash equilibrium with weakly undominated strategies as a solution concept. Since agents' preferences are endogenous, the notion of weak domination differs and more than one preference relation is considered. Formally:

DEFINITION 2: *The strategy* Q_i *weakly dominates the strategy* $\hat{Q}_i$ *if there is no profile* $\mathbf{Q}'_{-i} \in \mathcal{Q}_{-i}$ *such that* $g_i\left(\hat{Q}_i, \mathbf{Q}'_{-i}\right) P_i \left(Q_i, \mathbf{Q}'_{-i}\right) g_i \left(Q_i, \mathbf{Q}'_{-i}\right)$ *and there exists a profile* $\mathbf{Q}'_{-i} \in \mathcal{Q}_{-i}$ *such that* $g_i\left(Q_i, \mathbf{Q}'_{-i}\right) P_i \left(Q_i, \mathbf{Q}'_{-i}\right) g_i \left(\hat{Q}_i, \mathbf{Q}'_{-i}\right)$.

DEFINITION 3: *An equilibrium of the game* $\langle \mathcal{Q}, g, \mathbf{P} \rangle$ *is a profile of strategies* $\left(Q_i^*\right)_{i \in M \cup W}$ *such that:*
(1) For each $i \in \mathcal{M} \cup \mathcal{W}$, Q_i^* *is not weakly dominated by any other strategy* $Q_i \in \mathcal{Q}_i$.
(2) There is no strategy $Q_i \in \mathcal{Q}_i$ *such that* $g_i\left(Q_i, \mathbf{Q}^*_{-i}\right) P_i \left(Q_i^*, \mathbf{Q}^*_{-i}\right) g_i \left(Q_i^*, \mathbf{Q}^*_{-i}\right)$.

II. Analysis of the modified stable matching problem

A well-known result presented by Roth (1984) states that when agents' preferences are exogenous, every equilibrium of the game induced by the GSM results in a stable matching. We attempt to obtain an analogous result when agents' preferences are endogenous. In other words, we look for a matching mechanism $\langle \mathcal{Q}, g \rangle$ such that for any profile of preferences $\mathbf{P}$, if $\mathbf{Q}^*$ is an equilibrium of the mechanism's induced game $\langle \mathcal{Q}, g, \mathbf{P} \rangle$, then $g\left(\mathbf{Q}^*\right)$ is stable w.r.t $\mathbf{P}\left(\mathbf{Q}^*\right)$. We refer to this problem as the modified stable matching problem. Subsequently, we present a short description of the GSM and a simple example which demonstrates that Roth's (1984) result does not hold when agents' preferences are endogenous.

The GSM is a ranking mechanism that uses the Gale-Shapley algorithm as an outcome function. The following description of the Gale-Shapley algorithm is taken from Roth and Sotomayor (1990). To start, each man proposes to the first woman on his preference list of acceptable women. Each woman rejects the proposal of any man who is unacceptable to her, and if she receives more than one proposal from acceptable men, she rejects all but the one she most prefers. Any man whose proposal has not been rejected at this point is kept "engaged". At any step any man whose proposal has been rejected in the previous step proposes to his next choice, as long as there remains an acceptable woman to whom he has not yet proposed. If at any step of the algorithm a man has already proposed to, and been rejected by, all of the women he finds acceptable, then he issues no further proposals. Each woman receiving proposals rejects any from unacceptable men, and also rejects all but her most preferred among the group consisting of the new proposers together with any man she may have kept engaged from the previous step. The algorithm stops and marriages are consummated after any step in which no man is rejected.

EXAMPLE 1: *Let* $\mathcal{M} = \{m\}$, $\mathcal{W} = \{w\}$ *and consider the GSM. Let* $P_w\left(Q_m\right) = Q_m$ *and* $P_m\left(Q_w\right) = Q_w$. *In other words, man m wants to be matched with woman w if and only if she states that she does not want to be matched with him* ($Q_w = (w, m)$) *and the same applies for woman w. The strategies* $Q_w^* = (w, m)$ *and* $Q_m^* = (m, w)$ *form a unique equilibrium of the game that is induced by the GSM regardless of which side of the market makes the proposals. To see this, note that by the individual rationality* *property of the mechanism neither agent can gain by a deviation and that in the only situation in which $i \in \{m, w\}$ can influence the resulting matching i is better off single. In the matching μ that results from this equilibrium, both agents are single but prefer being matched to each other and therefore (m, w) blocks μ.*

The preferences appearing in Example 1 are not intuitive but are nonetheless used because they keep the analysis straightforward and clearly illustrate two issues. First, they demonstrate the importance of imposing a structure on the formation of preferences. Second, they demonstrate the intuition behind the failure of individually rational matching mechanisms when agents' preferences are endogenous. We formalize this intuition in proposition 1.

DEFINITION 4: *An individually rational matching mechanism is a matching mechanism $\langle \mathcal{Q}, g\rangle$ such that for each $i \in \mathcal{M} \cup \mathcal{W}$ there exists a message $Q_i \in \mathcal{Q}_i$ such that $g_i\left(Q_i, \mathbf{Q}_{-i}\right) = i$ for each $\mathbf{Q}_{-i} \in \mathcal{Q}_{-i}$.*

PROPOSITION 1: *For each individually rational matching mechanism $\langle \mathcal{Q}, g\rangle$, there exists a profile of messages $\mathbf{Q}^*$ and a profile of preferences $\mathbf{P}$ such that $\mathbf{Q}^*$ is an equilibrium of the mechanism's induced game $\langle \mathcal{Q}, g, \mathbf{P}\rangle$ and $g\left(\mathbf{Q}^*\right)$ is not stable w.r.t $\mathbf{P}\left(\mathbf{Q}^*\right)$.*

Although example 1 implies that without any restriction on the preferences formation the GSM cannot solve the modified stable matching problem, it is worthwhile asking whether there are markets in which a planner can use this mechanism and be sure that each of the equilibria of its induced game results in a stable matching (w.r.t the endogenous preferences). In what follows, we show that in cases in which only one side of the market includes agents whose preferences depend on the institutional setting, the answer to this question is positive.

One-Sided Social Preferences

We focus on cases in which only one side of the market includes agents whose preferences are endogenous (and, without loss of generality, we assume that men's preferences are not sensitive to the mechanism). This structure describes a situation in which the two sides of the market are not symmetric. For example, if one side of the market consists of firms and the other of workers, then there are situations in which firms are not affected by the mechanism, but workers are. Prior to the analysis, we present a lemma that will be helpful in what follows. It involves a straightforward proof that a result similar to the one proven by Dubins and Freedman (1981) and Roth (1982) holds in our modified environment.

LEMMA 1: *Let $\langle \mathcal{Q}, g, \mathbf{P}\rangle$ denote the game induced by the GSM with men making the proposals and suppose that for a man $m \in \mathcal{M}$, $P_m\left(\mathbf{Q}\right) = P_m\left(\mathbf{Q}'\right) = \bar{P}_m$ for each $\mathbf{Q}, \mathbf{Q}' \in \mathcal{Q}$. Then, $Q_m = \bar{P}_m$ (i.e. truthful revelation of m's preferences) weakly dominates any other strategy $Q'_m \in \mathcal{Q}_m$.*

We use this lemma to prove a simple result which states that if the side of the market making the proposals is chosen carefully, then every matching that is a result of an equilibrium of the GSM's induced game is stable w.r.t the agents' endogenous preferences. This claim uses standard matching theory arguments.

