

# Stability of marriage with externalities

Isa E. Hafalir

Accepted: 20 November 2007  
© Springer-Verlag 2008

**Abstract** In many matching problems, it is natural to consider that agents may have preferences not only over the set of potential partners but also over what other matches occur. Once such externalities are considered, the set of stable matchings will depend on what agents believe will happen if they deviate. In this paper, we introduce endogenously generated beliefs (which depend on the preferences). We introduce a particular notion of endogenous beliefs, called sophisticated expectations, and show that with these beliefs, stable matchings always exist.

**Keywords** Cooperative games · Matchings · Externalities

**JEL Classification** C71 · C78 · D62

## 1 Introduction

In the standard model of matching, introduced by [Gale and Shapley \(1962\)](#), agents on one side of the market, say men, are assumed to have preferences over agents on the other side of the market, say women, and vice-versa. The central concern is to identify matchings that are stable in the sense that no unmatched pair of agents prefer each other to their current matches. The reader is referred to [Roth and Sotomayor \(1990\)](#) for a detailed discussion of the literature.

---

I would like to thank the editor, William Thomson, two anonymous referees, Kalyan Chatterjee, Federico Echenique, Matthew Jackson, Tarik Kara, Semih Koray, and Manabu Toda for their comments and suggestions. I am very much indebted to Vijay Krishna for his guidance and support.

---

I. E. Hafalir (✉)  
Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA 15213, USA  
e-mail: isaemin@cmu.edu

Matching models are natural in many contexts—marriages, college admissions and labor markets. In many of these situations, some or all agents' preferences may be subject to externalities. A firm may care not only about the quality of its own employees but also about the quality of the employees of a competitor. Even in the marriage context, jealousy may play a role.

When externalities are present, a deviating pair needs to consider how other agents will react to the deviation. Consider a software firm that considers luring away a programmer from a competitor. How the rival firm reacts—perhaps by replacing the person with an even better programmer—will affect how profitable the deviation is. In such a situation, the expectations that the original firm has about the reactions of the rival become significant.

In an important paper, [Sasaki and Toda \(1996\)](#) were the first to consider matching problems with externalities. In their model of one-to-one matching, also called a marriage market, the idea of externalities is captured by specifying preferences over the complete matching rather than just over the set of agents on the other side of the market. The expectations of a deviating pair are specified via what they call *estimation functions*. An estimation function specifies the set of matchings among all other agents that the deviating agents consider to be possible. Given such estimation functions, a pair of man and woman *block* a matching if they are made better off under all matchings in their estimation functions when they form a pair. A matching that is not blocked is said to be *stable*. Notice that this formulation is non-Bayesian—that is, players do not assign probabilities to the different matchings, but rather deviate only if the deviation is profitable for all possible matchings that they consider to be possible. [Sasaki and Toda \(1996\)](#) show, however, that only the *universal* estimation function—one that considers all matchings to be possible—is compatible with the existence of a stable matching. In their model, however, the estimation functions are specified *exogenously*. In particular, the set of matchings that a deviating pair considers possible does not depend on the preferences of the other agents. But clearly the matchings among other agents that will result from the deviation will depend on these agents' preferences. Therefore, this dependence should be recognized by the deviating pair.<sup>1</sup> In this paper, we study endogenous estimation functions.

Recently, there has been a surge of literature about externalities in different cooperative/coalition formation frameworks. For instance, in coalition formation games with externalities, [Bloch \(1996\)](#), [Ray and Vohra \(1999\)](#) and [Maskin \(2003\)](#) proposed different ideas about which coalitions would form and how payoffs would be divided. In all these settings, the estimations are unique. That is, while making a decision, agents foresee what will happen in the future precisely. In addition, in these settings utility is transferable.

In our setting, we need to introduce estimation functions because there is no protocol (agents do not give their decisions in a fixed order, decision can be made and changed any time). Moreover, in our setting, utility is nontransferable.

<sup>1</sup> [Roy Chowdhury \(2005\)](#) studies a model in which the agents assume that a deviation will trigger no response from others (agents are allowed to remain single). Stable matchings then exist only under strong assumptions on preferences.

In this paper, we adapt [Sasaki and Toda \(1996\)](#) model but introduce *endogenous* estimations—that is, the set of matchings considered possible by a deviating pair depends on the preferences of other agents. We first provide a sufficient condition for estimation functions to be compatible with the existence of stable matchings ([Proposition 2](#)). We then investigate a specific but important estimation function which satisfies this condition. Specifically, we define a plausible notion of *sophisticated expectations* that are not universal, but still guarantee that the set of stable matchings is nonempty for all preference profiles ([Proposition 3](#)).

Sophisticated expectations are determined via an algorithm. The algorithm, defined formally in [Sect. 3.3](#), is based on the idea that given some estimation functions, one can induce in a natural way a matching game in which there are no externalities. This game has a nonempty set of stable matchings. Sophisticated agents would recognize this fact and append these stable matchings to their original estimations. The new estimation functions define a new induced problem and give a new set of stable matchings, which are then appended in the same way. The process continues until there are no additional matchings in the induced game. The resulting estimation function is what we call “sophisticated”.

The notion of sophisticated expectations is different from that of *rational expectations*. The latter notion leads to estimation functions that satisfy a kind of reduced game consistency common in cooperative game theory—essentially, only stable matchings of the reduced game are considered possible. But as anticipated by [Sasaki and Toda \(1996\)](#), rational expectations do not guarantee the existence of stable matchings.<sup>2</sup>

In [Sect. 4](#), we consider the question of deviations by larger coalitions—recall that the notion of stability is based on pairwise deviations. As is well-known, in matching models without externalities, allowing for deviations by larger coalitions does not affect the set of unblocked matchings. In other words, the core is the same as the set of stable matchings. This equivalence, however, does not extend to situations with externalities. Indeed, it is known that the core may be empty when externalities are present. We thus examine a more permissive notion, the *bargaining set*. Although the bargaining set may, in general, also be empty, we provide a sufficient condition on preferences that ensures that this is not so ([Proposition 4](#)).

## 2 The model and exogenous estimations

This paper models two-sided one-to-one matchings. These are called marriage markets in the literature, and so let  $M$  and  $W$  denote the finite sets of men and women and  $M \cap W = \emptyset$ . We suppose that there are equal numbers of each so that  $|M| = |W| = n$ .<sup>3</sup>

<sup>2</sup> The idea of rational expectations is applied in [Li \(1993\)](#) multiple principal agent model with externalities. Li obtains the existence of the equilibrium in the case of weak externalities, which reduces to the following in our model: agents have lexicographic preferences in which the priority is given to their matches. However, in this case the blocking decision of pairs does not depend on what others do after deviation in our setting. That is, effectively there will be no externalities.

<sup>3</sup> A more general model would allow for weak preferences, different sizes of the two sides of the market and some matches to be not individually rational. We opted to use a more basic model to focus on endogenous beliefs of the agents.

