Theoretical Economics 17 (2022), 1619–1649 1555-7561/20221619

# Stable matching under forward-induction reasoning

Luciano Pomatto
Division of the Humanities and Social Sciences, California Institute of Technology

A standing question in the theory of matching markets is how to define stability under incomplete information. This paper proposes an epistemic approach. Agents negotiate through offers, and offers are interpreted according to the highest possible degree of rationality that can be ascribed to their proponents. A matching is deemed "stable" if maintaining the current allocation is a rationalizable action for each agent. The main result shows an equivalence between this notion and "incomplete-information stability," a cooperative solution concept put forward by Liu, Mailath, Postlewaite, and Samuelson (2014) for markets with incomplete information.

Keywords. Matching, incomplete information, stability.

JEL classification. C78.

## 1. Introduction

Over the past decades, a vast literature has substantially broadened our conceptual understanding of matching markets. Much of the existing literature assumes complete information, that is, that the value of a matching is entirely known to the relevant parties. However, incomplete information is arguably commonplace in most environments.

The crucial difficulty in the study of matching markets with incomplete information lies in the notion of stability. Consider a job market where workers and firms are matched. Under complete information, a matching is *stable* if no pair of workers and firms are willing to reject the existing match to form more profitable partnerships. Consider now a market where there is uncertainty about the profitability of partnerships. Whether or not to leave the existing match is now a complex decision. One reason is that the actions taken to exit the default allocation (starting a negotiation, proposing an agreement, etc.) will typically reveal something about the parties involved. Another reason is that if the matching is to be deemed "stable,"then such actions should be unexpected. Hence, agents must revise their beliefs based on zero probability events. So, under incomplete information, a theory of stability must also incorporate a theory of beliefs.

This paper considers an epistemic approach to matching markets with incomplete information. We study a class of markets with transferable utility where agents on one

Luciano Pomatto: luciano@caltech.edu
I am indebted to Alvaro Sandroni and Marciano Siniscalchi for many useful discussions. I would like to the thank the referees for their comments and suggestions. I also thank Pierpaolo Battigalli, Yi-Chun Chen, Willemien Kets, George Mailath, Larry Samuelson, Rakesh Vohra, and several seminar audiences for helpful comments on earlier drafts.

© 2022 The Author. Licensed under the Creative Commons Attribution-NonCommercial License 4.0. Available at https://econtheory.org. https://doi.org/10.3982/TE3867

side of the market (e.g., workers) have private information about their characteristics (e.g., their skills), which are payoff-relevant for both sides. Each worker is assumed to know her payoff-type and each firm knows the type of the worker it is matched to. Notably, agents are not required to share a common prior. Instead, workers' beliefs are assumed to satisfy a simpler "grain of the truth" assumption, which postulates that agents assign at least positive probability to the actual profile of payoff-types.

A default allocation is given. It specifies how workers are matched to firms and at what wages. Firms have the opportunity to negotiate away from the current allocation. Negotiation is modeled as a noncooperative game and occurs through take-it-or-leave-it offers. If no offers are made, or all offers are rejected, then the default allocation is implemented. The approach taken in this paper is deliberately in between cooperative and noncooperative. As in the classical study of stability and the core, we abstract away from the process by which a certain allocation is formed. At the same time, to formalize players' beliefs and thought processes, we model deviations from a given allocation through a noncooperative game.

Consider a firm who receives an offer from another agent, named Ann. The firm cannot know with certainty whether accepting the offer is profitable. It must reach this decision by updating its belief about Ann's characteristics from the fact that she made an offer. Intuitively, it faces questions such as: what must be true about Ann for her to make this offer? What can we infer about her from the fact she is the only one who made an offer, and so forth. The approach taken in this paper is to follow the idea that offers are interpreted according to the highest degree of sophistication that can be ascribed to those who make them. This is formalized by assuming that players behave accordingly to a notion of extensive-form rationalizability (Pearce (1984)) for dynamic games with incomplete information due to Battigalli (2003) and Battigalli and Siniscalchi (2003), strong $\Delta$-rationalizability. Stability is defined by imposing three requirements on players' actions and beliefs. Informally:

1. Agents are rational and abstain from making offers;
2. Players expect no offer to be made by other agents; and
3. In case a player deviates and makes an offer, the offer is interpreted according to the highest degree of strategic sophistication that can be ascribed to its proponent.

If all three requirements are satisfied, then the default allocation is said to be *stable under forward induction*. Rationality is defined by requiring players' actions to be optimal (given their beliefs) at every history they act. Requirement (2) is formalized by the assumption that players assign probability 1, at the beginning of the game, to the event that other players will not make offers.

The third requirement is crucial and it is formalized through an iterative definition. Each player expects others to be rational and also expects others to believe, ex-ante, that no offer will be made. This belief is held at the beginning of the game and conditional on any offer, provided that the offer does not provide decisive proof against it. As a further step in their thought process, agents expect other players to believe in their opponents rationality and their surprise upon observing an offer. This iteration progresses through

higher orders. Each step leads players to rationalize the observed behavior according to a higher degree of sophistication. Requirement (3) is formalized by taking the limit of this iteration.

The main result of this paper, Theorem 1, characterizes the set of matching outcomes that are stable under forward induction. It shows that a matching outcome is stable under forward induction if and only if it is *incomplete-information stable*, a cooperative notion introduced by Liu, Mailath, Postlewaite, and Samuelson (2014). This notion satisfies two fundamental properties: existence and efficiency under standard supermodularity conditions. Through an explicit epistemic characterization of incomplete-information stability, the result provides a clearer understanding of what types of reasoning can lead to stability and efficiency in matching markets. At the same time, the paper highlights some important differences between the cooperative approach and the current approach based on forward-induction reasoning.

One such difference lies in the type of informational assumptions. Liu et al. (2014) assume that the matching and the profile of wages are common knowledge. In this paper, we make the weaker hypothesis that workers' beliefs about other agents' payoff-types, matches and wages, assign positive probability to the actual realization. A second important difference is in the criterion by which firms evaluate risk. A strict interpretation of incomplete-information stability suggests that firms evaluate a potential match with a worker of unknown type according to the worst-case payoff. In this paper, agents are assumed to be expected utility maximizers.

### 1.1 *Related literature*

This paper is linked to several strands of the literature. Starting with Wilson (1978), notions of core under incomplete information have been introduced by Vohra (1999), Dutta and Vohra (2005), Serrano and Vohra (2007), de Clippel (2007), Myerson (2007), and Peivandi (2013), among others. The current paper shares some similarities with Serrano and Vohra (2007), where blocking coalitions are formed noncooperatively, as equilibrium outcomes of a voting game.

A number of papers have studied matching under incomplete information.[1] Chakraborty, Citanna, and Ostrovsky (2010) study markets with one-sided incomplete information and interdependent valuations. They study a model where agents on one side of the market (colleges) receive informative signals about the quality of the agents on the opposite side (students). After a matching is realized, colleges can make rematching offers to students, and a matching is deemed stable if there exists a Bayesian Nash equilibrium in which colleges abstain from making offers. One important difference with the framework of this paper is that stability is defined as a property of a centralized mechanism producing the matching allocation, rather than a property of a matching outcome. A second important difference is in the choice of solution concept (Bayesian Nash equilibrium instead of rationalizability).

[1] See, among others, Roth (1989), Chade (2006), Ehlers and Massó (2007), Hoppe, Moldovanu, and Sela (2009), and Chade, Lewis, and Smith (2014).

Bikhchandani (2017) extends the analysis of Liu et al. (2014) to markets without transferable utility. The paper presents a notion of "Bayesian stability" for markets with two-sided incomplete information. This notion presupposes a common prior over types.