PROPOSITION 2: *Let $\langle \mathcal{Q}, g, \mathbf{P} \rangle$ denote the game induced by the GSM with men making the proposals and suppose that for each $m \in \mathcal{M}$ and $\mathbf{Q}, \mathbf{Q}' \in \mathcal{Q}$, $P_m(\mathbf{Q}) = P_m(\mathbf{Q}') = P_m$. If $\mathbf{Q}^*$ is an equilibrium of $\langle \mathcal{Q}, g, \mathbf{P} \rangle$, then $g(\mathbf{Q}^*)$ is stable w.r.t $\mathbf{P}(\mathbf{Q}^*)$.*

To see the intuition behind the proof, note that according to lemma 1 each man truthfully reveals his preferences in each equilibrium. Since there is only one relevant profile of men's strategies, the market collapses to a conventional marriage market and the classic matching theory arguments are applicable. Note that if women are making the proposals, dominant strategies need not exist and the stability of matchings produced in equilibria of the game induced by the GSM is not guaranteed.

III. Positive reciprocity

Two restrictions are now imposed on the preferences formation. Broadly speaking, the first restriction allows us to capture the idea that if agent i's message expresses that he (she) desires agent j, then this makes him (her) more attractive to j. The second restriction implies that the relative ranking of any two agents depends only on their messages. Formally:

DEFINITION 5: *Let $i \in \mathcal{I}$, $\mathcal{J} \neq \mathcal{I}$ and $\mathcal{I}, \mathcal{J} \in \{\mathcal{M}, \mathcal{W}\}$. Given a matching mechanism*

$\langle \mathcal{Q}, g \rangle$, the message $Q_i \in \mathcal{Q}_i$ is at least as favorable to agent $j \in \mathcal{J}$ as the message $Q'_i \in \mathcal{Q}_i$ if for each $\mathbf{Q}_{-i} \in \mathcal{Q}_{-i}$, $g_i(Q'_i, \mathbf{Q}_{-i}) = j$ implies $g_i(Q_i, \mathbf{Q}_{-i}) = j$.

- A preferences profile $\mathbf{P}$ satisfies *monotonicity* if $i P_j (Q_i, \mathbf{Q}_{-i})\, i'$ implies $i P_j (Q'_i, \mathbf{Q}_{-i})\, i'$ whenever Q'_i is at least as favorable to j as Q_i for each $i \in \mathcal{I}$, $i' \in \mathcal{I} \cup \{j\}$, $j \in \mathcal{J} \neq \mathcal{I}$ and $\mathcal{I}, \mathcal{J} \in \{\mathcal{M}, \mathcal{W}\}$.
- A preferences profile $\mathbf{P}$ satisfies *independence* if the following two conditions are met for each $i \in I$, $j \in \mathcal{J} \neq I$ and $\mathcal{I}, \mathcal{J} \in \{\mathcal{M}, \mathcal{W}\}$:
  - For any two message profiles $\mathbf{Q}, \mathbf{Q}' \in \mathcal{Q}$ such that $Q_i = Q'_i$ and $Q_{i'} = Q'_{i'}$, $i P_j (\mathbf{Q})\, i'$ if and only if $i P_j (\mathbf{Q}')\, i'$.
  - For any two message profiles $\mathbf{Q}, \mathbf{Q}' \in \mathcal{Q}$ such that $Q_i = Q'_i$, $i P_j (\mathbf{Q})\, j$ if and only if $i P_j (\mathbf{Q}')\, j$.

Social psychologists have presented experimental evidence that broadly supports the assumptions we make regarding preferences formation. Aronson and Worchel (1966), Curtis and Kim (1986) and Condon and Crano (1988) presented subjects with false evidence that other subjects have indicated that they like them. As a result, the "liked" subjects favored these other subjects and gave them more positive evaluations.[2]

Below are some examples of matching mechanisms, their interpretation and the comparability of different messages according to favorability for each example. First, we consider the GSM, which is used in our benchmark result (Roth, 1984).

EXAMPLE 2: *Let $\langle \mathcal{Q}, g \rangle$ denote the GSM with men making the proposals and consider two possible messages of woman w, Q_w and Q'_w. Suppose that $m Q_w w$ and $m Q'_w w$. Then, the message Q_w is at least as favorable to m as Q'_w if and only if $\mathcal{B}_m \left(Q'_w\right) \subseteq \mathcal{B}_m \left(Q_w\right)$, where $\mathcal{B}_m \left(Q_w\right) = \left\{m' \in \mathcal{M} : m Q_w m'\right\}$. Suppose that $m Q_w w$ and $w Q'_w m$. Then, Q_w is clearly at least as favorable to m as Q'_w. Note that not all messages are comparable according to favorability. In particular, a message Q_w, which declares that m is w's xth choice, is not necessarily at least as favorable to m as a message Q'_w, which declares that m is w's yth choice for $1 < x < y$. To see this, suppose that there are two men m' and m'' such that m' (m'') is stated to be w's first (last) choice under Q_w and w's last (first) choice under Q'_w. The two messages are not comparable w.r.t each man $m \in \mathcal{M} / \left\{m', m''\right\}$ and therefore man m's preferences are not restricted by $monotonicity$ in this case. The men's messages are comparable in a symmetric manner.*

In the next example, we consider a bidding mechanism in which each man has a budget that he allocates among the women. Each woman states whether she is willing to be matched. One can interpret the men as students who are bidding for seminar participation and the women as professors who are to decide whether to give the seminar or not. In this context, a professor might be either embarrassed or admired if the sum of the bids for his course is publicly revealed.

EXAMPLE 3: *In this bidding mechanism, each $m \in \mathcal{M}$ assigns a real number $x_{mw} \geq 0$ to each woman $w \in \mathcal{W}$ such that $\sum_{w \in W} x_{mw} = 1$. Each woman $w \in \mathcal{W}$ declares her decision $Q_w \in \{yes, no\}$. Let $\mathcal{W}^1 = \{w \in \mathcal{W} : Q_w = yes\}$, $\mathcal{M}^1 = \mathcal{M}$. At each step $k = 1, 2, ...$ the mechanism's outcome function matches one couple by choosing the highest bid x_{mw} made by a man $m \in \mathcal{M}^k$ for a woman $w \in \mathcal{W}^k$ (with ties broken by the agents' indices), matches w with m and removes them from the lists of available agents, such that $\mathcal{M}^{k+1} = \mathcal{M}^k / \{m\}$ and $\mathcal{W}^{k+1} = \mathcal{W}^k / \{w\}$. The outcome function stops at step h, at which $\mathcal{W}^h = \emptyset$ or $\mathcal{M}^h = \emptyset$.*
When $|\mathcal{W}| \geq 3$, not all of the men's strategies are comparable according to favorability. For example, suppose that $\mathcal{W} = \{w_1, w_2, w_3\}$. The bidding strategy $(0, 0.45, 0.55)$ is (not) at least as favorable to w_3 (w_2) as the bidding strategy $(0.1, 0.4, 0.5)$. As for the women's strategies, for each $m \in \mathcal{M}$, yes is at least as favorable for m as no. The model's behavioral assumption implies that if a professor states that she wants to give a seminar, it could make her seminar more attractive to students, while the opposite is impossible.

[2]A caveat in the use of these studies as evidence in our context is that the effect of signaling cannot be isolated from the intrinsic reciprocity.

The next example involves a serial dictatorship mechanism which illustrates that "cheap talk" has no influence on the formation of preferences according to our criterion.