A bijection  $\mu : M \cup W \rightarrow M \cup W$  is called a **matching** if (i)  $\mu(\mu(a)) = a$  for all  $a \in M \cup W$ ; (ii)  $\mu(m) \in W$  for all  $m \in M$  and  $\mu(w) \in M$  for all  $w \in W$ .

Thus  $(m, w) \in \mu$  means that  $m$  and  $w$  are paired in the matching  $\mu$ . Let  $A(M, W)$  be the set of all matchings and  $A(m, w) = \{\mu \in A(M, W) \mid (m, w) \in \mu\}$  denote the set of matchings where  $m$  and  $w$  are paired. Each  $a \in M \cup W$  has a *strict* preference ordering  $\succ_a$  over  $A(M, W)$ . Note that this is the most general way of representing externalities since agents have preferences over the complete set of matchings. Let  $\succ$  denote a preference profile, that is,  $\succ = \{\succ_a \mid a \in M \cup W\}$ . The triplet  $(M, W, \succ)$  is called a **matching problem with externalities**.

Let  $\varphi_m(w) \subseteq A(m, w)$  be the set of matchings which  $m$  considers *possible* when  $w$  is paired with him. Similarly, let  $\varphi_w(m) \subseteq A(m, w)$  be the set of matchings which  $w$  considers possible when  $m$  is paired with her. Sasaki and Toda (1996) call  $\varphi_m$  and  $\varphi_w$  **estimation functions**. They denote an **estimation profile** of agents by  $\varphi = \{\varphi_a \mid a \in M \cup W\}$ .

Given an estimation profile  $\varphi$ , a matching  $\mu$  is  $\varphi$ -**admissible** if for all pairs  $(m, w) \in \mu$ ,

$$\mu \in \varphi_m(w) \quad \text{and} \quad \mu \in \varphi_w(m), \quad (1)$$

and a matching  $\mu$  is **blocked** by a pair  $(m, w) \notin \mu$  if for all  $\mu' \in \varphi_m(w)$  and for all  $\mu'' \in \varphi_w(m)$ ,

$$\mu' \succ_m \mu \quad \text{and} \quad \mu'' \succ_w \mu. \quad (2)$$

A matching  $\mu$  is  $\varphi$ -**stable** if it is  $\varphi$ -admissible and is not blocked. Let  $S_\varphi(M, W, \succ)$  denote the set of all  $\varphi$ -stable matchings.<sup>4</sup>

Sasaki and Toda (1996) establish that if the estimation functions for a pair  $(m, w)$  are such that for at least one of them, the estimation function is not the set of all matchings, then there exists a preference profile such that the set of  $\varphi$ -stable matchings is empty.

**Proposition 1** *For all  $n \geq 3$ , if there exists a pair  $(m, w)$  such that either  $\varphi_m(w) \neq A(m, w)$  or  $\varphi_w(m) \neq A(m, w)$ , then there exists a preference profile  $\succ$  such that  $S_\varphi(M, W, \succ) = \emptyset$ .*

As the statement of the proposition makes apparent, the estimation functions are assumed to be exogenously given—in particular, they do not depend on preferences. This seems unnatural, however, since the set of potential matchings that  $m$  estimates as being likely when he is paired with  $w$  may well depend on the preferences of the other  $2n - 2$  agents. As an extreme case, suppose that there is another pair  $(m', w')$  whose preferences are not subject to any external effects and each member of the pair considers the other to be the best mate. Then it seems natural that every estimation function for both  $m$  and  $w$  should only allow  $m'$  to be paired with  $w'$ . Thus, estimation functions should not be required to be independent of preferences.

For instance, it seems natural that, given a preference profile  $\succ$ , only estimation functions satisfying the following minimal condition (which is weaker than the above extreme case) should be admitted.

<sup>4</sup> In the rest of the paper, the term “stability” is sometimes used without specifying the estimation function. The estimation functions are omitted whenever there is no confusion.

**Definition 1** An estimation  $\varphi$  has *No Matched-Couple Veto Property (NMCVP)* if the following condition is satisfied: Let  $(m, w), (m', w') \in \mu$  for some  $\mu \in A(M, W)$ . If for all  $a \in M \cup W - \{m, m', w, w'\}$  and all  $\mu^a \in A(m, w) \setminus A(a, \mu(a))$ ,  $\mu \succ_a \mu^a$ , then  $\mu \in \varphi_m(w) \cap \varphi_w(m)$ .

In words, NMCVP means that among  $n - 1$  couples that would be formed after a deviation, if  $n - 2$  of these couples prefer their mates in a specific matching  $\mu$  to any other matching, then the  $(n - 1)$ st couple would not have any option but to form a pair after the deviation. Thus,  $\mu$  should be recognized as a plausible matching by the deviators.

If we require that estimation functions satisfy NMCVP, then Proposition 1 no longer holds.<sup>5</sup> It is because the constructed preference profile in its proof makes  $\varphi$  violate NMCVP. While natural, NMCVP is by itself not the only property that we would like estimation functions to satisfy. It is just one natural assumption which, we think, should be satisfied by estimation functions. However, NMCVP is not a sufficient condition for the existence of stable matchings (as we will show that the rational expectations satisfy NMCVP, but may result in an empty set of stable matchings), and we do not know whether it is a necessary condition for the existence of stable matching.

Sasaki and Toda (1996) also establish that for universal estimations (with  $\varphi_m(w) = \varphi_w(m) = A(m, w)$ ), the set of stable matchings is nonempty. But like Proposition 1, this result also relies on the fact that agents consider all matchings to be possible even though these may be “irrational” given the preference profile.

We will show below that given any preference profile, there exist endogenously generated estimation functions—which do not always generate the set of all matchings and satisfy NMCVP—such that the resulting set of stable matchings is nonempty. We develop a procedure for finding such estimation functions.

### 3 Endogenous estimations

In this section, we proceed as follows. Fix a particular preference profile  $\succ$ . We suppose that if  $m$  and  $w$  are paired, then they have the *same* set of feasible matchings. Thus, with a slight abuse of notation we denote by  $\varphi(m, w)$ , the set of matchings considered feasible by both  $m$  and  $w$ , when  $m$  is paired with  $w$ . Formally,  $\varphi_m(w) = \varphi_w(m) \equiv \varphi(m, w)$ . We demonstrate that there are endogenously generated estimation functions satisfying the assumption of equal expectations which result in a nonempty set of stable matchings.

We denote an estimation profile by  $\varphi = \{\varphi(m, w) \mid (m, w) \in M \times W\}$ . Note that  $\varphi(m, w)$  depends on preferences but since  $\succ$  is assumed to be fixed, the dependence of  $\varphi$  on  $\succ$  is suppressed. Let  $S_\varphi(M, W)$  denote the set of stable matchings.

<sup>5</sup> This, in turn, implies that the only exogenous estimation function satisfying NMCVP is the universal estimations.