Chen and Hu (2019) provide an alternative foundation for incomplete-information stability. They adopt a partitional model, and assume that agents optimize according to a max-min criterion. Stability is formulated as the joint requirement that a matching is not blocked and that the absence of blocking pairs does not reveal any new information. In addition, they establish that any dynamic process that allows randomly chosen blocking pairs to rematch will converge to a stable allocation.

Liu (2020) introduces a cooperative notion of stability that captures some important concepts of game-theoretic equilibrium analysis. Unlike in Liu et al. (2014), agents share a common prior and stability is formulated as a property of a matching function mapping types profiles to allocations. Agents update their beliefs upon the realization of a matching allocation. Beliefs are subsequently updated when participating to a blocking pair, in the event where one occurs. The paper studies different criteria of beliefs updating.

This paper builds upon the literature on forward-induction reasoning. Extensive form rationalizability was introduced in Pearce (1984), while the best rationalization principle was first formalized in Battigalli (1996). Common strong belief in rationality was defined and characterized in Battigalli and Siniscalchi (2002), and in Battigalli and Siniscalchi (2003) for games with payoff uncertainty. The implications of common strong belief in rationality are also studied in Battigalli and Friedenberg (2013) and Battigalli and Prestipino (2013).

This paper is also related to the literature on forward-induction refinements of equilibrium concepts in signaling games, where forward-induction inferences made upon observing a message are based on a candidate equilibrium outcome that is a priori expected by the players. This sort of logic plays an important role in Banks and Sobel (1987) and Cho and Kreps (1987). It is also at the core of the work of Sobel, Stole, and Zapater (1990). Their paper applies extensive form rationalizability to signaling games by replacing a given equilibrium path with an action for the sender, which yields the equilibrium payoff to all players. Whether or not the "equilibrium" action is rationalizable in the modified game is shown to depend on whether the equilibrium survives the iterated intuitive criterion. Another related paper is Battigalli and Siniscalchi (2003), where it is shown that common strong belief in rationality and in a fixed distribution over terminal nodes of a signaling game characterizes self-confirming equilibria satisfying the iterated intuitive criterion.

There are several significant differences between this paper and the contributions of Sobel, Stole, and Zapater (1990) and Battigalli and Siniscalchi (2003). In this paper, players have heterogeneous beliefs over the payoff-types of the informed players. In addition, unlike signaling games, the information structure of the blocking game we consider in this paper does not have a product structure. Finally, the main result of the paper, the characterization of Theorem 2, does not share similarities with other results in the literature. The idea that players may rationalize past behavior has a long history

in game theory. The idea of forward induction goes back to Kohlberg (1981). Solution concepts expressing different forms of forward induction were introduced in Kohlberg and Mertens (1986), Banks and Sobel (1987), Cho and Kreps (1987), Van Damme (1989), Reny (1992), Govindan and Wilson (2009), and Man (2012), among others.

## 2. Two-sided matching markets

We consider a two-sided matching environment with transferable utility, following Crawford and Knoer (1981) and Liu et al. (2014). A set of *agents* is divided in two groups, denoted by $I$ and $J$. For concreteness, $I$ is referred to as the set of *workers* and $J$ as the set of *firms*. We assume $|I| \geq 2$. Each worker is endowed with a *payoff-type* belonging to a finite set $W$. Each firm $j \in J$ is also endowed with a payoff-type belonging to a finite set $F$. We denote by $\mathbf{w} \in W^I$ and $\mathbf{f} \in F^J$ the corresponding *profiles* of attributes.

A *matching function* is a map $\mu : I \to J \cup \{\emptyset\}$ that is injective on $\mu^{-1}(J)$. If $\mu(i) = j$, then worker $i$ is hired by firm $j$. If $\mu(i) = \emptyset$, then worker $i$ is unemployed. Similarly, if $\mu^{-1}(j) = \emptyset$ then no worker is hired by firm $j$. A worker is assigned to at most one firm and a firm can hire at most one worker.

A match between a worker of type $w$ and a firm of type $f$ gives rise, in the absence of monetary transfers, to a payoff of $\nu(w, f)$ for the worker and $\phi(w, f)$ for the firm. Following Mailath, Postlewaite, and Samuelson (2013), we refer to $\nu$ and $\phi$ as *premuneration values*. The premuneration values of an unmatched worker or firm is equal to 0. To have a unified notation for both matched and unmatched agents, let $\nu(w, \mathbf{f}_\emptyset) = 0$ for every $w \in W$ and $\phi(\mathbf{w}_\emptyset, f) = 0$ for every $f \in F$.

Associated to a matching function is a payment scheme $\mathbf{p}$ specifying for each pair $(i, \mu(i))$ of matched agents a transfer $\mathbf{p}_{i,\mu(i)} \in \mathbb{R}$ from firm $\mu(i)$ to worker $i$. Unmatched workers receive no payments. We use the notation $\mathbf{p}_{i,\emptyset} = \mathbf{p}_{\emptyset,j} = 0$ for every $i$ and $j$. Under the matching $\mu$ and payment scheme $\mathbf{p}$, the utility of worker $i$ and firm $j$ is given by

$$\nu(\mathbf{w}_i, \mathbf{f}_{\mu(i)}) + \mathbf{p}_{i,\mu(i)} \quad \text{and} \quad \phi(\mathbf{w}_{\mu^{-1}(j)}, \mathbf{f}_j) - \mathbf{p}_{\mu^{-1}(j),j},$$

respectively.

A *matching outcome* is a tuple $(\mathbf{w}, \mathbf{f}, \mu, \mathbf{p})$ specifying workers' and firms' payoff-types and an *allocation* $(\mu, \mathbf{p})$ consisting of a matching function and a payment scheme. A matching outcome is individually rational if it provides nonnegative payoffs to all workers and firms.

A *default allocation* or *status quo*, $(\mu, \mathbf{p})$ is given. Agents have the opportunity to negotiate and abandon the status quo in favor of new partnerships, but if no agreement is reached, then the default allocation remains in place. If the matching outcome $(\mathbf{w}, \mathbf{f}, \mu, \mathbf{p})$ is common knowledge, then this is the setting studied by Shapley and Shubik (1971) and Crawford and Knoer (1981). In this case, a matching outcome $(\mathbf{w}, \mathbf{f}, \mu, \mathbf{p})$ is *complete-information stable* if it is individually rational and there is no worker $i$, firm $j$ and payment $q$ such that

$$\begin{aligned} &\nu(\mathbf{w}_i, \mathbf{f}_j) + q > \nu(\mathbf{w}_i, \mathbf{f}_{\mu(i)}) + \mathbf{p}_{i,\mu(i)} \quad \text{and} \\ &\phi(\mathbf{w}_i, \mathbf{f}_j) - q > \phi(\mathbf{w}_{\mu^{-1}(j)}, \mathbf{f}_j) - \mathbf{p}_{\mu^{-1}(j),j}. \end{aligned}$$

As shown by Shapley and Shubik (1971), for any profiles $\mathbf{w}$ and $\mathbf{f}$ there always exists an allocation with the property that the resulting matching outcome is complete-information stable, and every stable outcome is efficient.[2]

### 2.1 *Incomplete information*

The standard framework is now altered by relaxing the assumption of complete information. We consider markets where agents have only partial information regarding other agents' types as well as the current allocation. We study markets with one-sided, interim, incomplete information.