EXAMPLE 4: *The serial dictatorship mechanism $\langle \mathcal{Q}, g \rangle$ operates in stages. At each stage $k = 1, 2, ...$ the outcome function assigns woman w_k her first-ranked man among the men who are available at stage k. In this mechanism, each pair of men's strategies Q_m and Q'_m are degenerately comparable by favorability since they do not have any strategic meaning. It follows that the women's profile of preferences is constant under this mechanism.*

Since we have imposed a limiting structure on the formation of preferences, it is appropriate to check whether Roth's (1984) result holds under these restrictions. Example 5 shows that it does not.

EXAMPLE 5: *Consider the GSM with men making the proposals and let $\mathcal{M} = \{m_1, m_2, m_3\}$, $\mathcal{W} = \{w_1, w_2, w_3\}$. Let $P_{m_1} = (w_1, w_2, w_3, m_1)$, $P_{w_1} = (m_1, m_2, m_3, w_1)$,*

$$P_{m_2}(\mathbf{Q}) = \begin{pmatrix} w_1, w_2, w_3, m_2 & \text{if } Q_{w_2} = (m_2, ...) \\ w_1, w_3, w_2, m_2 & \text{otherwise} \end{pmatrix},$$

$$P_{m_3}(\mathbf{Q}) = \begin{pmatrix} w_1, w_3, w_2, m_3 & \text{if } Q_{w_3} = (m_3, ...) \\ w_1, w_2, w_3, m_3 & \text{otherwise} \end{pmatrix},$$

$$P_{w_2}(\mathbf{Q}) = \begin{pmatrix} m_1, m_2, m_3, w_2 & \text{if } Q_{m_2} = (w_2, ...) \\ m_1, m_3, m_2, w_2 & \text{otherwise} \end{pmatrix},$$

$$P_{w_3}(\mathbf{Q}) = \begin{pmatrix} m_1, m_3, m_2, w_3 & \text{if } Q_{m_3} = (w_3, ...) \\ m_1, m_2, m_3, w_3 & \text{otherwise} \end{pmatrix}.$$

According to this profile, there is a "consensus" among the agents regarding the two agents (m_1, w_1) and each $i \in \{m_2, m_3, w_2, w_3\}$ gives a "bonus" to one of the agents for publicly ranking him first. This profile satisfies $monotonicity$ and $independence$. The following profile of strategies $\mathbf{Q}^*$ forms an equilibrium of the mechanism's induced game: $Q^*_{m_1} = (w_1, w_2, w_3, m_1)$, $Q^*_{m_2} = (w_1, w_2, w_3, m_2)$, $Q^*_{m_3} = (w_1, w_3, w_2, m_3)$, $Q^*_{w_1} = (m_1, m_2, m_3, w_1)$, $Q^*_{w_2} = (m_1, m_2, m_3, w_2)$, $Q^*_{w_3} = (m_1, m_3, m_2, w_3)$. The matching resulting from $\mathbf{Q}^*$ is $g(\mathbf{Q}^*) = ((m_1, w_1), (m_2, w_2), (m_3, w_3))$, which is blocked by (m_2, w_3) and (m_3, w_2). To see that there are no profitable deviations from $\mathbf{Q}^*$, note that by the stability property of the mechanism no agent can do better. None of the strategies in use in $\mathbf{Q}^*$ is dominated by an argument similar to the one used in proposition 3.

The GSM allows m_2, m_3, w_2, w_3 to coordinate on an unstable Nash equilibrium. In a world with no motives related to the matching process, the exclusion of dominated strategies implies that men state their true preferences (see Dubins and Freedman (1981) and Roth (1982)), which rules out these unstable equilibria. The motives in the model presented here eliminate the men's dominant strategies and enable coordination on such unstable equilibria.

The following result states not only that the GSM fails to solve the modified stable matching problem, but also that no other stable matching mechanism can do so. Although they are not strategy-proof, it is beneficial to explore this class of mechanisms since it has received a great deal of attention, both in matching theory and in practice.[3] In order to prove proposition 3, we introduce a profile of preferences and a profile of strategies inspired by example 5, such that the profile of strategies is an equilibrium of any stable matching mechanism's induced game and results in a matching that is not stable w.r.t the endogenous preferences induced by it.

DEFINITION 6: *A stable matching mechanism is a ranking mechanism* $\langle \mathcal{Q}, g \rangle$ *such that for each* $\mathbf{Q} \in \mathcal{Q}$*,* $g(\mathbf{Q})$ *is stable w.r.t* $\mathbf{Q}$*.*

PROPOSITION 3: *Let* $|\mathcal{M}|, |\mathcal{W}| \geq 3$*. For each stable matching mechanism* $\langle \mathcal{Q}, g \rangle$*, there exists a profile of strategies* $\mathbf{Q}^*$ *and a profile of preferences* $\mathbf{P}$ *that satisfies monotonicity and independence such that* $\mathbf{Q}^*$ *is an equilibrium of the mechanism's induced game* $\langle \mathcal{Q}, g, \mathbf{P} \rangle$ *and* $g(\mathbf{Q}^*)$ *is not stable w.r.t* $\mathbf{P}(\mathbf{Q}^*)$*.*

Ergin and Sonmez (2006) provide a result which is similar to that of Roth (1984). They show that although the Boston mechanism, which is a monotonic priority matching mechanism, is not a stable matching mechanism, the game induced by it Nash equilibrium with undominated strategies implements the stable correspondence when agents' preferences are exogenous. The following proposition states that when agents' preferences are endogenous, a matching that results from an equilibrium of a game that is induced by a monotonic priority matching mechanism need not be stable even if $monotonicity$ and $independence$ are satisfied by the agents' preferences.

A priority matching mechanism is a ranking mechanism. We adopt Ergin and Sonmez's (2006) framework and define a (k, l) match to be one between a man and a woman such that the man submits a ranking in which the woman is ranked kth and the woman submits a ranking in which the man is ranked lth. Given a marriage market with n men and m women, a match priority is a one-to-one function $\pi : \{1, ..., n\} \times \{1, ..., m\} \to \{1, ..., nm\}$. A priority matching mechanism determines its outcome in nm steps:

- Step 1: Form any feasible and acceptable $\pi^{-1}(1)$ match.
- Step 2: Form any feasible and acceptable $\pi^{-1}(2)$ match.
- $\vdots$
- Step nm: Form any feasible and acceptable $\pi^{-1}(nm)$ match.

A match priority π is monotonic if $(k, l) \leq (k', l')$ implies $\pi(k, l) \leq \pi(k', l')$. A priority matching mechanism is monotonic if it is induced by a monotonic match priority.

[3] Stable matching mechanisms are not strategy-proof even in the conventional setting.

PROPOSITION 4: *Let* $|\mathcal{M}|, |\mathcal{W}| \geq 3$*. For each monotonic priority matching mechanism* $\langle \mathcal{Q}, g \rangle$*, there exists a profile of strategies* $\mathbf{Q}^*$ *and a profile of preferences* $\mathbf{P}$ *that satisfies monotonicity and independence such that* $\mathbf{Q}^*$ *is an equilibrium of the mechanism's induced game* $\langle \mathcal{Q}, g, \mathbf{P} \rangle$ *and* $g(\mathbf{Q}^*)$ *is not stable w.r.t* $\mathbf{P}(\mathbf{Q}^*)$*.*

The intuition behind Ergin and Sonmez's implementation result is that for one side of the market truthful revelation is a dominant strategy, so that in equilibrium agents on that side of the market must use this dominant strategy. The existence of matching-process-related motives eliminates dominant strategies and allows agents to coordinate on an equilibrium that results in an unstable matching.