### 3.1 Rational expectations

One natural way to formulate the notion of an endogenous estimation is via a “reduced game” consistency condition. Specifically, suppose the pair  $(m, w)$  is formed and “exit”. Then we have a reduced game with sets  $M' = M - \{m\}$  and  $W' = W - \{w\}$ . Let  $S_\varphi(M', W')$  denote the set of stable matchings in this game where preferences  $\succ'_a$  are restrictions of  $\succ_a$  to the set of matchings of the remaining agents—that is, not including  $m$  and  $w$ . Agents are said to have **rational expectations**<sup>6</sup> estimations if the estimation of  $m$  and  $w$  is just  $S_\varphi(M', W')$ .

Formally, the rational expectations estimation function  $\rho(M, W)$  is defined inductively as follows.

When  $n = 2$ , there are no externalities. This is because if a pair  $(m, w)$  forms, the only possibility for the other couple is to form a pair. So for  $n = 2$ , the set of stable matchings  $S(M, W)$  is nonempty (Gale and Shapley 1962) and does not depend on any estimation function. This means that when  $n = 3$ , consistency demands that we set  $\rho(m, w) = S(M', W')$ , where  $M' = M - \{m\}$  and  $W' = W - \{w\}$ . Now for  $n = 3$ , denote by  $S_\rho(M, W)$  the set of stable matchings.

If this set is not empty, we proceed to the next step. For  $n = 4$ , we set  $\rho(m, w) = S_\rho(M', W')$ , where again  $M' = M - \{m\}$  and  $W' = W - \{w\}$  and so on.

This notion is well defined, however, only if at every stage the set of stable matchings is nonempty. But, as shown by Sasaki and Toda (1996), this may not be the case. The following example for  $n = 3$  shows that the rational expectations estimation function does not guarantee the existence of stable matchings.

*Example 1* Let  $n = 3$ . Then we have six different matchings, and each agent has preferences over these. Suppose that agents assign utilities 1 to 6 starting from the least preferred matching to the most preferred matching (these are only ordinal). Consider the following preferences:

	1	5	3		4	1	6		3	6	1
	$m_1$	$m_2$	$m_3$		$m_1$	$m_2$	$m_3$		$m_1$	$m_2$	$m_3$
$\mu_1$	$w_1$	$w_2$	$w_3$	$\mu_2$	$w_2$	$w_3$	$w_1$	$\mu_3$	$w_3$	$w_1$	$w_2$
	3	2	4		3	5	2		3	4	4
	5	2	5		6	4	2		2	3	4
	$m_1$	$m_2$	$m_3$		$m_1$	$m_2$	$m_3$		$m_1$	$m_2$	$m_3$
$\mu_4$	$w_1$	$w_3$	$w_2$	$\mu_5$	$w_3$	$w_2$	$w_1$	$\mu_6$	$w_2$	$w_1$	$w_3$
	1	6	5		1	6	5		1	6	2

For  $n = 2$ , there are no externalities, and so the rational expectations estimations can be easily obtained. We can confirm that  $\rho(m_1, w_1) = \{\mu_4\}$ ,  $\rho(m_1, w_2) = \{\mu_6\}$ ,  $\rho(m_1, w_3) = \{\mu_5\}$ ,  $\rho(m_2, w_1) = \{\mu_3\}$ ,  $\rho(m_2, w_2) = \{\mu_1\}$ ,  $\rho(m_2, w_3) = \{\mu_2\}$ ,  $\rho(m_3, w_1) = \{\mu_5\}$ ,  $\rho(m_3, w_2) = \{\mu_3\}$ , and  $\rho(m_3, w_3) = \{\mu_1\}$ .

<sup>6</sup> The rational expectations notion is defined both in Li (1993) and Sasaki and Toda (1996). We follow their terminology.

For  $\rho(m, w) = S(M', W')$ ,  $\mu_1$  is blocked by  $(m_2, w_1)$ ;  $\mu_2$  is blocked by  $(m_2, w_1)$ ;  $\mu_3$  is blocked by  $(m_3, w_3)$ ;  $\mu_4$  is blocked by  $(m_2, w_1)$ ;  $\mu_5$  is blocked by  $(m_3, w_3)$ ;  $\mu_6$  is blocked by  $(m_2, w_2)$ . So the set of stable matchings is empty.

We can also confirm that although rational expectations might result in an empty set of stable matchings, rational expectations (when it is well defined) satisfies NMCVP. Suppose that  $m$  and  $w$  are paired, and among the remaining agents, each  $a \in M \cup W - \{m, m', w, w'\}$  satisfies  $\mu \succ_a \mu^a$  for all  $\mu^a \in A(m, w) \setminus A(a, \mu(a))$  and for some  $\mu \in A(m, w) \cap A(m', w')$ . Then,  $\mu \in \rho(m, w)$ . This is because, in the *reduced problem* with sets  $M' = M - \{m\}$  and  $W' = W - \{w\}$ , no  $a \in M \cup W - \{m, m', w, w'\}$  would deviate from  $\mu$ , hence,  $m'$  or  $w'$  cannot deviate either (since they cannot find partners willing to pair with them). Therefore,  $\mu$  is a stable matching in the reduced problem.

### 3.2 Estimations with nonempty set of stable matchings

In this section, we give a sufficient condition for estimation functions to be compatible with the existence of stable matchings.

We can define an induced preference profile of agents over the agents on the other side as follows. The ranking that a particular  $m \in M$  assigns to a  $w \in W$ , is the same as the ranking of the worst matching in  $\varphi(m, w)$  among all other worst matchings in the collection  $\varphi(m, w')$ ,  $w' \in W$ . This defines an *induced problem* without externalities. Formally, given  $\varphi(m, w)$ , for all  $m \in M$  and all  $w \in W$ , let  $\mu_m^\varphi(w)$  be the worst matching for  $m$  in  $\varphi(m, w)$ , and let  $\mu_w^\varphi(m)$  be the worst matching for  $w$  in  $\varphi(m, w)$ . That is,  $\mu_m^\varphi(w)$  and  $\mu_w^\varphi(m)$  are the unique matchings satisfying  $\mu \succ_m \mu_m^\varphi(w)$  for all  $\mu \in \varphi(m, w) - \{\mu_m^\varphi(w)\}$ , and  $\mu \succ_w \mu_w^\varphi(m)$  for all  $\mu \in \varphi(m, w) - \{\mu_w^\varphi(m)\}$  respectively.