We are given a finite set $\mathbf{M}$ of possible matching outcomes. We refer to $\mathbf{M}$ as the *market*. For simplicity, each $\mathbf{m} \in \mathbf{M}$ is assumed to be individually rational. Players' information about the matching outcome is modeled as a profile $(\mathcal{P}_k)_{k \in I \cup J}$ of information partitions on $\mathbf{M}$. For every $\mathbf{m} \in \mathbf{M}$ and player $k$, we denote by $\mathcal{P}_k(\mathbf{m}) \subseteq \mathbf{M}$ the information available to $k$ when the actual outcome is $\mathbf{m}$.

Fix a matching outcome $\mathbf{m} = (\mathbf{w}, \mathbf{f}, \mu, \mathbf{p}) \in \mathbf{M}$. For every firm $j$, we assume

$$\mathcal{P}_j(\mathbf{m}) = \left\{(\tilde{\mathbf{w}}, \mathbf{f}, \mu, \mathbf{p}) \in \mathbf{M} : \tilde{\mathbf{w}}_{\mu^{-1}(j)} = \mathbf{w}_{\mu^{-1}(j)}\right\}.$$

Hence, each firm knows the current profile $\mathbf{f}$ of firms' types, the allocation $(\mu, \mathbf{p})$, and the type of the worker it is matched to, if any. Workers, in contrast, are only required to possess minimal information about the environment. For every worker $i$, define

$$\mathcal{P}_i^*(\mathbf{m}) = \left\{(\tilde{\mathbf{w}}, \mathbf{f}, \tilde{\mu}, \tilde{\mathbf{p}}) \in \mathbf{M} : \tilde{\mathbf{w}}_i = \mathbf{w}_i,\ \tilde{\mu}(i) = \mu(i) \text{ and } \tilde{\mathbf{p}}_{i,\mu(i)} = \mathbf{p}_{i,\mu(i)}\right\}.$$

That is, under the information partition $\mathcal{P}_i^*$, each worker $i$ knows the profile $\mathbf{f}$, her payoff-type $\mathbf{w}_i$, her match $\mu(i)$, and wage. We assume that for each worker, her information partition $\mathcal{P}_i$ satisfies $\mathcal{P}_i(\mathbf{m}) \subseteq \mathcal{P}_i^*(\mathbf{m})$ for every $\mathbf{m}$. That is, $\mathcal{P}_i^*$ is a *lower bound* on the amount of information available to $i$. This allows for a fairly general formulation.

In addition to the information specified by the partitions, agents entertain probabilistic beliefs about what they do not know. Beliefs will be described in Section 4.

## 3. The blocking game

This section introduces a simple noncooperative game by which players negotiate over new partnerships to abandon the status quo allocation. Negotiations occur through take-it-or-leave-it offers.

[2]As is well known, if a matching outcome is stable, then the allocation belongs to the core. This equivalence has no obvious counterpart under incomplete information. Going from pairwise stability to group stability raises a number of issues, the main one being what information agents are allowed to share within a coalition (Wilson (1978)). An additional open question is how to model forward-induction reasoning in the context of coalitional deviations.

### 3.1 *Model*

In this noncooperative game, the set of players is $I \cup J$. We are given a matching outcome $\mathbf{m} = (\mathbf{w}, \mathbf{f}, \mu, \mathbf{p})$ belonging to $\mathbf{M}$. The game is played in two stages. In each stage, actions are played simultaneously and the game has observable actions; hence, first-stage choices become public information at the beginning of the second stage.

In the first stage, each worker $i$ can *abstain* or make an offer $(j, q)$, where $j$ is a firm other than $\mu(i)$ and $q$ belongs to $Q$, a fixed finite subset of $\mathbb{R}$. Informally, an offer $(j, q)$ means that worker $i$ is willing to break the status quo and form a new partnership with firm $j$ at a wage $q$.

The assumption of a discrete currency $Q \subseteq \mathbb{R}$ will avoid introducing measurability assumptions on the strategies of the firms. It will also guarantee the existence of optimal strategies for any choice of workers' beliefs. We assume $Q$ to be a sufficiently fine grid.[3] The results are not sensitive to the particular specification of $Q$.

In the second stage, each firm that has received at least one offer chooses between rejecting all offers or accepting one.

Payoffs are defined as follows. For every offer $(j, q)$ by worker $i$ that has been accepted, call the resulting combination $(i, j, q)$ a *blocking offer*. For every blocking offer $(i, j, q)$, worker $i$ is matched to firm $j$ at a wage $q$ and the two agents receive payoffs $\nu(\mathbf{w}_i, \mathbf{f}_j) + q$ and $\phi(\mathbf{w}_i, \mathbf{f}_j) - q$, respectively. If worker $i$ is not part of a blocking offer but $\mu(i)$ is, then $i$ receives a payoff of 0 (i.e., $i$ becomes unmatched). Similarly, if firm $j$ is not part of a blocking offer but $\mu^{-1}(j)$ is then $j$ receives a payoff of 0. All the other agents remain matched according to the original allocation $(\mu, \mathbf{p})$ and obtain the corresponding payoffs.

### 3.2 *Discussion*

The game has two features that play an important role in the analysis. The first is that offers are binding: an offer that is accepted is immediately implemented. The second is that inaction preserves the status quo. That is, if no offers are made then the original allocation $(\mu, \mathbf{p})$ is applied. Both features make the game close in spirit to assumptions that are implicit in the interpretation of the core under complete information (see, for instance, the discussion in Myerson (1997)).

It should be emphasized that in this game incomplete information is analyzed at the interim stage. In particular, there is no ex ante stage at which workers plan their actions conditional on every realized matching outcome.

We now introduce some auxiliary notation that will be useful in what follows. Let $H$ denote the set of all nonterminal histories and denote by $\emptyset$ the empty (or *initial*) history. Each history $h \in H$ other than $\emptyset$ describes what offers, if any, have been made and to what firms. For each firm, denote by $H_j$ the set of histories where $j$ has received at least one offer. A strategy of worker $i$ is an element $s_i \in \{a\} \cup (J \setminus \mu(i)) \times Q$, where $a$ corresponds to abstaining from making offers. A strategy of firm $j$ is represented by a function $s_j : H_j \to I \cup \{r\}$, where $r$ corresponds to rejecting all offers received by $j$, and $s_j$ has the

[3] See Section A.1 for a formal statement of this assumption.

property that if $s_j(h) \neq r$ then $s_j(h)$ belongs to the set of workers who made an offer to $j$ at history $h$. The set of strategies of each player $k$ is denoted by $S_k$. For every history $h$ and player $k$, we denote by $S_{-k}(h)$ the set of strategies in $S_{-k}$ that lead to history $h$ for some $s_k \in S_k$.

## 4. Example

In this section, we present an example that illustrates some of the ideas underlying the main result. In this example, both stability under forward induction and incomplete-information stability will lead to the conclusion that a given matching outcome is not stable. However, the sort of iterative reasoning described by the two procedures will be qualitatively different.

There are two workers, $a$ and $b$, and two firms, $A$ and $B$. A match between a worker of type $w \in \mathbb{R}$ and a firm of type $f \in \mathbb{R}$ leads to premuneration values $\phi(w, f) = \nu(w, f) = w \cdot f$. The market consists of the two matching outcomes $\mathbf{m_1}$ and $\mathbf{m_2}$, described in Figure 1. In both outcomes, worker $a$ is matched to firm $A$ and worker $b$ to firm $B$. Worker $a$'s type is 1 and firm $A$'s type is 2, and the two are matched at a wage of 0. In this market, the only uncertainty is about the type of worker $b$, which in outcome $\mathbf{m_2}$ is equal to $b_2 = 2$ and in outcome $\mathbf{m_1}$ it is equal to $b_1 = 1$.