The modified Gale-Shapley mechanism

In this subsection, we solve the modified stable matching problem using a modified GSM which we denote by $\langle \bar{\mathcal{Q}}, \bar{g} \rangle$. The mechanism is specified as follows: For each $m \in \mathcal{M}$, let $\bar{\mathcal{Q}}_m = \mathcal{L}_m$ and for each $w \in \mathcal{W}$ let $\bar{\mathcal{Q}}_w = \mathcal{M} \cup \{w\}$. In other words, in the modified GSM each man states a strict ranking of the women and the option of remaining unmatched and each woman states at most one man who is an acceptable match for her. Let $\bar{g}$ be the Gale-Shapley algorithm (with men making the proposals) such that for each $w \in \mathcal{W}$ the algorithm considers only $Q_w \in \bar{\mathcal{Q}}_w$ as an acceptable match for w. Note that the modified GSM is not a stable matching mechanism since it only considers the women's stated first choices.

PROPOSITION 5: *Suppose that* $\mathbf{Q}^*$ *is an equilibrium of the game that is induced by the modified GSM* $\langle \bar{\mathcal{Q}}, \bar{g}, \mathbf{P} \rangle$*. If* $\mathbf{P}$ *satisfies* $monotonicity$ *and* $independence$*, then* $\bar{g}(\mathbf{Q}^*)$ *is stable w.r.t* $\mathbf{P}(\mathbf{Q}^*)$*.*

It turns out that the modified GSM solves the stable matching problem when agents' preferences are endogenous. Although at first glance it would appear that the message space of ranking mechanisms is not sufficiently rich to express the agents' endogenous preferences which results in the failure of conventional matching mechanisms, imposing additional restrictions on the message space does solve the modified stable matching problem. These restrictions prevent women from making statements that are not necessary for the matching process. In addition, they make it a dominant strategy for men to behave as in the most optimistic scenario, in which they are every woman's stated first choice.

In a similar manner, whenever $monotonicity$ and $independence$ are satisfied by agents on both sides of the market, one can solve the modified stable matching problem with a mechanism that restricts the message space of men and uses the Gale-Shapley algorithm with women making the proposals.[4] Since the proof of proposition 5 uses $monotonicity$ and $independence$ only w.r.t the men's preferences, it follows that if a

[4]The modified GSM is not symmetric. Since agents' preferences are mechanism-dependent, the set of matchings that can be supported by an equilibrium depends on the side of the market which makes the proposals.

planner has reason to suspect that *monotonicity* and *independence* are not satisfied by the preferences of agents on one side of the market, then he can still choose the side of the market making the proposals such that the modified GSM will produce only stable matchings.

It is of interest to evaluate the performance of the modified GSM in the conventional setup, in which agents' preferences do not depend on the mechanism. Specifically, the next proposition, together with proposition 5, shows that under the conventional settings the modified GSM preserves the classic properties of the GSM.

PROPOSITION 6: *Consider the modified GSM and suppose that for each* $i \in \mathcal{M} \cup \mathcal{W}$ *and* $\mathbf{Q}, \mathbf{Q}' \in \mathcal{Q}$, $P_i(\mathbf{Q}) = P_i(\mathbf{Q}') = P_i'$. *Let* μ *be a matching that is stable w.r.t* $\mathbf{P}'$. *There exists a profile of strategies* $\mathbf{Q}^*$ *such that* $\mathbf{Q}^*$ *is an equilibrium of* $\langle \bar{\mathcal{Q}}, \bar{g}, \mathbf{P}' \rangle$ *and* $\bar{g}(\mathbf{Q}^*) = \mu$.

Since the conventional environment is a special case of this model, proposition 5 establishes that each matching which is the result of an equilibrium of the modified GSM's induced game is stable w.r.t the agents' preferences and that the modified GSM is strategy-proof for the men. By proposition 6, the modified GSM implements the stable correspondence and therefore all of the GSM's classic properties are preserved by its modification.

IV. Concluding remarks

The model presented is the first attempt to incorporate behavioral motives into matching theory. In recent years, matching theory has became increasingly applied in practice. The theoretical analysis presented here examines the mechanisms and concepts that have been the focus of this literature for the past thirty years. It is essentially a generalization of the two-sided matching problem and a robustness check for the mechanisms that are used in practice. The results of the model's analysis imply that privacy implications need to be taken into consideration. It appears that stable matching mechanisms perform well when information about the matching process is kept private. There are various settings in which one might want the matching process to be transparent; however, in that case, conventional mechanisms will fail.

Even if the matching process is kept confidential, the sensitivity of preferences to the mechanism may be problematic since some of the messages may be deduced from the mechanism's outcomes. For example, a student who was put on a waiting list knows that he is not top-ranked. An intern who is assigned to a hospital that he ranked fifth will likely understand that he was not the first choice of his first choice. Moreover, it seems plausible that some of the rankings will leak out of the system. For example, an intern may find out the actual rankings of the hospital he was assigned to.

Another important theoretical remark concerns the revelation principle. The matching mechanisms considered in this work are not direct mechanisms.[5] In this paper's environment, an agent i's "type" is a full description of the function P_i. When agent's

[5] Conventional matching mechanisms are not direct mechanisms either.

preferences depend on the mechanism, the function P_i changes when the mechanism is changed and therefore the revelation principle is not well-defined and cannot be used.

REFERENCES

Aronson, Elliot and Worchel, Philip (1966): "Similarity versus liking as determinants of interpersonal attractiveness," Psychonomic Science, Vol 5(4), 157-158.

Bierbrauer, Felix and Netzer, Nick (2012): "Mechanism design and intentions," University of Zurich Working Paper.

Condon, John W. and Crano, William D. (1988): "Inferred evaluation and the relation between attitude similarity and interpersonal attraction," Journal of Personality and Social Psychology, 54.5: 789.

Curtis, Rebecca C. and Miller, Kim (1986): "Believing another likes or dislikes you: Behaviors making the beliefs come true," Journal of Personality and Social Psychology, 51(2), 284-290.

Dubins, Lester and Freedman, David A. (1981): "Machiavelli and the Gale-Shapley algorithm," American Mathematical Monthly, 88, 485-94.

Ergin, Haluk and Sonmez, Tayfun (2006): "Games of school choice under the Boston mechanism,"Journal of Public Economics, Elsevier, vol. 90(1-2), pages 215-237, January.

Gale, David and Shapley, Lloyd (1962): "College admissions and the stability of marriage," American Mathematical Monthly, 69, 9–14.

Gale, David and Sotomayor, Marilda (1985): "Ms Machiavelli and the stable matching problem," American Mathematical Monthly, 92, 261-8.

Glazer, Jacob and Rubinstein, Ariel (1998): "Motives and implementation: On the design of mechanisms to elicit opinions," Journal of Economic Theory, Elsevier, vol. 79(2), pages 157-173, April.

Gradwohl, Ronen (2013): "Privacy in implementation," Working paper.

Roth, Alvin E. (1982): "The economics of matching: stability and incentives," Mathematics of Operations Research, 7, 617-28.

Roth, Alvin E. (1984): "Misrepresentation and stability in the marriage problem," Journal of Economic Theory, Elsevier, vol. 34(2), pages 383-387, December.

Roth, Alvin E. and Sotomayor, Marilda (1990):"Two-sided matching: A study in game theoretic modeling and analysis," Cambridge Univ. Press, London/New York.