Define the preference without externality  $\succ^\varphi$  as:

$$w \succ_m^\varphi w' \text{ iff } \mu_m^\varphi(w) \succ_m \mu_m^\varphi(w') \quad \text{and} \quad m \succ_w^\varphi m' \text{ iff } \mu_w^\varphi(m) \succ_w \mu_w^\varphi(m') \quad (3)$$

Let  $I^\varphi$  denote the set of stable matchings of  $(M, W, \succ^\varphi)$ . That is:

$$\begin{aligned} I^\varphi &= \{ \mu \in A : \nexists (m, w) \notin \mu \text{ s.t. } w \succ_m^\varphi \mu(m) \quad \text{and} \quad m \succ_w^\varphi \mu(w) \} \\ &= \{ \mu \in A : \nexists (m, w) \notin \mu \text{ s.t. } \mu_m^\varphi(w) \succ_m \mu_m^\varphi(\mu(m)) \quad \text{and} \quad \mu_w^\varphi(m) \succ_w \mu_w^\varphi(\mu(w)) \} \end{aligned} \quad (4)$$

Define

$$I^\varphi(m, w) = \{ \mu \in A(m, w) : \mu \in I^\varphi \} \quad (5)$$

to be the projection of the set of stable matchings onto  $M' \times W'$ . One should not confuse  $I^\varphi(m, w)$  with  $\varphi(m, w)$ . The former is the set of stable matchings for the preference  $\succ^\varphi$  in which  $m$  and  $w$  are paired, and the latter is the set of matchings in the estimation function of  $m$  and  $w$ .

The next proposition establishes that any estimation function  $\varphi$  such that for every preference profile, there exists a matching  $\mu$  which is both  $\varphi$ -admissible and in the

set of stable matchings of the induced problem, results in a nonempty set of stable matchings  $S_\varphi$ .

**Proposition 2** *For an estimation function  $\varphi$ , if  $\mu \in \varphi(m, w) \cap I^\varphi(m, w)$  for all  $(m, w) \in \mu$ , then  $\mu \in S_\varphi(M, W)$ .*

*Proof* If  $\mu \in \varphi(m, w) \cap I^\varphi(m, w)$  for all  $(m, w) \in \mu$ , then  $\mu \in \varphi(m, w)$  for all  $(m, w) \in \mu$ . So admissibility is obviously satisfied.

We now show that  $\mu$  is also unblocked by contradiction. Assume that there is  $(m, w) \notin \mu$  such that  $\mu_m^\varphi(w) \succ_m \mu$  and  $\mu_w^\varphi(m) \succ_w \mu$ . Since  $\mu \in \varphi(m, \mu(m))$  and  $\mu \in \varphi(\mu(w), w)$  because of admissibility, we have  $\mu \succeq_m \mu_m^\varphi(\mu(m))$  and  $\mu \succeq_w \mu_w^\varphi(\mu(w))$ . Transitivity of preferences implies that  $\mu_m^\varphi(w) \succ_m \mu_m^\varphi(\mu(m))$  and  $\mu_w^\varphi(m) \succ_w \mu_w^\varphi(\mu(w))$ , which contradicts  $\mu \in I^\varphi(m, w)$ . Hence,  $\mu \in S_\varphi(M, W)$ .  $\square$

In the next section we introduce a special but an important estimation function which satisfies the necessary condition of Proposition 2.

### 3.3 Sophisticated expectations

In our model, agents are assumed to be very cautious—they do not block if the current matching is better than the worst that could happen after they deviate. Moreover, we suppose that there is no commitment—agents can block anytime, even after they block a matching once. Therefore, when  $m$  and  $w$  block a matching and form a pair, other agents would not necessarily restrict their estimation functions to  $A(m, w)$ . Note that if there were such a commitment, the only natural estimation function would be rational expectations, and in this case, there might be no stable matchings.

Consider  $m$  and  $w$  who are about to block. Consider  $\mu \in A(m, w)$  and suppose that (given agents' estimation functions) for no unmatched pair  $(m', w')$ , the worst matching in  $\varphi(m', w')$  is better than the worst matching in  $\varphi(m', \mu(m'))$  for  $m'$  and similarly, the worst matching in  $\varphi(m', w')$  is better than the worst matching in  $\varphi(\mu(w'), w')$  for  $w'$ . Then, we assume that  $\mu$  should be in the estimation function  $\varphi(m, w)$ . We think that this is a reasonable assumption because agents are cautious. They consider the worst outcome (of course, in their estimation function) while deciding whom they are going to be paired with. Hence, if a matching is sustainable using this criterion, it should be a matching one should consider to be possible. By imposing this assumption on the estimation function, we define another notion, *sophisticated expectations*, and denote it by  $\sigma(m, w)$ .<sup>7</sup> We will show that unlike the rational expectations, this notion guarantees that the resulting set of stable matchings is nonempty.

Take any stable matching  $\mu$  of the induced game  $(M, W, \succ^\varphi)$ .<sup>8</sup> Then, from the assumption on estimation functions above,  $m$  and  $w$  should have this matching in their estimation function  $\varphi(m, w)$ . Hence,  $m$  considers every matching in the set of

<sup>7</sup> As it will be clear from the formal definition, sophisticated estimations is the minimal estimation function which includes rational estimations and satisfies this assumption.

<sup>8</sup> We know that there exists such matchings from Gale and Shapley (1962).

stable matchings of the induced game to be also possible and appends them to his original estimation.

The notion is defined inductively. For  $n = 2$ , the estimation function is a singleton. And we know that the set of stable matchings is nonempty for  $n = 2$ . With the assumption that  $S_\sigma(M', W')$ , the set of stable matchings for  $|M'| = |W'| = l < n$ , is nonempty for all  $(M', W')$ , we want to show that  $S_\sigma(M, W)$ , the set of stable matchings for  $|M| = |W| = l + 1$ , is also nonempty.

So suppose that  $S_\sigma(M', W')$  is nonempty for all  $M'$  and  $W'$  satisfying  $|M'| = |W'| = l < n$ . Write  $\sigma^1(m, w) = S_\sigma(M', W')$  and call  $\sigma^1$  a **1st degree estimation**. Now suppose that if the pair  $(m, w)$  deviates, then  $m$  will believe that as a result of this deviation, the *worst* matching in  $\sigma^1(m, w)$  from his perspective will result. This induces a preference ordering for  $m$  over  $W$  (similarly for  $w$  over  $M$ ). Thus a matching game *without* externalities, called the **1st order induced problem** is defined. This induced game has a set of stable matchings, denoted by  $I^{\sigma^1}$ . For sophisticated agents, it is natural to assume that they will have  $I^{\sigma^1}$  also in their estimation functions. The set  $I^{\sigma^1}$ , together with  $\sigma^1(m, w)$  form a new estimation function. We call this estimation function **2nd degree estimation** and denote it by  $\sigma^2(m, w)$ .

Then similarly, everybody will take into account the worst matching in  $\sigma^2(m, w)$  as a result of their deviation. Again, this will generate an induced preference over the opposite sex and thus define a matching game without externalities. Again, this will have a set of stable matchings, denoted by  $I^{\sigma^2}$  and so on. Agents who are sophisticated enough will proceed in this way until the induced stable matching gives nothing new. We denote the final estimation function by  $\sigma(\cdot, \cdot)$ . We call it the **sophisticated estimation function**. We will show that sophisticated expectations estimation function is compatible with the existence of stable matchings.