*Incomplete-information stability* The matching outcome $\mathbf{m_2}$ is not incomplete-information stable. The first iteration eliminates the outcome $\mathbf{m_1}$. The reason is that worker $b$ and firm $A$ can form a blocking pair at transfer $q = -1/2$. Such a blocking pair increases $b_1$'s payoff from 0 to 3/2, and increases $A$'s payoff from 2 to 5/2 (obviously, it would increase $A$'s payoff even if her type was $b_2$).

In its second iteration, incomplete-information stability stipulates that a firm, when part of a blocking pair, evaluates a deviation from the current matching by restricting attention to matching outcomes that have not been previously eliminated. Having ruled out the outcome $\mathbf{m_1}$ in the first step, in the second step the outcome $\mathbf{m_2}$ is now eliminated as well by considering a blocking pair between worker $b_2$ and firm $A$ at a transfer

| | | | | |
|---|---|---|---|---|
| worker: | $a$ | $b$ | $a$ | $b$ |
| worker payoffs: | 2 | 0 | 2 | 4 |
| worker types, **w**: | 1 | 1 | 1 | 2 |
| payment, **p**: | 0 | $-4$ | 0 | $-4$ |
| firm types, **f**: | 2 | 4 | 2 | 4 |
| firm payoffs: | 2 | 8 | 2 | 12 |
| firm: | $A$ | $B$ | $A$ | $B$ |
| matching outcome: | $\mathbf{m_1}$ | | $\mathbf{m_2}$ | |

Figure 1. A market consisting of two matching outcomes, $\mathbf{m_1}$ and $\mathbf{m_2}$. Workers and firms are ordered by columns. In both outcomes, worker $a$ is matched to firm $A$, and worker $b$ to firm $B$, and the types of $a$, $A$, and $B$ are, respectively, 1, 2, and 4.

$q \in (0, 2)$. Any such transfer provides the worker with a payoff strictly greater than 4, and the firm with a payoff strictly greater than 2.

Key to the argument is the inference made by firm $A$ and worker $b_2$ when the two agents are involved in a candidate blocking pair at a proposed transfer $q \in (0, 2)$. Incomplete-information stability suggests the following line of reasoning for firm $A$: "Suppose worker $b$ were of type 1. This would invalidate the assumption that the matching is stable, since $b$ could have formed an "obvious" blocking pair, which would have increased her payoff, and would have increased my payoff regardless of her type. Hence, $b$'s type must be 2. So, I agree to break the current matching and match with $b$."

*The role of beliefs* If blocking pairs are formed through an explicit negotiation, then we encounter a difficulty in formalizing the inference described in the previous paragraph: It is not intuitively obvious whether worker $b$, if of low type, would indeed choose to make a low offer such as $q = -1/2$. While such an offer would be accepted by firm $A$ regardless of $A$'s beliefs, the worker could demand a higher wage from $A$ in the hope of being mistaken for a high type.

To see this, suppose $b$'s type is $b_2$, and consider an offer to $A$ at wage $q$. A wage $q > 2$ would lead to the offer being rejected with certainty. A wage $q < 0$ would make the deviation unprofitable for the worker even if the offer was accepted. Hence, the range $q \in (0, 2)$ describes all wages that type $b_2$ could conceivably offer to $A$.

Whether the low type would be more likely to make a low rather than a high offer should, intuitively, depends on her beliefs about the behavior of firm $A$. This issue does not arise under incomplete-information stability, since beliefs do not enter explicitly in its definition.

*Stability under forward induction* We now analyze the example above by applying stability under forward induction. For this example, it is enough to consider a simplified version of the blocking game. We assume, without loss of generality, that worker $a$ can only abstain from making offers. Worker $b$ has three possible actions: she can either abstain, she can make an offer to firm $A$ at a "low" wage $\underline{q} = -1/2$, or she can make an offer to firm $A$ at a "high" wage $\overline{q} = 1$. Firm $A$ has four possible strategies: conditional on receiving an offer $q \in \{\underline{q}, \overline{q}\}$, the firm can accept it, denoted by $accept(q)$, or reject it, denoted by $reject(q)$. Thus, a market-strategy pair is a tuple that describes the type of worker $b$, her action, and the strategy of firm $A$. The set of all such pairs is

$$\{b_1, b_2\} \times \{abstain, \underline{q}, \overline{q}\} \times \big(\{accept(\underline{q}), reject(\underline{q})\} \times \{accept(\overline{q}), reject(\overline{q})\}\big). \tag{1}$$

We now give an informal overview of the logic described by stability under forward induction. In the blocking game, each player holds beliefs about other players' strategies and types, both at the beginning of the game and conditional on an offer. This belief is assumed to be consistent with players' information: each player assigns probability 1 to their type, to the fact that the market is that of Figure 1, to the type of the agent they are matched to, etc. For every $n \in \mathbb{N}$, we call a market-strategy pair, that is, an element of (1), *$n$-rationalizable* if for each player their strategy is a best response to a belief that satisfies the following properties:

(i) at the beginning of the game, it assigns probability one to pairs that are $(n-1)$-rationalizable;

(ii) for players other than worker $b$, their beliefs assign probability one to the event that $b$ will not make offers;

(iii) conditional on an unexpected offer, consider the highest $k \leq n$ such that the offer is part of a market-strategy pair that is $k$-rationalizable. The firm's conditional belief must then assign probability one to $k$-rationalizable pairs.

It is standard to verify that that an $n$-rationalizable pair is also $(n-1)$-rationalizable. Hence, this gives an elimination procedure. We will call a market *stable under forward induction* if abstaining is $n$-rationalizable for every $n$. We now show that both markets $\mathbf{m}_1$ and $\mathbf{m}_2$ are not stable under forward induction.

*Step 1.* In this first step, we eliminate all pairs where the high type $b_2$ makes a low-wage offer, since, even if accepted, it would make her worse off. We also eliminate all pairs where firm $A$ rejects a low-wage offer. All other pairs are 1-rationalizable.

*Step 2.* By the previous step, at the beginning of the game, worker $b$, regardless of her type, must assign probability 1 to firm $A$ accepting a low offer. This rules out pairs where $b_1$ abstains, as she must anticipate that the offer will be accepted. The resulting set of 2-rationalizable pairs in which $b$'s type is 1 is

$$\{b_1\} \times \{\underline{q}, \overline{q}\} \times \big(\{accept(\underline{q})\} \times \{accept(\overline{q}), reject(\overline{q})\}\big) \tag{2}$$

while the set of 2-rationalizable pairs in which $b$'s type is high is

$$\{b_2\} \times \{abstain, \overline{q}\} \times \big(\{accept(\underline{q})\} \times \{accept(\overline{q}), reject(\overline{q})\}\big). \tag{3}$$

No more pairs can be eliminated. For example, it is 2-rationalizable for type $b_1$ to make a high-wage offer, since it is 1-rationalizable for firm $A$ to accept it (under the belief that the offer comes from a high type).

*Step 3.* This step and the next one are where forward-induction reasoning plays a role. The key observation is that firm $B$ knows the true type of worker $b$. Hence, by the previous step, if $b$'s type is low, there can be no belief for $B$ that assigns probability 1 to pairs that are 2-rationalizable and to the event that $b$ abstains from making an offer. Hence, all pairs where $b$'s type is low are eliminated. The set of 3-rationalizable pairs is (3). In particular, it is 3-rationalizable for $b_2$ to abstain since it is 2-rationalizable for $A$ to reject offer $\overline{q}$.

*Step 4.* In this step, we rule out the pairs where firm $A$ rejects a high offer. This is the key step in showing that $\mathbf{m_2}$ is not stable under forward induction. As shown in the previous step, a market-strategy pair where $b_2$ plays $\overline{q}$ is 3-rationalizable, while any pair that involve type $b_1$ is at most 2-rationalizable. Thus, conditional on receiving a high offer, firm $B$ must assign probability 1 to the event that the offer was made by the high type. This follows from requirement (iii).