Spiegler, Ran (2002): "Equilibrium in justifiable strategies: A model of reason-based choice in extensive-form games," Review of Economic Studies, Wiley Blackwell, vol. 69(3), pages 691-706, July.

APPENDIX - PROOFS

PROPOSITION 1: *For each individually rational matching mechanism* $\langle \mathcal{Q}, g \rangle$*, there exists a profile of messages* $\mathbf{Q}^*$ *and a profile of preferences* $\mathbf{P}$ *such that* $\mathbf{Q}^*$ *is an equilibrium of the mechanism's induced game* $\langle \mathcal{Q}, g, \mathbf{P} \rangle$ *and* $g(\mathbf{Q}^*)$ *is not stable w.r.t* $\mathbf{P}(\mathbf{Q}^*)$*.*

PROOF:

Let $\langle \mathcal{Q}, g \rangle$ be an individually rational matching mechanism and let $m' \in \mathcal{M}$, $w' \in \mathcal{W}$. For each $i \in \mathcal{M} \cup \mathcal{W}$ let $Q_i^* \in \mathcal{Q}_i$ denote a message such that $g_i\left(Q_i^*, \mathbf{Q}_{-i}\right) = i$ for each $\mathbf{Q}_{-i} \in \mathcal{Q}_{-i}$. Consider a profile of preferences $\mathbf{P}$ in which $P_{m'}(\mathbf{Q}) = \left(w', m', ...\right)$ if $Q_{w'} = Q_{w'}^*$ and $P_{m'}(\mathbf{Q}) = \left(m', ...\right)$ otherwise, and $P_{w'}(\mathbf{Q}) = \left(m', w', ...\right)$ if $Q_{m'} = Q_{m'}^*$ and $P_{w'}(\mathbf{Q}) = \left(w', ...\right)$ otherwise. For each $j \in \mathcal{M} \cup \mathcal{W} / \left\{m', w'\right\}$ and $\mathbf{Q} \in \mathcal{Q}$ let $P_j(\mathbf{Q}) = (j, ...)$. Consider the profile of messages $\left(Q_i^*\right)_{i \in M \cup W}$. First, note that $g(\mathbf{Q}^*) = \left(m', w', ...\right)$ and that the pair $\left(m', w'\right)$ blocks $g(\mathbf{Q}^*)$ w.r.t $\mathbf{P}(\mathbf{Q}^*)$. Secondly, by the individual rationality property, for each $i \in \mathcal{M} \cup \mathcal{W}$, $\mathbf{Q}_i^*$ is a best response to any profile $\mathbf{Q}_{-i} \in \mathcal{Q}_{-i}$ and therefore it is not weakly dominated and $\mathbf{Q}^*$ is an equilibrium of the mechanism's induced game $\langle \mathcal{Q}, g, \mathbf{P} \rangle$.

LEMMA 1: *Let* $\langle \mathcal{Q}, g, \mathbf{P} \rangle$ *denote the game induced by the GSM with men making the proposals and suppose that for a man* $m \in M$*,* $P_m(\mathbf{Q}) = P_m\left(\mathbf{Q}'\right) = \bar{P}_m$ *for each* $\mathbf{Q}, \mathbf{Q}' \in \mathcal{Q}$*. Then,* $Q_m = \bar{P}_m$ *(i.e. truthful revelation of m's preferences) weakly dominates any other strategy* $Q'_m \in \mathcal{Q}_m$*.*

PROOF:

Let $\langle \mathcal{Q}, g, \mathbf{P} \rangle$ denote the game induced by the GSM with men making the proposals and suppose that for a man $m \in \mathcal{M}$, $P_m(\mathbf{Q}) = P_m\left(\mathbf{Q}'\right) = \bar{P}_m$ for each $\mathbf{Q}, \mathbf{Q}' \in \mathcal{Q}$. Let $\mathbf{Q}_{-m} \in \mathcal{Q}_{-m}$ be an arbitrary profile of strategies and consider a strategy $Q_m \neq \bar{P}_m$. Assume by negation that $g_m(Q_m, \mathbf{Q}_{-m})\, \bar{P}_m g_m\left(\bar{P}_m, \mathbf{Q}_{-m}\right)$. Suppose that $g_m(Q_m, \mathbf{Q}_{-m}) \in \mathcal{W}$. By the algorithm, if m reports $\bar{P}_m$, then he proposes to $g_m(Q_m, \mathbf{Q}_{-m})$ and is rejected by her at some stage, a contradiction. Suppose that $g_m(Q_m, \mathbf{Q}_{-m}) = m$. This is a contradiction since if m reports $\bar{P}_m$, he does not propose to women who are an unacceptable match for him.

It is left to show that there exists a profile $\mathbf{Q}_{-m} \in \mathcal{Q}_{-m}$ such that $g_m\left(\bar{P}_m, \mathbf{Q}_{-m}\right) \bar{P}_m g_m$ $(Q_m, \mathbf{Q}_{-m})$. Since $Q_m \neq \bar{P}_m$, there are three possible cases. In the first, there are two women $w, w' \in \mathcal{W}$ such that $w Q_m w'$, $w' \bar{P}_m w$ and $w \bar{P}_m m$. In that case, the strategy $\bar{P}_m$ does strictly better than the strategy Q_m against a profile of strategies $\mathbf{Q}'_{-m}$ such that $Q'_i = (m, i, ...)$ for $i \in \left\{w, w'\right\}$ and $Q'_j = (j, ...)$ for each $j \in \mathcal{W} / \left\{w, w'\right\}$. In the two complementary cases, there is a woman $w \in \mathcal{W}$ such that $m \bar{P}_m w$ $(w \bar{P}_m m)$ and $w Q_m m$ $(m Q_m w)$. In those cases, the strategy $\bar{P}_m$ does strictly better than the strategy Q_m against a profile of strategies $\mathbf{Q}'_{-m}$ such that $Q'_w = \left(m, w', ...\right)$ and $Q'_j = (j, ...)$ for each $j \in \mathcal{M} \cup \mathcal{W} / \{m, w\}$. It follows that Q_m is weakly dominated by $\bar{P}_m$.

PROPOSITION 2: *Let $\langle \mathcal{Q}, g, \mathbf{P} \rangle$ denote the game induced by the GSM with men making the proposals and suppose that for each $m \in M$ and $\mathbf{Q}, \mathbf{Q}' \in \mathcal{Q}$, $P_m(\mathbf{Q}) = P_m(\mathbf{Q}') = P_m$. If $\mathbf{Q}^*$ is an equilibrium of $\langle \mathcal{Q}, g, \mathbf{P} \rangle$, then $g(\mathbf{Q}^*)$ is stable w.r.t $\mathbf{P}(\mathbf{Q}^*)$.*

PROOF:

Let $\langle \mathcal{Q}, g, \mathbf{P} \rangle$ denote the game induced by the GSM with men making the proposals and suppose that for each $m \in \mathcal{M}$ and $\mathbf{Q}, \mathbf{Q}' \in \mathcal{Q}$, $P_m(\mathbf{Q}) = P_m(\mathbf{Q}') = P_m$. Assume by negation that $\mathbf{Q}^*$ is an equilibrium of $\langle \mathcal{Q}, g, \mathbf{P} \rangle$ and that $g(\mathbf{Q}^*)$ *is not stable w.r.t* $\mathbf{P}(\mathbf{Q}^*)$. It cannot be that $g(\mathbf{Q}^*)$ is blocked by any individual $i \in \mathcal{M} \cup \mathcal{W}$ since in that case, i could profitably deviate to $Q_i' = (i, ...)$. It must be that $g(\mathbf{Q}^*)$ is blocked by a pair (m, w). Since $\mathbf{Q}^*$ is an equilibrium of $\langle \mathcal{Q}, g, \mathbf{P} \rangle$, by lemma 1, $Q_m = P_m$. Since (m, w) form a blocking pair, $w P_m g_m(\mathbf{Q}^*)$ and therefore w can submit a ranking $Q_w' = (m, w, ...)$ and marry m. Since (m, w) form a blocking pair w.r.t $\mathbf{P}(\mathbf{Q}^*)$, it must be that $m P_w(\mathbf{Q}^*) g_w(\mathbf{Q}^*)$ which is a contradiction of $\mathbf{Q}^*$ being an equilibrium of $\langle \mathcal{Q}, g, \mathbf{P} \rangle$.