Formally, let  $\sigma^1(m, w) = S_\sigma(M', W')$  and for  $k = 1, 2, 3, \dots$  inductively define,

$$\sigma^{k+1}(m, w) = \sigma^k(m, w) \cup I^{\sigma^k}(m, w).$$

(see the definitions (3)–(5))

Since  $S_\sigma(M', W')$  was assumed to be nonempty,  $\sigma^k$  and  $I^{\sigma^k}$  are well defined and nonempty (Gale and Shapley 1962).

Note that for all  $(m, w) \in M \times W$  and for  $k = 1, 2, 3, \dots$

$$\sigma^k(m, w) \subseteq \sigma^{k+1}(m, w) \quad \text{and} \quad \sigma^k(m, w) \subseteq A(m, w).$$

Thus the  $\sigma^k$  sequence of sets is monotone and the set of all matchings is finite. Thus it has a limit. Now, let

$$\sigma(m, w) = \lim_{k \rightarrow \infty} \sigma^k(m, w),$$

and denote by  $I^\sigma(m, w)$  the set of stable matchings for preferences  $\succ^\sigma$ . Note that

$$I^\sigma(m, w) \subset \sigma(m, w), \quad \forall (m, w) \in M \times W. \tag{6}$$

**Proposition 3** For the estimation function  $\sigma$ , the set of stable matchings  $S_\sigma$  is non-empty. In fact,  $I^\sigma \subset S_\sigma$ .

*Proof* The proof directly follows from Proposition 2. This is because  $I^\sigma(m, w) \subset \sigma(m, w)$ , and hence, every  $\mu \in I^\sigma(m, w)$  is a stable matching.  $\square$

Note that this estimation function also satisfies NMCVP, since it includes the rational expectations and the rational expectations satisfy the condition.

### 3.4 Discussion of sophisticated expectations

One may wonder whether  $\sigma(m, w) = A(m, w)$  and so whether the result above follows from Sasaki and Toda’s existence result. The next example demonstrates however, that we may have  $\sigma(m, w) \subsetneq A(m, w)$  and so the set of stable matchings is nonempty even though the estimation function does not include all matchings.

Consider Example 1 again. Recall that there exists no stable matching for the rational expectations estimations,  $\rho$ .

*Example 2* Suppose  $n = 3$  and consider the utility assignments of Example 1.

Let us first find  $\sigma^1(m, w) = S_\sigma(M', W')$ .

$$\begin{aligned} \sigma^1(m_1, w_1) &= \{\mu_4\}, & \sigma^1(m_1, w_2) &= \{\mu_6\}, & \sigma^1(m_1, w_3) &= \{\mu_5\} \\ \sigma^1(m_2, w_1) &= \{\mu_3\}, & \sigma^1(m_2, w_2) &= \{\mu_1\}, & \sigma^1(m_2, w_3) &= \{\mu_2\} \\ \sigma^1(m_3, w_1) &= \{\mu_5\}, & \sigma^1(m_3, w_2) &= \{\mu_3\}, & \sigma^1(m_3, w_3) &= \{\mu_1\} \end{aligned}$$

So with  $\mu_{m_i}^{\sigma^1}(w_j)$  and  $\mu_{w_i}^{\sigma^1}(m_j)$  values, we have the following induced problem with preferences  $\succ^{\sigma^1}$ :

	5	5	3		2	1	2		6	6	1
$\mu_1$	$m_1$	$m_2$	$m_3$		$m_1$	$m_2$	$m_3$		$m_1$	$m_2$	$m_3$
	$w_1$	$w_2$	$w_3$	$\mu_2$	$w_2$	$w_3$	$w_1$	$\mu_3$	$w_3$	$w_1$	$w_2$
	1	2	4		1	5	5		1	4	4
	5	1	1		6	5	2		2	6	3
$\mu_4$	$m_1$	$m_2$	$m_3$		$m_1$	$m_2$	$m_3$		$m_1$	$m_2$	$m_3$
	$w_1$	$w_3$	$w_2$	$\mu_5$	$w_3$	$w_2$	$w_1$	$\mu_6$	$w_2$	$w_1$	$w_3$
	1	5	4		1	2	5		1	4	4

In the induced problem with preferences  $\succ^{\sigma^1}$ , the set of stable matchings is  $I^{\sigma^1} = \{\mu_6\}$  as  $\mu_1$  is blocked by  $(m_2, w_1)$ ,  $\mu_2$  is blocked by  $(m_2, w_2)$ ,  $\mu_3$  is blocked by  $(m_3, w_3)$ ,  $\mu_4$  is blocked by  $(m_3, w_1)$ , and  $\mu_5$  is blocked by  $(m_3, w_3)$ . So,  $\sigma^2(m_i, w_j) = \sigma^1(m_i, w_j)$  except for  $(m_2, w_1)$  and  $(m_3, w_3)$ . Thus,

$$\begin{aligned} \sigma^2(m_1, w_1) &= \{\mu_4\}, & \sigma^2(m_1, w_2) &= \{\mu_6\}, & \sigma^2(m_1, w_3) &= \{\mu_5\} \\ \sigma^2(m_2, w_1) &= \{\mu_3, \mu_6\}, & \sigma^2(m_2, w_2) &= \{\mu_1\}, & \sigma^2(m_2, w_3) &= \{\mu_2\} \\ \sigma^2(m_3, w_1) &= \{\mu_5\}, & \sigma^2(m_3, w_2) &= \{\mu_3\}, & \sigma^2(m_3, w_3) &= \{\mu_1, \mu_6\}. \end{aligned}$$

Hence, with  $\mu_{m_i}^{\sigma^2}(w_j)$  and  $\mu_{w_i}^{\sigma^2}(m_j)$  values, we have the following induced problem with preferences  $\succ^{\sigma^2}$ :

	5	5	3		2	1	2		6	3	1
	$m_1$	$m_2$	$m_3$		$m_1$	$m_2$	$m_3$		$m_1$	$m_2$	$m_3$
$\mu_1$	$w_1$	$w_2$	$w_3$	$\mu_2$	$w_2$	$w_3$	$w_1$	$\mu_3$	$w_3$	$w_1$	$w_2$
	1	2	2		1	5	5		1	4	4
	5	1	1		6	5	2		2	3	3
	$m_1$	$m_2$	$m_3$		$m_1$	$m_2$	$m_3$		$m_1$	$m_2$	$m_3$
$\mu_4$	$w_1$	$w_3$	$w_2$	$\mu_5$	$w_3$	$w_2$	$w_1$	$\mu_6$	$w_2$	$w_1$	$w_3$
	1	5	4		1	2	5		1	4	2