In more intuitive terms, consider a high-wage offer by worker $b$. Under forward-induction reasoning, firm $A$ interprets the offer by maintaining the highest possible degree of belief in the event that other players are rational and that the offer was unexpected (by everyone other than worker $b$). In particular, and this is the key aspect, firm

$A$ must take into account that the offer was unexpected to firm $B$, even though the same firm *knew* $b$'s actual type. What is the "best" possible explanation that, ex ante, could have justified firm $B$'s belief that worker $b$ was going to abstain? Such an explanation depends on $b$'s type.

Consider the case where $b$'s type is $b_1$. Then firm $B$ must have thought that $b$ believed firm $A$ was irrational. If not, then $B$ would have expected $A$ to accept a low offer, making abstaining a nonoptimal strategy. Now consider the case where $b$'s type is $b_2$. In this case, firm $B$ could have expected $b$ to abstain, as a best response to the belief that firm $A$ would have rejected a high offer under the (incorrect) belief that $b$'s type was 1. The latter explanation assigns a higher degree of rationality to $b$'s belief.

We obtain that the set of 4-rationalizable pairs is

$$\{b_2\} \times \{abstain, \overline{q}\} \times \big(\{accept(\underline{q})\} \times \{accept(\overline{q})\}\big).$$

*Step 5.* The previous step implies that at the beginning of the game, worker $b$ must assign probability 1 to the fact that $A$ will accept a high offer. This makes it impossible for abstaining to be 5-rationalizable. This concludes the argument that $\mathbf{m}_2$ is not stable under forward induction.

Both solution concepts rest on the idea that agents believe a matching to be "as stable as possible," even when taking actions that can lead to a breakdown of the current match. Incomplete information stability operationalizes this idea by stipulating that a firm, whenever it is part of a blocking pair, considers possible only those matching outcomes that have not previously been ruled out at earlier stages of the elimination procedure. In stability under forward induction, upon receiving an offer, a firm restricts attention only to those matching outcomes compatible with the highest possible level of rationalizability.

Finally, it can be useful to compare the logic in the previous example with the one obtained by applying forward-induction reasoning in signaling games. The two are obviously related, as the epistemic conditions behind the notion of rationalizability applied here are analogous to the epistemic conditions used by Battigalli and Siniscalchi (2003) to characterize iterated intuitive criterion in signaling games. In a signaling game, the iterated intuitive criterion is used to test an equilibrium, that is, a distribution over senders' types, messages, and receivers' actions. Forward-induction reasoning requires agents to believe ex ante and, as much as possible, conditional on an unexpected offer, that players are rational and expect others to behave according to the fixed equilibrium distribution.

The main difference with stability under forward induction is that the latter is a property of a single matching outcome $\mathbf{m}$. It refers to a realizations of payoff types, rather than to a distribution over payoff types. In particular, this solution concept does not assign an "equilibrium" action to types, as $b_1$ in the example above, who are not part of a stable allocation.

## 5. Rationalizability

We now formally define stability under forward induction. We first collect some preliminary definitions.

*Conditional beliefs* A *conditional probability system* for player $k$ is a collection of conditional probabilities[4]

$$\beta_k = \big(\beta_k(\cdot|h)\big)_{h\in H} \in \prod_{h\in H} \Delta\big(\mathbf{M}\times S_{-k}(h)\big)$$

with the property that each $\beta_k(\cdot|h)$ is derived from $\beta_k(\cdot|\emptyset)$ by conditioning. That is, for every history $h$ and set $E_{-k}\subseteq \mathbf{M}\times S_{-k}$,

$$\beta_k\big(\mathbf{M}\times S_{-k}(h)|\emptyset\big)>0 \quad \text{implies } \beta_k(E_{-k}|h)=\frac{\beta_k\big(E_{-k}\cap\big(\mathbf{M}\times S_{-k}(h)\big)|\emptyset\big)}{\beta_k\big(\mathbf{M}\times S_{-k}(h)|\emptyset\big)}$$

where $\emptyset$ denotes the empty history.

*Beliefs and information* Players' beliefs are required to conform to the information agents possess about the current matching outcome. Formally, given $\mathbf{m}\in\mathbf{M}$ and a player $k$, a conditional probability system $\beta_k$ *is consistent with k's information* if $\beta_k(\mathcal{P}_k(\mathbf{m})\times S_{-k}(h)|h)=1$ for all $h\in H$. So, under this assumption, players are certain, at every history, of the information described by their partition. We maintain the assumption of consistency throughout the paper.[5]

Given outcome $\mathbf{m}\in\mathbf{M}$, a conditional probability system $\beta_k$ satisfies the *grain of truth assumption* if $\beta_k(\{\mathbf{m}\}\times S_{-k}|\emptyset)>0$. The assumption requires player $k$ to assign strictly positive probability, at the beginning of the game, to the actual matching outcome. It will be sufficient to require *workers*' beliefs to satisfy the grain of truth assumption. Formally, given $\mathbf{m}\in\mathbf{M}$ and a conditional probability system $\beta_k$ we say that $\beta_k$ is *consistent* if it consistent with $k$'s information and, in case $k\in I$, it satisfies the grain of truth assumption.

*Stability* Given a player $k$, a conditional probability system $\beta_k$ *believes in no competing offers* if the initial probability $\beta_k(\cdot|\emptyset)$ assigns probability 1 to each worker $i\neq k$ not making offers. The assumption expresses the idea that if a current matching is deemed to be stable, then players will not expect others to initiate a negotiation to deviate from the match.

*Optimality* Given a player $k$, a strategy $s_k$, and a pair $(\mathbf{m},s_{-k})$ in $\mathbf{M}\times S_{-k}$, let $U_k(s_k,s_{-k},\mathbf{m})$ denote the resulting payoff for player $k$. A strategy $s_k$ is *sequentially optimal under* $\beta_k$ if at every history $h$ where $k$ is asked to act, the action specified by $s_k$ maximizes the expectation of $U_k$ with respect to $\beta_k(\cdot|h)$.[6]

In addition to sequential optimality, we assume that given a conditional probability system $\beta_i$, a worker $i$ makes offers only if they are not indifferent between making offers

[4]For every finite set $S$, we denote by $\Delta(S)$ the set of probability measures on $S$.

[5]Weaker notions of consistency can, however, be easily accommodated into our framework. It would be natural, for example, to consider the case where players do not know the partner's type in the status quo allocation, but instead assign probability greater than $1-\varepsilon$ to the correct type. Two different notions of consistency would correspond different notions of stability under forward induction.

[6]Since players move at most once along each path of play, it is sufficient to define optimality, as we do here, in terms of one-shot deviations.

and abstaining. This tie-breaking assumption rules out cases where a worker makes an offer they expect will be rejected with probability 1. To simplify the language, we call a strategy $s_k$ *optimal under* $\beta_k$ if it sequentially optimal under $\beta_k$, and in case $k$ is a worker, it satisfies the tie-breaking assumption described above.

### 5.1 *Rationalizability and stability*

We now define our main solution concept. For the next definition, given a subset $\Psi \subseteq \mathbf{M} \times S$ we denote by $\Psi_k$ and $\Psi_{-k}$ the projection of $\Psi$ on, respectively, $\mathbf{M} \times S_k$ and $\mathbf{M} \times S_{-k}$.