PROPOSITION 3: *Let $|\mathcal{M}|, |\mathcal{W}| \geq 3$. For each stable matching mechanism $\langle \mathcal{Q}, g \rangle$, there exists a profile of strategies $\mathbf{Q}^*$ and a profile of preferences P that satisfies monotonicity and independence such that $\mathbf{Q}^*$ is an equilibrium of the mechanism's induced game $\langle \mathcal{Q}, g, \mathbf{P} \rangle$ and $g(\mathbf{Q}^*)$ is not stable w.r.t $\mathbf{P}(\mathbf{Q}^*)$.*

PROOF:

This claim is proved by embedding a profile of preferences that is similar to the one presented in example 5 within a larger market. Let $\langle \mathcal{Q}, g \rangle$ be a stable matching mechanism and consider six agents $\{m_1, m_2, m_3, w_1, w_2, w_3\}$ such that $P_{m_1} = (w_1, m_1, ...)$, $P_{w_1} = (m_1, w_1, ...)$,

$$P_{m_2}(\mathbf{Q}) = \begin{pmatrix} w_1, w_2, w_3, m_2, ... & \text{if } Q_{w_2} = (m_2, ...) \\ w_1, w_3, w_2, m_2, ... & \text{otherwise} \end{pmatrix},$$

$$P_{m_3}(\mathbf{Q}) = \begin{pmatrix} w_1, w_3, w_2, m_3, ... & \text{if } Q_{w_3} = (m_3, ...) \\ w_1, w_2, w_3, m_3, ... & \text{otherwise} \end{pmatrix},$$

$$P_{w_2}(\mathbf{Q}) = \begin{pmatrix} m_1, m_2, m_3, w_2, ... & \text{if } Q_{m_2} = (w_2, ...) \\ m_1, m_3, m_2, w_2, ... & \text{otherwise} \end{pmatrix},$$

$$P_{w_3}(\mathbf{Q}) = \begin{pmatrix} m_1, m_3, m_2, w_3, ... & \text{if } Q_{m_3} = (w_3, ...) \\ m_1, m_2, m_3, w_3, ... & \text{otherwise} \end{pmatrix}.$$

Suppose that for each $i \in \mathcal{M} \cup \mathcal{W} / \{m_1, m_2, m_3, w_1, w_2, w_3\}$ and $\mathbf{Q} \in \mathcal{Q}$, $P_i(\mathbf{Q}) = (i, ...)$. Clearly, $independence$ is satisfied by the profile of preferences $\mathbf{P}$. To see that $monotonicity$ is also satisfied, we only need to check agents whose preferences are not constant. Since there is symmetry between m_2, m_3, w_2 and w_3, we only present the argument for man m_2. The only case in which m_2 reverses his preference between two women is a case in which woman w_2 switches between two strategies: a strategy Q_{w_2} which states that m_2 is her first choice and another strategy Q'_{w_2} which does not state that. To see that the strategy Q'_{w_2} is not at least as favorable to m_2 as the strategy Q_{w_2}, consider a profile of strategies $\mathbf{Q}_{-w_2}$ in which w_2 is the first choice of each $m \in \mathcal{M}$. By the stability of $\langle \mathcal{Q}, g\rangle$, Q_{w_2} matches w_2 and m_2 while Q'_{w_2} does not. It follows that $\mathbf{P}$ satisfies *monotonicity*.

Consider the following profile of strategies: $Q^*_{m_1} = (w_1, m_1, ...)$, $Q^*_{m_2} = (w_1, w_2, w_3, m_2, ...)$, $Q^*_{m_3} = (w_1, w_3, w_2, m_3, ...)$, $Q^*_{w_1} = (m_1, w_1, ...)$, $Q^*_{w_2} = (m_1, m_2, m_3, w_2, ...)$, $Q^*_{w_3} = (m_1, m_3, m_2, w_3, ...)$, and $Q^*_i = (i, ...)$ for each $i \in \mathcal{M} \cup \mathcal{W} / \{m_1, m_2, m_3, w_1, w_2, w_3\}$. By the stability of $\langle \mathcal{Q}, g\rangle$, $g(\mathbf{Q}^*) = ((m_1, w_1), (m_2, w_2), (m_3, w_3), ...)$. Note that agent $k \in \mathcal{M} \cup \mathcal{W} / \{m_2, m_3, w_2, w_3\}$ is matched to his first choice and therefore cannot gain by a deviation. By the stability of $\langle \mathcal{Q}, g\rangle$ and given $\mathbf{Q}^*_{-j}$, each agent $j \in \{m_2, m_3, w_2, w_3\}$ cannot deviate and marry another agent. It follows that $\mathbf{Q}^*$ is a Nash equilibrium of $\langle \mathcal{Q}, g, \mathbf{P}\rangle$.

It is left to show that each $Q_i \in \mathbf{Q}^*$ is not dominated. For each $i \in \mathcal{M} \cup \mathcal{W}$, a strategy Q_i that ranks i's first choice first is not dominated by a strategy that does not, since it does strictly better against a profile of strategies in which i is everyone else's first choice. Again, due to the symmetry between m_2, m_3, w_2 and w_3, the following argument is presented only for man m_2. For m_2, ranking w_2 second (and w_1 first) does strictly better than any other strategy that ranks w_1 first (and w_2 lower than second) against a profile of strategies in which m_1 and w_1 report each other as first choices and each $w \in \mathcal{W} / \{w_1\}$ ranks m_2 as her first choice. A similar argument applies to $Q^*_{m_2}$ ranking w_3 third and other women as unacceptable. It follows that $\mathbf{Q}^*$ does not involve the use of weakly dominated strategies and therefore it is an equilibrium and $g(\mathbf{Q}^*)$ is blocked by (m_2, w_3) and (m_3, w_2).