In the induced problem with preferences  $\succ^{\sigma^2}$ , the set of stable matchings is  $I^{\sigma^2} = \{\mu_1\}$  as  $\mu_2$  is blocked by  $(m_2, w_2)$ ,  $\mu_3$  is blocked by  $(m_3, w_3)$ ,  $\mu_4$  is blocked by  $(m_2, w_1)$ ,  $\mu_5$  is blocked by  $(m_3, w_3)$ , and  $\mu_6$  is blocked by  $(m_2, w_2)$ . So  $\sigma^3(m_i, w_j) = \sigma^2(m_i, w_j)$  except for  $(m_1, w_1)$ . Thus,

$$\begin{aligned} \sigma^3(m_1, w_1) &= \{\mu_4, \mu_1\}, & \sigma^3(m_1, w_2) &= \{\mu_6\}, & \sigma^3(m_1, w_3) &= \{\mu_5\} \\ \sigma^3(m_2, w_1) &= \{\mu_3, \mu_6\}, & \sigma^3(m_2, w_2) &= \{\mu_1\}, & \sigma^3(m_2, w_3) &= \{\mu_2\} \\ \sigma^3(m_3, w_1) &= \{\mu_5\}, & \sigma^3(m_3, w_2) &= \{\mu_3\}, & \sigma^3(m_3, w_3) &= \{\mu_1, \mu_6\} \end{aligned}$$

So with  $\mu_{m_i}^{\sigma^3}(w_j)$  and  $\mu_{w_i}^{\sigma^3}(m_j)$ , we have the following induced problem with preferences  $\succ^{\sigma^3}$ :

	1	5	3		2	1	2		6	3	1
	$m_1$	$m_2$	$m_3$		$m_1$	$m_2$	$m_3$		$m_1$	$m_2$	$m_3$
$\mu_1$	$w_1$	$w_2$	$w_3$	$\mu_2$	$w_2$	$w_3$	$w_1$	$\mu_3$	$w_3$	$w_1$	$w_2$
	1	2	2		1	5	5		1	4	4
	1	1	1		6	5	2		2	3	3
	$m_1$	$m_2$	$m_3$		$m_1$	$m_2$	$m_3$		$m_1$	$m_2$	$m_3$
$\mu_4$	$w_1$	$w_3$	$w_2$	$\mu_5$	$w_3$	$w_2$	$w_1$	$\mu_6$	$w_2$	$w_1$	$w_3$
	1	5	4		1	2	5		1	4	2

In the induced problem with preferences  $\succ^{\sigma^3}$ , the set of stable matchings is  $I^{\sigma^3} = \{\mu_1\}$  as  $\mu_2$  is blocked by  $(m_2, w_2)$ ,  $\mu_3$  is blocked by  $(m_3, w_3)$ ,  $\mu_4$  is blocked by  $(m_2, w_1)$ ,  $\mu_5$  is blocked by  $(m_3, w_3)$ , and  $\mu_6$  is blocked by  $(m_2, w_2)$ . Since for all  $(m_i, w_j) \in \mu_1, \mu_1 \in \sigma^3(m_i, w_j)$  we have  $S_2(m, w) = S_3(m, w) = S_4(m, w) = \dots$  for all  $(m, w) \in M \times W$ .

So we have  $\sigma(m, w) = \sigma^3(m, w)$ . And given  $\sigma$ , we have  $\mu_1 \in S_\sigma$ .

For the above example, although the universal and sophisticated estimation functions are different, they both give the same set of stable matchings (which has only  $\mu_1$  in it). One may then wonder whether this is in general true. The answer is negative. For the example given by [Sasaki and Toda \(1996\)](#) on p. 102, the set of stable

matchings given by universal estimations has three matchings  $\{\mu_1, \mu_4, \mu_6\}$ , whereas sophisticated expectations give the singleton set of stable matchings  $\{\mu_1\}$ . Note that if a matching is not stable with universal estimations, then it is not stable with sophisticated expectations. The example shows that the converse is not true; a matching can be stable with universal estimations but not stable with sophisticated expectations.

In the sophisticated expectations, we suppose that agents never drop a matching from  $\sigma^k$  once it is included in earlier rounds. If we suppose that agents drop matchings which are not stable  $(k - 1)$  st order induced problem from the  $k$ th order estimation  $\sigma^k$ , then we might obtain cycles, which do not give any estimation functions. Consider Example 2, and suppose that  $\mu_6$  is dropped from  $\sigma^3$ . Then we can check that  $\mu_1$  and  $\mu_6$  should be added to  $\sigma^4$ . This in turn implies that  $\mu_6$  is dropped from  $\sigma^5$ . Therefore, we obtain  $\sigma^5 = \sigma^3$  and a cycle occurs.

We suppose that agents cannot know each other's estimation functions for sure and they are cautious.<sup>9</sup> When the agents do not know each other's estimation functions, they can make guesses. We suppose that the agents consider the stable matchings of the reduced problem and (first order) induced problem to be plausible. And they know that other agents might also have these matchings in their estimations. The new estimations would give a new induced problem and cause new stable matchings to be added. Since they do not know others' estimations functions for sure and they are cautious, they do not drop matchings which are not stable in the new induced problem.

In the induced problem, we suppose that agents always care about the worst cases. Defining an induced problem supposing that agents care about the worst cases only after deviations would be very difficult, since there is no natural status quo. Nevertheless, a stable matching in the induced problem (which is in the estimation functions of its pairs) is an element of  $S_\varphi(M, W)$ . Therefore, it is natural to assume that the agents consider the stable matchings of the induced problem to be plausible.

Proposition 2 demonstrates that any matching which is both "considered as possible" by all pairs of the matching and "stable in the induced problem" is stable. The proposition also shows that it is not necessary for all agents to be equally sophisticated. It suffices that each agent be sophisticated enough to deduce that the set of stable matchings of the induced problem are possible ( $I^\varphi(m, w) \subset \varphi(m, w)$ ). This result may be used to find other plausible estimation functions than sophisticated expectations. While only an example, the sophisticated estimation functions are particularly useful nevertheless, because they can be constructed by using an *explicit* algorithm (via  $\sigma^k$  and  $I^{\sigma^k}$ ).

One natural question is to characterize minimal preference dependent estimation function that guarantees the nonemptiness of set of stable matchings. However, the answer is not obvious. First note that (because of cautious behavior) two estimations of a specific agent would be equivalent to each other if the worst matching for that

<sup>9</sup> If the agents knew each other's estimation functions for sure, then it would be natural for agents to have only matchings which are stable in their estimation functions. That is,  $\varphi(m, w) = S_\varphi(m, w) \equiv \{\mu \in A(m, w) : \mu \in S_\varphi(M, W)\}$ . However, there might be no estimation function satisfying this equality. For instance, in Example 2, the only stable matching could be—if there is any— $\mu_1$  (since with universal estimations,  $\mu_1$  is the only stable matching). But then,  $\varphi(m, w)$  is not even nonempty for  $(m, w) \notin \mu_1$ , for instance for  $(m_2, w_1)$ .

agent in the two estimation functions are the same. That is, the only relevant matching in an estimation is the worst matching while giving a blocking decision. Secondly, any estimation function giving more than one stable matching for some preference profile can be made smaller so as to give only one stable matching. If we think of estimation functions as functions of the preference profiles, it could be defined in any way wanted, and minimal estimation function would not give us any intuitive answer. The question is whether there are intuitively appealing estimation functions which are smaller than sophisticated expectations. We are, however, not aware of any such estimation functions. Sophisticated expectations is one intuitively appealing estimation function that guarantees the existence of stable matchings.<sup>10</sup>

#### 4 The core and the bargaining set

In marriage markets without externalities, the set of stable matchings and the core are equivalent. This is because blocking agents care only about their own matches. Put another way, in such marriage markets, a blocking coalition is not affected by the complementary coalition.