DEFINITION 1. Let $\mathfrak{R}^0 = \mathbf{M} \times S$. Inductively, for every $n \geq 1$ define $\mathfrak{R}^n$ to be set of pairs $(\mathbf{m}, s) \in \mathbf{M} \times S$ such that for each player $k$ there exists a consistent conditional probability system $\beta_k$ such that the following hold:

**(P1-$n$)** $s_k$ is optimal under $\beta_k$;
**(P2-$n$)** $\beta_k$ believes in no competing offers;
**(P3-$n$)** $\beta_k(\mathfrak{R}_{-k}^{n-1}|\emptyset) = 1$; and
**(P4-$n$)** for all $h \in H$ and $m \in \{0, \ldots, n-1\}$,

$$\text{if } \big(\mathcal{P}_k(\mathbf{m}) \times S_{-k}(h)\big) \cap \mathfrak{R}_{-k}^m \neq \emptyset \text{ then } \beta_k\big(\mathfrak{R}_{-k}^m | h\big) = 1. \tag{4}$$

A pair $(\mathbf{m}, s)$ is *$n$-rationalizable* if it belongs to $\mathfrak{R}^n$. The set of rationalizable outcome-strategy pairs is defined as $\mathfrak{R}^\infty = \bigcap_{n \geq 0} \mathfrak{R}^n$.

DEFINITION 2. An outcome $\mathbf{m} \in \mathbf{M}$ is *stable under forward induction* if $(\mathbf{m}, \textit{abstain}) \in \mathfrak{R}_i^\infty$ for each worker $i$. That is, if it is rationalizable for every worker to abstain from making offers.

Definition 2 is an instance of Battigalli's (2003) notion of strong $\Delta$-rationalizability (see also Battigalli and Siniscalchi (2003)). We now describe the logic underlying the definition.[7]

Consider a pair $(\mathbf{m}, s)$ consisting of a matching outcome and a profile of strategies. The pair is $n$-rationalizable if for each player $k$ we can find conditional beliefs $\beta_k$ so that $\beta_k$ and $s_k$ satisfy four basic conditions. Properties (P1-$n$) and (P2-$n$) establish that players are rational and expect others not to engage in negotiation. As $n$ goes to infinity, (P3-$n$) implies that rationality and belief in no competing offers are almost common belief at the beginning of the game.

Property (P4-$n$) is crucial and disciplines beliefs conditional upon observing unexpected offers. Consider a history $h$ reached after a worker made an unexpected offer to firm $k$. Notice that $\mathfrak{R}^1 \supseteq \cdots \supseteq \mathfrak{R}^{n-1}$ constitute increasingly stringent assumptions on players' beliefs and behavior. By (4), conditional on the offer, firm $k$ assigns probability 1 to the strongest assumption $\mathfrak{R}^m$ that, by satisfying $(\mathcal{P}_k(\mathbf{m}) \times S_{-k}(h)) \cap \mathfrak{R}_{-k}^m \neq \emptyset$, has not been refuted by the observed offer *and* $k$'s information $\mathcal{P}_k(\mathbf{m})$ about the market. Hence,

[7] In Section B in the Appendix, we show that Definition 1 is equivalent to the definition of $\Delta$-rationalizability in Battigalli and Siniscalchi (2003), once the latter is adapted to the present framework.

(P4-$n$) captures the idea that players interpret offers according to the highest possible degree of sophistication that can be attached to their proponents and by maintaining, as much as possible, the assumption that offers were ex ante unexpected.

Property (P4-$n$) expresses forward-induction reasoning. Following Battigalli and Siniscalchi (2002), say that a player "strongly believes" an event if she believes the event at the beginning of the game and at every history where the event is not contradicted by the evidence. Upon observing an offer, when $n = 2$, property (P4-$n$) requires players to strongly believe the event "other players are rational and did not expect the offer." When $n = 3$, each player strongly believes that "other players are rational, did not expect the offer, and strongly believe that others are rational and did not expect the offer," and so on. The results in Battigalli and Prestipino (2013) can be adapted to show that at each $n$, Definition 2 captures the implications of, informally, (i) rationality, (ii) consistency, (iii) belief in noncompeting offers and $n$ orders of strong belief in (i)–(iii).

Finally, a matching outcome $\mathbf{m}$ is deemed to be stable under forward induction if, under $\mathbf{m}$, abstaining is a rationalizable strategy for every worker. It should be remarked that abstaining from making offers is not required to be the *only* rationalizable strategy. This makes stability under forward induction a relatively permissive solution concept.

## 6. Incomplete-information stability

A notion of stability under incomplete information was introduced by Liu et al. (2014). Its definition takes the form of an iterative elimination procedure defined over the set of matching outcomes.

Definition 3. Let $\Lambda^0 = \mathbf{M}$. Inductively, for each $\ell \in \mathbb{N}$ define $\Lambda^\ell$ as the set of all outcomes $(\mathbf{w}, \mathbf{f}, \mu, \mathbf{p}) \in \Lambda^{\ell-1}$ such that there is no $i \in I$, $j \in J$, and $q \in \mathbb{R}$ such that

$$\nu(\mathbf{w}_i, \mathbf{f}_j) + q > \nu(\mathbf{w}_i, \mathbf{f}_{\mu(i)}) + \mathbf{p}_{i,\mu(i)} \tag{5}$$

and

$$\phi\big(\mathbf{w}'_i, \mathbf{f}_j\big) - q > \phi(\mathbf{w}_{\mu^{-1}(j)}, \mathbf{f}_j) - \mathbf{p}_{\mu^{-1}(j),j} \tag{6}$$

for all $\mathbf{w}' \in \mathbf{W}$ such that $\mathbf{m}' = (\mathbf{w}', \mathbf{f}, \mu, \mathbf{p})$ satisfies

$$\mathbf{m}' \in \Lambda^{\ell-1}, \tag{7}$$

$$\mathcal{P}_j\big(\mathbf{m}'\big) = \mathcal{P}_j(\mathbf{m}), \quad \text{and} \tag{8}$$

$$\nu\big(\mathbf{w}'_i, \mathbf{f}_j\big) + q > \nu\big(\mathbf{w}'_i, \mathbf{f}_{\mu(i)}\big) + \mathbf{p}_{i,\mu(i)}. \tag{9}$$

$\Lambda^\ell$ is the set of matching outcomes that are *level $\ell$ incomplete-information stable*. The set of *incomplete-information stable* matching outcomes is $\Lambda^\infty = \bigcap_{\ell=1}^{\infty} \Lambda^\ell$.

In Liu et al. (2014), the set $\Lambda^0$ is set equal to the set of all individually rational outcomes, rather than a finite set $\mathbf{M}$ as in Definition 4. The discretization $\Lambda^0 = \mathbf{M}$ simplifies the statements of our main results and avoids measurability considerations. A matching

is stable in the definition of Liu et al. (2014) if and only if it is stable (as defined above) for some market $\mathbf{M}$.[8]

An outcome $(\mathbf{w}, \mathbf{f}, \mu, \mathbf{p})$ is eliminated in the first iteration if it is possible to find a worker $i$, a firm $j$, and a wage $q$ so that the two agents can form a new partnership that is profitable for the worker and gives the firm a higher payoff than the original allocation $(\mu, \mathbf{p})$ for all types $\mathbf{w}'_i$ that satisfy restrictions (7)–(9). When $\ell = 1$, this amounts to considering type profiles $\mathbf{w}'$ that do not contradict the fact that $j$ knows the type of the worker they are matched to, and such that the the partnership, if agreed upon, would be profitable for the worker. Successive iterations shrink the set of types that satisfy (7). In the $\ell$th step of the procedure, the same reasoning is applied to the set of matching outcome that have survived $\ell - 1$ steps of the elimination process.