PROPOSITION 4: *Let $|\mathcal{M}|, |\mathcal{W}| \geq 3$. For each monotonic priority matching mechanism $\langle \mathcal{Q}, g\rangle$, there exists a profile of strategies $\mathbf{Q}^*$ and a profile of preferences $\mathbf{P}$ that satisfies monotonicity and independence such that $\mathbf{Q}^*$ is an equilibrium of the mechanism's induced game $\langle \mathcal{Q}, g, \mathbf{P}\rangle$ and $g(\mathbf{Q}^*)$ is not stable w.r.t $\mathbf{P}(\mathbf{Q}^*)$.*

PROOF:

Let $\langle \mathcal{Q}, g\rangle$ be a monotonic priority matching mechanism and let $|\mathcal{M}|, |\mathcal{W}| \geq 3$. Consider the following profile of strategies $\mathbf{Q}^*$: For $j \in \mathcal{M} \cup \mathcal{W} / \{m_1, m_2, w_1, w_2\}$, $Q^*_j = (j, ...)$, $Q^*_{m_1} = (w_1, w_2, m_1, ...)$, $Q^*_{m_2} = (w_2, m_2, ...)$, $Q^*_{w_1} = (m_1, w_1, ...)$ and $Q^*_{w_2} = (m_2, m_1, w_2, ...)$. By the monotonicity of $(\mathcal{Q}, g)$, $(1, 1)$ matches are formed before any others and therefore $g(\mathbf{Q}^*) = ((m_1, w_1), (m_2, w_2), m_3, w_3, ...)$ is the matching that results from $\mathbf{Q}^*$.

Consider the following profile of preferences: $P_j(\mathbf{Q}) = (j, ...)$ for each $\mathbf{Q} \in \mathcal{Q}$ and $j \in \mathcal{M} \cup \mathcal{W} / \{m_1, m_2, w_1, w_2\}$. For each $\mathbf{Q} \in \mathcal{Q}$ and $i \in \{m_2, w_1\}$, $P_i = (g_i(\mathbf{Q}^*), i, ...)$. Man m_1's preferences are $P_{m_1}(\mathbf{Q}) = (w_2, w_1, m_1, ...)$ if w_2 reports that he is at least her second choice and $P_{m_1}(\mathbf{Q}) = (w_1, w_2, m_1, ...)$ otherwise. Woman w_2's preferences are $P_{w_2}(\mathbf{Q}) = (m_1, m_2, w_2, ...)$ if m_1 reports that she is at least his second choice and $P_{w_2}(\mathbf{Q}) = (m_2, m_1, w_2, ...)$ otherwise. Clearly, *independence* is satisfied by $\mathbf{P}$. By the monotonicity of $\langle \mathcal{Q}, g \rangle$, *monotonicity* is satisfied by $\mathbf{P}$. Also, by the monotonicity of $\langle \mathcal{Q}, g \rangle$, $(1, 1)$ matches are formed before any others and therefore $\mathbf{Q}^*$ is a Nash equilibrium of $\langle \mathcal{Q}, g, \mathbf{P} \rangle$. It is easy to see that the pair (m_1, w_2) blocks $g(\mathbf{Q}^*)$. It is left to verify that $\mathbf{Q}^*$ does not include dominated strategies.

For each $i \in \mathcal{M} \cup \mathcal{W} / \{m_1, w_2\}$, Q_i^* cannot be dominated by another strategy Q_i' since it is a best response to any profile of opponents' strategies. It is left to verify that $Q_{w_2}^*$ and $Q_{m_1}^*$ are not dominated. Since the argument for $Q_{w_2}^*$ is identical, we only need to check $Q_{m_1}^*$. The strategy $Q_{m_1}^*$ cannot be dominated by a strategy Q_{m_1}' that reports $w \in \mathcal{W} / \{w_1, w_2\}$ as an acceptable match since $Q_{m_1}^*$ does strictly better than Q_{m_1}' against a profile of strategies in which each $w' \in \mathcal{W} / \{w\}$ reports m_1 as an unacceptable match and $Q_w = (m_1, w, ...)$. Also, the strategy $Q_{m_1}^*$ cannot be dominated by a strategy Q_{m_1}' that reports that $w \in \{w_1, w_2\}$ is an unacceptable match since $Q_{m_1}^*$ does strictly better than Q_{m_1}' against a profile of strategies in which each $w' \in \mathcal{W} / \{w\}$ reports m_1 as an unacceptable match and $Q_w = (m_1, w, ...)$. Finally, $Q_{m_1}^*$ cannot be dominated by a strategy Q_{m_1}' that ranks w_2 first since $Q_{m_1}^*$ does strictly better than Q_{m_1}' against a profile of strategies in which each $w \in \mathcal{W} / \{w_1, w_2\}$ reports m_1 as an unacceptable match, w_1 and w_2 report that m_1 is their third choice and each $m \in \mathcal{M} / \{m_1\}$ reports each $w \in \{w_1, w_2\}$ as an unacceptable match. It follows that $\mathbf{Q}^*$ is a an equilibrium of the mechanism's induced game $\langle \mathcal{Q}, g, \mathbf{P} \rangle$ and $g(\mathbf{Q}^*)$ is not stable w.r.t $\mathbf{P}(\mathbf{Q}^*)$.

PROPOSITION 5: *Suppose that* $\mathbf{Q}^*$ *is an equilibrium of the game that is induced by the modified GSM* $\langle \bar{\mathcal{Q}}, \bar{g}, \mathbf{P} \rangle$. *If* $\mathbf{P}$ *satisfies monotonicity and independence, then* $\bar{g}(\mathbf{Q}^*)$ *is stable w.r.t* $\mathbf{P}(\mathbf{Q}^*)$.

PROOF:

Consider the modified GSM $\langle \bar{\mathcal{Q}}, \bar{g} \rangle$. For each man $m \in \mathcal{M}$, define $T_m \in \bar{\mathcal{Q}}_m$ to be m's strict preferences over $\mathcal{W} \cup \{m\}$ given that for each $w \in \mathcal{W}$, $Q_w = m$. We now claim that any strategy $Q_m \in \bar{\mathcal{Q}}_m$ that ranks women who are an acceptable match according to T_m differently than T_m or reports women who are an acceptable (unacceptable) match according to T_m as an unacceptable (acceptable) match is weakly dominated by the strategy $Q_m = T_m$.

Consider an arbitrary profile $\mathbf{Q}_{-m} \in \bar{\mathcal{Q}}_{-m}$ and assume by negation that there exists a strategy $Q_m' \in \bar{\mathcal{Q}}_m$ such that $\bar{g}_m(Q_m', \mathbf{Q}_{-m}) \, P_m(Q_m', \mathbf{Q}_{-m}) \, \bar{g}_m(T_m, \mathbf{Q}_{-m})$. It cannot be that $\bar{g}_m(Q_m', \mathbf{Q}_{-m}) = m$ since by *independence*, $m T_m \bar{g}_m(T_m, \mathbf{Q}_{-m})$. Also, it cannot be that $\bar{g}_m(T_m, \mathbf{Q}_{-m}) = m$ since it implies that $Q_{\bar{g}_m(Q_m', \mathbf{Q}_{-m})} = m$ and by *independence*, $\bar{g}_m(Q_m', \mathbf{Q}_{-m}) \, T_m m$ which is a contradiction. It follows that $\bar{g}_m(Q_m', \mathbf{Q}_{-m}), \bar{g}_m(T_m, \mathbf{Q}_{-m}) \in \mathcal{W}$. Moreover, it must be that $Q_{\bar{g}_m(Q_m', \mathbf{Q}_{-m})} = Q_{\bar{g}_m(T_m, \mathbf{Q}_{-m})} = m$. By *independence*, it must be that $\bar{g}_m(Q_m', \mathbf{Q}_{-m}) \, T_m \bar{g}_m(T_m, \mathbf{Q}_{-m})$. It follows that under T_m, m proposes to $\bar{g}_m(Q_m', \mathbf{Q}_{-m})$ before proposing to $\bar{g}_m(T_m, \mathbf{Q}_{-m})$ and is rejected by $\bar{g}_m(Q_m', \mathbf{Q}_{-m})$, which is a contradiction since $Q_{\bar{g}_m(Q'_m, \mathbf{Q}_{-m})} = m$. It follows that T_m is a best response to any $\mathbf{Q}_{-m} \in \mathcal{Q}_{-m}$.