Once externalities are considered, however, the equivalence of the two notions is not immediate. A blocking coalition is affected by the complementary coalition. Because of this, the core notion also requires the use of estimation functions. One can define the estimation functions of agents when they form a (not necessarily a pair) coalition.

Ideally, one would like to define endogenously generated estimation functions for coalitions in the same manner as in that in the previous sections. But as an initial investigation, we assume (as do Sasaki and Toda 1996) that members of a blocking coalition “estimate” all matchings to be possible; that is, the estimation functions are universal. Universal estimations provide the best chance for the core to be nonempty—if the core with universal estimations is empty then it is empty with any other estimations. But Sasaki and Toda (1996) provide an example in which the core is empty with universal estimations.

The notion of a bargaining set is more permissive than that of the core [see the survey by Maschler (1992) and Klijn and Masso (2003)]. It requires that blocks which are not credible—in the sense that some members of the block have better block options—are ruled out. Below we extend the standard notion of a bargaining set to allow for externalities, again assuming universal estimations.

**Definition 2** The pair  $[(M', W'), \mu']$  **blocks** a matching  $\mu$  if coalitions  $M' \subseteq M$  and  $W' \subseteq W$ , with  $|M'| = |W'|$ ,  $\mu' \in A(M', W')$ , and for all  $\mu^* \in A(M - M', W - W')$ , we have

$$(\mu' \cup \mu^*) \succ_a \mu \quad \text{for all } a \in (M', W'),$$

where  $(\mu' \cup \mu^*)(a) = \mu'(a)$  for all  $a \in (M', W')$  and  $(\mu' \cup \mu^*)(a) = \mu^*(a)$  for all other  $a$ .

<sup>10</sup> As an anonymous referee pointed out, one could find different algorithms (similar to  $\sigma^k$ ) which also guarantee the existence of stable matchings.

In other words, a matching is **blocked** by a coalition of agents, if there is a matching that they can achieve themselves, such that whatever the complementary agents do, the members of the coalition are all better off. The **Core** is the set of all matchings which are not blocked.

**Definition 3** The pair  $[(M'', W''), \mu'']$  is a **counter block** against the block  $[(M', W'), \mu']$  at  $\mu$ , if both  $[(M'', W''), \mu'']$  and  $[(M', W'), \mu']$  block  $\mu$ ,  $(M'' \cup W'') \cap (M' \cup W') \neq \emptyset$ , and for all  $a \in (M'' \cup W'') \cap (M' \cup W')$ , there exists a  $\mu^a \in A(M - M', W - W')$  such that for all  $\mu^* \in A(M - M'', W - W'')$ , we have

$$(\mu'' \cup \mu^*) \succ_a (\mu' \cup \mu^a).$$

A block is not **credible** if it is counter-blocked. In other words, a block  $[(M', W'), \mu']$  is **counter-blocked by**  $[(M'', W''), \mu'']$  if all agents in the intersection  $(M'' \cup W'') \cap (M' \cup W')$  prefer the new block  $[(M'', W''), \mu'']$  to the old block  $[(M', W'), \mu']$ . The **bargaining set** is the set of all matchings such that every block is counter-blocked. Let us denote bargaining set of  $(M, W)$  by  $B(M, W)$

Even though the bargaining set is larger than the core, it too may be empty.

*Example 3* Suppose  $n = 3$  and consider the following utility assignments:

	5	5	5		6	6	1		3	3	2
$\mu_1$	$m_1$	$m_2$	$m_3$	$\mu_2$	$m_1$	$m_2$	$m_3$	$\mu_3$	$m_1$	$m_2$	$m_3$
	$w_1$	$w_2$	$w_3$		$w_2$	$w_3$	$w_1$		$w_3$	$w_1$	$w_2$
	5	3	5		4	6	6		3	3	6
	4	1	3		1	4	6		2	2	4
$\mu_4$	$m_1$	$m_2$	$m_3$	$\mu_5$	$m_1$	$m_2$	$m_3$	$\mu_6$	$m_1$	$m_2$	$m_3$
	$w_1$	$w_3$	$w_2$		$w_3$	$w_2$	$w_1$		$w_2$	$w_1$	$w_3$
	4	2	5		1	2	1		1	2	4

The matching  $\mu_1$  has only one block,  $[(m_1, w_2), (m_2, w_3)]$ . The matching  $\mu_2$  also has only one block,  $[(m_3, w_2)]$ . The matching  $\mu_3$  has three blocks,  $[(m_1, w_1)]$ ,  $[(m_3, w_3)]$  and  $[(m_1, w_1), (m_3, w_3)]$ . Among these, the last one has no counter blocks. The matching  $\mu_4$  has two blocks,  $[(m_3, w_3)]$  and  $[(m_1, w_1), (m_3, w_3)]$ . Among these the latter has no counter blocks. The matching  $\mu_5$  has three blocks,  $[(m_1, w_1)]$ ,  $[(m_1, w_1), (m_2, w_2)]$  and  $[(m_1, w_2), (m_2, w_3)]$ . Among these the last one has no counter blocks. The matching  $\mu_6$  has 8 blocks:  $[(m_1, w_1), (m_2, w_2), (m_3, w_3)]$  and all 1 and 2 combinations, and  $[(m_1, w_2), (m_2, w_3)]$ . Among these the last one has no counter blocks. Hence, the bargaining set (and hence, also the core) is empty.

In the example, different agents consider different matchings to be the “worst”. As the following proposition shows, if all agents agree on the worst matching, then the bargaining set is nonempty.<sup>11</sup>

<sup>11</sup> Although the existence of a unanimously worst matching assumption is a strong one, it can be satisfied in matching models in which both sides of the market have some characteristics that a match is good only if both partners have the same characteristics. For instance, matching software engineers with law firms and lawyers with software firms would be the worst matching for all agents in the market.

**Proposition 4** *If there exists a matching which is the worst for all agents, then  $B(M, W) \neq \emptyset$ .*

*Proof* Let there be  $n$  couples and let  $\underline{\mu}$  denote the worst matching (every agent prefers any other matching to the matching  $\underline{\mu}$ ). We show that if no matching  $\mu$  other than  $\underline{\mu}$  is in  $B(M, W)$ , then  $\underline{\mu} \in B(M, \bar{W})$ . Suppose that no matching  $\mu \neq \underline{\mu}$  is in the bargaining set. Note that, for all nonempty coalitions  $(M', W')$  with  $|M'| = |W'|$ , for all matchings  $\mu' \in A(M', W')$  such that  $\mu'$  is not equal to the projection of  $\underline{\mu}$ —that is, for some  $a \in M' \cup W'$ ,  $\mu'(a) \neq \underline{\mu}(a)$ ,  $[(M', W'), \mu']$  blocks  $\underline{\mu}$ . Below, we show that there is a counter-block for every block against  $\underline{\mu}$ .