As shown by Liu et al. (2014), incomplete-information stability satisfies two significant properties. First, any complete-information stable matching is also incomplete-information stable. Hence, for every pair of types profiles $\mathbf{w}$ and $\mathbf{f}$ there exists an outcome $(\mathbf{w}, \mathbf{f}, \mu, \mathbf{p})$ that is stable. Moreover, under standard supermodularity assumptions every stable matching is efficient. Call an outcome $(\mathbf{w}, \mathbf{f}, \mu, \mathbf{p})$ (ex post) *efficient* if it achieves a maximal total surplus across all matching outcomes, keeping $\mathbf{w}$ and $\mathbf{f}$ fixed. Liu et al. (2014) show that if $W$ and $F$ are subsets of $\mathbb{R}$ and $\nu$ and $\phi$ are strictly increasing and strictly supermodular, then every incomplete-information stable matching outcome is efficient.[9]

## 7. Characterization theorems

The next theorem characterizes the set of matching outcomes that are stable under forward induction.

Theorem 1. *A matching outcome $\mathbf{m} \in \mathbf{M}$ is stable under forward induction if and only if it is incomplete-information stable.*

Theorem 1 provides epistemic foundations for incomplete-information stability, which can be interpreted as the outcome of noncooperative negotiation under the assumption that players revise their beliefs according to forward-induction reasoning. The two solution concepts describe different principles under which agents on the uninformed side adjust their beliefs when interpreting deviations from the status quo allocation. Under incomplete-information stability, firms evaluate a candidate blocking pair according to a worst-case scenario. Under stability under forward induction, offers are interpreted according to the best rationalization principle.

To provide an intuition for the result, consider first a matching outcome that fails to be 1-level incomplete-information stable. For instance, the matching outcome $\mathbf{m}_1$ in the example of Section 4, where a worker of type $b_1$ can form a blocking pair with firm $A$ at a transfer $\underline{q}$. The existence of this blocking pair makes abstaining a nonrationalizable action for the worker. In particular, abstaining cannot be a 2-rationalizable

[8]This follows immediately from Lemma 1 and Proposition 2 in Liu et al. (2014).

[9]See Liu et al. (2014) for a formal and more general statement.

strategy: under a belief that satisfies the conditions of rationalizability, the worker must believe that $A$ is rational, and hence that the firm will accept an offer at wage $\underline{q}$. More generally, in any market, the set $\Lambda^1$ of matching outcomes that are 1-level incomplete-information stable coincides with the set of matching outcomes under which abstaining is a 3-rationalizable strategy for each worker (as implied by Theorem 2 below). This is a natural finding. The underlying high-level intuition is that a deviation is profitable even for the worst worker type if and only if it must be accepted by every rational firm, and hence no belief can sustain abstaining as a rational action.[10]

The relation between the two solution concepts is more subtle when they are iterated to higher orders. We focus here on one basic principle relating the two solution concepts. Roughly speaking, both notions ask agents to believe that a matching is "as stable as possible." The way this is built in the definition of incomplete-information stability is evident: at every step of the iteration, the profitability of a blocking pair is checked only with respect to those matching outcomes that have not been ruled out in earlier steps.

The same principle is captured, in a different way, by stability under forward induction. Upon receiving an offer $s_i$, a firm restricts attention only to those matching outcomes $\mathbf{m}$ that make $s_i$ compatible with the highest level of rationalizability. An important observation is that if $\mathbf{m}$ is such that the pair $(\mathbf{m}, s_i)$ is $n$-rationalizable, then it must be that for the same $\mathbf{m}$ it is an $(n-1)$-rationalizable strategy for all workers to abstain from making offers. This is formally proved in Lemma 4 in the Appendix. Hence, intuitively, firms interpret offers by maintaining, as much as possible, the hypothesis that the matching is stable. This observation follows from the fact that at the beginning of the game players do not expect offers to be made, and from the assumption that for every worker there exists another agent who assigns positive probability to the worker's true type. The latter can be a firm, when the worker is matched, or other workers under the grain of truth assumption.

This observation can already be seen at work in the example of Section 4. When firm $A$ receives the high offer $\overline{q}$ from worker $b$, $A$'s conditional belief assigns probability 1 to those matching outcomes that make the offer $\overline{q}$ a 3-rationalizable action. The only possible outcome is the one where $b$'s type is high, because it is the one compatible with firm $B$'s belief that $b$ will abstain.

We can compare the two procedures not only in the final predictions $\mathfrak{R}^\infty$ and $\Lambda^\infty$, but also at each step of the two iterations. For every $n$, we consider the set

$$\mathfrak{S}^n = \left\{\mathbf{m} \in \mathbf{M} : (\mathbf{m}, a) \in \mathfrak{R}_i^n \text{ for every } i \in I\right\}.$$

So, $\mathfrak{S}^n$ is the collection of matching outcomes with the property that maintaining the status quo is, for every worker, an $n$-rationalizable strategy.[11]

Comparing the two sequences $(\Lambda^n)$ and $(\mathfrak{S}^n)$ allows to relate degrees of forward-induction reasoning and levels of stability. Suppose $\mathfrak{S}^n \subseteq \Lambda^\ell$. Then an outside observer

[10]This is analogous to the standard result that in game, an action is undominated if and only if it is optimal with respect to some belief.

[11]The set $\mathfrak{S}^n$ is in general a strict subset of the projection of $\mathfrak{R}^n$ on $\mathbf{M}$. This is because it is possible for a strategy $s_i$ to satisfy $(\mathbf{m}, s_i) \in \mathfrak{R}_i^n$ even if $(\mathbf{m}, a) \notin \mathfrak{R}_i^n$.

who knows agents play $n$-rationalizable strategies will be able to infer that any matching that is not blocked is incomplete-information stable at level $\ell$. Conversely, suppose $\Lambda^\ell \subseteq \mathfrak{S}^n$. In this case, observing a matching outcome belonging to $\Lambda^\ell$ does not reject the hypothesis that players play $n$-rationalizable strategies.

The next result establishes universal bounds relating the two elimination procedures that apply to any market.

THEOREM 2. *For every $\ell \in \mathbb{N}$, $\mathfrak{S}^{3\ell} \subseteq \Lambda^\ell \subseteq \mathfrak{S}^{1+2\ell}$ for every level $\ell$.*

Theorem 1 is an immediate corollary of the result.

## 8. Extensions and discussion

### 8.1 *Offers and rejection*

Stability under forward induction requires abstaining to be a rationalizable strategy for every worker. Alternatively, it may be natural to deem "stable" a matching where any profitable offer, if made, would be rejected. The next theorem shows an equivalence between these two notions. For the next result, we say that a strategy $s_j$ rejects the unilateral offer $s_i = (j, q)$ if $s_j$ rejects offer $s_i$ when the latter is the only offer made by any worker.

THEOREM 3. *A matching outcome $\mathbf{m} = (\mathbf{w}, \mathbf{f}, \mu, \mathbf{p})$ is stable under forward induction if and only if there is a rationalizable pair $(\mathbf{m}, s) \in \mathfrak{R}^\infty$ such that for every firm $j$, the strategy $s_j$ rejects any unilateral offer $s_i = (j, q)$ such that $\nu(\mathbf{w}_i, \mathbf{f}_j) + q > \nu(\mathbf{w}_i, \mathbf{f}_{\mu(i)}) + \mathbf{p}_{i,\mu(i)}$.*

The result shows that a matching is stable if and only if we can find for every firm $j$ a rationalizable strategy $s_j$ that rejects any offer that, if accepted, would be profitable for the worker proposing it.

### 8.2 *Strict stability*

The fact that a matching outcome is stable does not imply that abstaining is, for every worker, the *only* rationalizable strategy. We call an outcome $\mathbf{m} = (\mathbf{w}, \mathbf{f}, \mu, \mathbf{p})$ *strictly stable* if $\mathfrak{R}_i^\infty = \{(\mathbf{m}, a_i)\}$ for every $i$. We now show that strict stability is an unsuitably strong notion of stability. The next result provides a characterization.