It is left to show that there exists a profile $\mathbf{Q}_{-m} \in \mathcal{Q}_{-m}$ such that T_m does strictly better than Q'_m against $\mathbf{Q}_{-m}$. There are three possible cases. First, suppose that $Q'_m \neq T_m$ reports a woman $w \in \mathcal{W}$ who is an unacceptable (acceptable) match by T_m as an acceptable (unacceptable) match. By $independence$, w is unacceptable (acceptable) for m whenever $Q_w = m$. Clearly, T_m does strictly better than Q'_m against a profile $\mathbf{Q}'_{-m} \in \mathcal{Q}_{-m}$ in which $Q'_w = m$ and $Q'_{w'} \neq m$ for each $w' \in \mathcal{W}/\{w\}$. For the complementary case, suppose that there are two women $w, w' \in \mathcal{W}$ such that $w' T_m m$, $w T_m w'$ and $w' Q'_m w$. The strategy T_m does strictly better than the strategy Q'_m against a profile $\mathbf{Q}^*_{-m} \in \mathcal{Q}_{-m}$ such that $Q^*_w = Q^*_{w'} = m$ and $Q^*_{\bar{w}} \neq m$ for each $\bar{w} \in \mathcal{W}/\left\{w, w'\right\}$ since by $independence$, $w T_m w'$ implies $w P_m \left(Q'_m, \mathbf{Q}^*_{-m}\right) w'$. It follows that T_m weakly dominates any other strategy $Q'_m \in \bar{\mathcal{Q}}_m$.

Let $\mathbf{Q}^*$ be an equilibrium of the mechanism's induced game $\left\langle \bar{\mathcal{Q}}, \bar{g}, \mathbf{P}\right\rangle$. Suppose that $\bar{g}(\mathbf{Q}^*)$ is not stable w.r.t $\mathbf{P}(\mathbf{Q}^*)$. By the individual rationality property of the mechanism, $\bar{g}(\mathbf{Q}^*)$ must be blocked by a pair (m, w). It follows that $w P_m \left(Q^*_w, \mathbf{Q}^*_{-w}\right) \bar{g}_m(\mathbf{Q}^*)$. By $monotonicity$, $w P_m \left(m, \mathbf{Q}^*_{-w}\right) \bar{g}_m(\mathbf{Q}^*)$ and by $independence$, $w T_m \bar{g}_m(\mathbf{Q}^*)$. It follows that $m = \bar{g}_w\left(m, \mathbf{Q}^*_{-w}\right)$ which contradicts $\mathbf{Q}^*$ being an equilibrium of $\left\langle \bar{\mathcal{Q}}, \bar{g}, \mathbf{P}\right\rangle$ since woman w can deviate profitably and marry man m.

PROPOSITION 6: *Consider the modified GSM and suppose that for each* $i \in \mathcal{M} \cup \mathcal{W}$ *and* $\mathbf{Q}, \mathbf{Q}' \in \mathcal{Q}$, $P_i(\mathbf{Q}) = P_i\left(\mathbf{Q}'\right) = P'_i$. *Let* μ *be a matching that is stable w.r.t* $\mathbf{P}'$. *There exists a profile of strategies* $\mathbf{Q}^*$ *such that* $\mathbf{Q}^*$ *is an equilibrium of* $\left\langle \bar{\mathcal{Q}}, \bar{g}, \mathbf{P}'\right\rangle$ *and* $\bar{g}(\mathbf{Q}^*) = \mu$.

PROOF:

Consider the modified GSM $\left\langle \bar{\mathcal{Q}}, \bar{g}\right\rangle$ and suppose that for each $i \in \mathcal{M} \cup \mathcal{W}$ and $\mathbf{Q}, \mathbf{Q}' \in \mathcal{Q}$, $P_i(\mathbf{Q}) = P_i\left(\mathbf{Q}'\right) = P'_i$. Let μ be a matching that is stable w.r.t $\mathbf{P}'$. Let $Q^*_m = P'_m$ for each $m \in \mathcal{M}$ and let $Q^*_w = \mu(w)$ for each $w \in \mathcal{W}$ such that $\mu(w) \neq w$. For each $w \in \mathcal{W}$ such that $\mu(w) = w$, let Q^*_w be a man $m \in \mathcal{M}$ who is an acceptable match for her w.r.t P'_w. If such a man does not exist, then let $Q^*_w = w$.

We now prove that $\bar{g}(\mathbf{Q}^*) = \mu$ by showing that $m = \mu(w)$ if and only if $m = \bar{g}_w(\mathbf{Q}^*)$. First, suppose that $m \neq \mu(w)$, then by construction of $\mathbf{Q}^*$, $m = \bar{g}_w(\mathbf{Q}^*)$ implies that $w P'_m \mu(m)$, $\mu(w) = w$ and $m P'_w w$ which contradicts the stability of μ w.r.t $\mathbf{P}'$. Secondly, suppose that $w = \mu(m)$ and $w \neq \bar{g}_m(\mathbf{Q}^*)$. By the construction of $\mathbf{Q}^*$, $\bar{g}_m(\mathbf{Q}^*) = w' \neq w$ implies that $m P'_{w'} w'$, $w' P'_m w$ and $w' = \mu\left(w'\right)$ which violates the stability of μ w.r.t $\mathbf{P}'$. Also, by the construction of $\mathbf{Q}^*$, $\bar{g}_m(\mathbf{Q}^*) = m$ implies that $m P'_m w$ which contradicts the stability of μ w.r.t $\mathbf{P}'$.

It remains to prove that $\mathbf{Q}^*$ is an equilibrium of $\left\langle \bar{\mathcal{Q}}, \bar{g}, \mathbf{P}'\right\rangle$. It follows from Dubins and Freedman (1981) and Roth (1982) that each $Q^*_m \in \mathbf{Q}^*$ is a best reply to $\mathbf{Q}^*_{-m}$ and that $Q^*_m \in \mathbf{Q}^*$ is not dominated by any other $Q_m \in \mathcal{Q}_m$. Each $Q^*_w \in \mathbf{Q}^*$ such that $Q^*_w \in \mathcal{M}$ is not dominated since it does strictly better than any other strategy against a profile of strategies in which $Q_{Q^*_w} = \left(w, Q^*_w, ...\right)$ and $Q_m = (m, ...)$ for each $m \in \mathcal{M}/\left\{Q^*_w\right\}$. By the construction of $\mathbf{Q}^*$, each $Q^*_w \in \mathbf{Q}^*$ such that $Q^*_w = w$ is a best response to any

$\mathbf{Q}_{-w} \in \mathcal{Q}_{-w}$ since each $m \in \mathcal{M}$ is an unacceptable match for w. The stability of μ w.r.t $\mathbf{P}'$ along with $\mathbf{Q}^*_{\mathcal{M}} = \mathbf{P}'_{\mathcal{M}}$ imply that if a woman $w \in \mathcal{W}$ deviates from Q^*_w to $Q'_w \in \mathcal{Q}_w$ she cannot marry a man $m \in \mathcal{M}$ such that $m P'_w \mu(w)$ and therefore neither of the women has a profitable deviation. It follows that $\mathbf{Q}^*$ is an equilibrium of $\langle \bar{\mathcal{Q}}, \bar{g}, \mathbf{P}' \rangle$.