First, note that for all blocks against  $\underline{\mu}$  involving fewer than  $n - 1$  couples, there is a counter-block. This is because the agents who are not in the blocking set can form more than one matching among themselves, and all possible matchings exert a negative externality on the blocking agents. Therefore, by fixing a matching among themselves, the nonblocking agents make the blocking agents better off, and since the matching they are blocking is the worst matching for all agents, they would be members of blocking coalition for the new block. Formally, suppose  $|M'| < n - 1$  and  $[(M', W'), \mu']$  blocks  $\underline{\mu}$ , then  $[(M, W), \mu]$  (with  $|M| = n$ ) is a counter-block against  $[(M', W'), \mu']$  for some  $\mu$  with  $\mu(a) = \mu'(a)$  for all  $a \in M' \cup W'$ .

Secondly, consider a block against  $\underline{\mu}$  involving  $n$  couples,  $[(M, W), \mu]$ . From our assumption,  $\mu$  has a block, say  $[(M', \bar{W}'), \mu']$ . Then, for all  $a \in M' \cup W'$ , we have  $(\mu' \cup \mu'') \succ_a \mu$  for all  $\mu'' \in A(M - M', W - W')$ . But then,  $[(M', W'), \mu']$  is a counter-block against  $[(M, W), \mu]$  at  $\underline{\mu}$ . This is because,  $[(M', W'), \mu']$  is a block against  $\underline{\mu}$  and all agents who are in both of the blocks  $[(M, W), \mu]$  and  $[(M', W'), \mu']$  are better off in the second block. The analysis for a block against  $\underline{\mu}$  involving  $n - 1$  couples is exactly like the  $n$  couples case, since when  $n - 1$  couples block, the non-blocking agents can form only one matching. Hence, the resulting blocking matching is unique, as in the  $n$  couples case. □

Interestingly, there are problems in which *only* the worst matching  $\underline{\mu}$  is in the bargaining set. Consider the following example.

*Example 4* Suppose  $n = 3$  and consider the following utility assignments:

	3	2	3		5	3	5		2	5	6
$\mu_1$	$m_1$	$m_2$	$m_3$		$m_1$	$m_2$	$m_3$		$m_1$	$m_2$	$m_3$
	$w_1$	$w_2$	$w_3$	$\mu_2$	$w_2$	$w_3$	$w_1$	$\mu_3$	$w_3$	$w_1$	$w_2$
	6	2	2		4	4	5		6	3	6
	4	6	4		6	4	2		1	1	1
$\mu_4$	$m_1$	$m_2$	$m_3$		$m_1$	$m_2$	$m_3$		$m_1$	$m_2$	$m_3$
	$w_1$	$w_3$	$w_2$	$\mu_5$	$w_3$	$w_2$	$w_1$	$\mu_6$	$w_2$	$w_1$	$w_3$
	4	3	3		5	5	2		1	1	1

Observe that  $\mu_6$  is the worst matching for every agent. Matchings other than  $\mu_6$  are not in the bargaining set:  $\mu_1$  has five blocks;  $[(m_2, w_3)]$ ,  $[(m_3, w_2)]$ ,  $[(m_1, w_2), (m_2, w_3)]$ ,  $[(m_1, w_3), (m_2, w_2)]$ , and  $[(m_2, w_3), (m_3, w_2)]$ . The last two do not have any counter

blocks. Each of the matchings  $\mu_2, \mu_3, \mu_4$  and  $\mu_5$  has one block;  $[(m_1, w_3), (m_2, w_2)], [(m_1, w_1)], [(m_1, w_2), (m_3, w_1)],$  and  $[(m_2, w_1), (m_3, w_2)]$  respectively. Every block of  $\mu_6$  has a counter-block. Hence,  $\mu_6$  is the only matching in the bargaining set.

The example illustrates a phenomenon associated with the notion of a bargaining set. A Pareto inefficient outcome—in this case, the worst matching—is the only one that survives the “credible” block test.

## 5 Concluding remarks

When externalities are present, the expectations that agents in a coalition hold regarding the complementary coalition are crucial in determining what outcomes are stable. Marriage problems with externalities seem to be a relatively simple context in which one may begin to explore what kinds of expectations are natural.

In this paper, we have presented a model in which the expectations that agents in a (pairwise) coalition hold regarding the complementary coalition are endogenously determined—that is, they are consistent with the preferences of agents in the complementary coalition. Sasaki and Toda (1996) showed that rational expectations are, in general, incompatible with the existence of a set of stable matchings. They also showed that exogenous expectations are incompatible (except when they are all inclusive). We have defined a notion of sophisticated expectations that have the following features: (a) they are endogenously generated; (b) they are not all inclusive; and (c) they lead to a nonempty set of stable matchings. We have also identified a general condition on estimation functions that is sufficient to guarantee a nonempty set of stable matchings.

We can reinterpret our result in terms of a rationality of the agents as follows.<sup>12</sup> If the blocking agents are rational and believe that other agents are rational too, they should not worry about a negative externality exerted on them by another man-woman couple that would never have wanted to pair in the first place. We recursively construct matchings in the reduced problem that pessimistic but rational agents may worry about and show that the set of stable matchings is nonempty.

It remains for future work to see if these ideas can, perhaps, be extended to other coalitional settings.

## References

- Bloch F (1996) Sequential formation of coalitions in games with externalities and fixed payoff division. *Games Econ Behav* 14:90–123
- Gale D, Shapley L (1962) College admission and the stability of marriage. *Am Math Monthly* 69:9–15
- Klijn F, Masso J (2003) Weak stability and a bargaining set for the one-to-one matching model. *Games Econ Behav* 42:91–100
- Li S (1993) Competitive matching equilibrium and multiple principle-agent models. University of Minnesota, Mimeo
- Maschler M (1992) The bargaining set, kernel and nucleolus: a survey. In: Aumann R, Hart S (eds) *Handbook of game theory*. Elsevier, Amsterdam
- Maskin E (2003) Bargaining, coalitions and externalities. Working Paper, Institute for Advanced Study

<sup>12</sup> We are grateful to an anonymous referee for suggesting this interpretation.

- Ray D, Vohra R (1999) A theory of endogenous coalition structures. *Games Econ Behav* 26:286–336
- Roth A, Sotomayor M (1990) *Two-sided matching: a study in game theoretic modelling and analysis*. Cambridge University Press, Cambridge
- Roy Chowdhury P (2005) Marriage markets with externalities. In: Mohan SR, Neogy SK (eds) *Operations research with economic and industrial applications: emerging trends*. Anamaya Publishers, New Delhi, pp 148–155
- Sasaki H, Toda M (1996) Two-sided matching problems with externalities. *J Econ Theory* 70:93–108