THEOREM 4. *Consider a market $\mathbf{M}$. A matching outcome $\mathbf{m} = (\mathbf{w}, \mathbf{f}, \mu, \mathbf{p})$ in $\mathbf{M}$ is strictly stable if and only if $\mathbf{m} \in \Lambda^\infty$ and there is no worker $i$, firm $j$, and payment $q$ such that*

$$\nu(\mathbf{w}_i, \mathbf{f}_j) + q > \nu(\mathbf{w}_i, \mathbf{f}_{\mu(i)}) + \mathbf{p}_{i,\mu(i)}$$

*and*

$$\phi(\mathbf{w}'_i, \mathbf{f}_j) - q \geq \phi(\mathbf{w}_{\mu^{-1}(j)}, \mathbf{f}_j) - \mathbf{p}_{\mu^{-1}(j),j}$$

*for some $\mathbf{w}' \in \mathbf{W}$ such that $\mathbf{m}' = (\mathbf{w}', \mathbf{f}, \mu, \mathbf{p}) \in \Lambda^\infty$, $\mathcal{P}_j(\mathbf{m}') = \mathcal{P}_j(\mathbf{m}')$, and $\nu(\mathbf{w}'_i, \mathbf{f}_j) + q > \nu(\mathbf{w}'_i, \mathbf{f}_{\mu(i)}) + \mathbf{p}_{i,\mu(i)}$.*

A strict stable matching outcome is incomplete-information stable. In addition, under strict stability, a worker $i$ and a firm $j$ can block a matching outcome $\mathbf{m}$ as long as there is *some* payoff profile $\mathbf{w}'$ that makes the combination $(i, j, q)$ profitable for firm $j$ and such that the resulting outcome $(\mathbf{w}', \mathbf{f}, \mu, \mathbf{p})$ is incomplete-information stable. An immediate implication of the result is that a strict stable matching outcome must be complete-information stable. In addition, it is possible to construct markets $\mathbf{M}$ that contain multiple complete-information stable outcomes but no strict stable outcomes.

## Appendix A

### A.1 *Definition of Q*

We now make formal the assumption, introduced in Section 3.1, that the set $Q$ is a sufficiently fine grid. To this end, notice that since $W$ and $F$ are finite we can find a large enough open interval $(\alpha, \gamma) \subseteq \mathbb{R}$ such that it is without loss of generality to restrict the attention, in the definition of incomplete-information stability, to payments $q$ that belong to $(\alpha, \gamma)$. Given $\varepsilon > 0$, we call a finite set $A \subseteq [\alpha, \gamma]$ an *$\varepsilon$-grid* if every open subinterval of $(\alpha, \gamma)$ of diameter $\varepsilon$ intersects $A$. We assume that $Q$ is an $\varepsilon$-grid where $\varepsilon \leq \varepsilon^*$ and the bound $\varepsilon^* > 0$ is described below.

For every $\mathbf{m} \in \mathbf{M} \backslash \Lambda^\infty$, let $n_{\mathbf{m}} \geq 0$ be such that $\mathbf{m} \in \Lambda^{n_{\mathbf{m}}} \backslash \Lambda^{n_{\mathbf{m}}+1}$. For every $\mathbf{m} \in \mathbf{M} \backslash \Lambda^\infty$, consider the set $P_{\mathbf{m}}$ of pairs $(i, j)$ such that for some payment $q \in (\alpha, \gamma)$ the combination $(i, j, q)$ has the property that it $n_{\mathbf{m}}$-blocks the outcome $\mathbf{m}$. For every $(i, j) \in P_{\mathbf{m}}$, select one such payment $q(i, j, \mathbf{m})$ such that $(i, j, q(i, j, \mathbf{m}))$ blocks $\mathbf{m}$.

Because the definition of incomplete-information stability involves only strict inequalities then for each $q(i, j, \mathbf{m})$, there exists a small enough $\varepsilon(i, j, \mathbf{m}) > 0$ such that any $q' \in \mathbb{R}$ that is at distance at most $\varepsilon(i, j, \mathbf{m})$ from $q$ has the properties that $q'$ belongs to $(\alpha, \beta)$ and $(i, j, q')$ also $n_{\mathbf{m}}$-blocks the outcome $\mathbf{m}$. We define $\varepsilon^*$ be the minimal $\varepsilon(i, j, \mathbf{m})$ across all payments $q(i, j, \mathbf{m})$.

By construction, the bound $\varepsilon^*$ has the following property. For every $\varepsilon \leq \varepsilon^*$ and every $\varepsilon$-grid $A$, if there exists a combination $(i, j, q)$ that $n_{\mathbf{m}}$-blocks an outcome $\mathbf{m} \in \mathbf{M}$ then there exists $q' \in A$ such that the combination $(i, j, q')$ also $n_{\mathbf{m}}$-blocks the same outcome.

### A.2 *Preliminaries*

Given any subset $\Psi \subseteq \mathfrak{R}$ and player $k$, denote by $\Psi_k$ and $\Psi_{-k}$ the projections of $\Psi$ on $\mathbf{M} \times S_k$ and $\mathbf{M} \times S_{-k}$, respectively. For every $k$ and conditional probability system (henceforth, CPS) $\beta_k$, we will denote by $\beta_{k,h}$ the probability measure $\beta_k(\cdot|h)$.

As shown in the next lemma, for a given worker $i$ it is enough to consider an initial probability $\rho_i \in \Delta(\mathbf{M} \times S_{-i})$ rather than a fully specified conditional probability system.

Lemma 1. *Fix $n \geq 0$ , $\mathbf{m} \in \mathbf{M}$, $i \in I$, and a strategy $s_i \in S_i$. Let $\rho_i \in \Delta(\mathbf{M} \times S_{-i})$ be a probability measure such that $\rho_i(\{\mathbf{m}\} \times S_{-i}) > 0$, $\rho_i(\mathcal{P}_i(\mathbf{m}) \times S_{-i}) = 1$ and $s_i$, and $\rho_i$ satisfy properties* (P1-$n$)–(P3-$n$). *Then there exists a CPS $\beta_i$ such that $\beta_{i,\emptyset} = \rho_i$ and $s_i$ and $\beta_i$ satisfy properties* (P1-$n$)–(P4-$n$).

PROOF. The CPS $\beta_i$ is easily defined as follows. Let $\beta_{i,\emptyset} = \rho_i$. Denote by $H^A_{-i}$ be the set of histories following no offers from workers other than $i$. For every $h \in H^A_{-i}$, let $\beta_{i,h} = \beta_{i,\emptyset}$. Now consider all histories $h \notin H^A_{-i}$ such that $h \neq \emptyset$ and (4) holds for $m = n-1$. For every such history, define $\beta_{i,h}$ to satisfy $\beta_{i,h}((\{\mathbf{m}\} \times S_{-j}(h)) \cap \mathfrak{R}^m_{-i}) = 1$. Proceeding inductively, we can decrease $m$ and repeat the argument at every step. Because $\mathfrak{R}^0_{-i} = \mathfrak{R}_{-i}$, then for every history there exists $m \leq n-1$ such that (4) holds. So, we obtain a collection of conditional probabilities $\beta_i = (\beta_{i,h})_{h \in H}$. We need to verify that $\beta_i$ is a well-defined CPS. Because $\beta_{i,\emptyset}$ assigns probability 1 to no offer being made by other workers, only histories in $H^A_{-i}$ have initial strictly positive probability. For every such history $h$, we have $\beta_{i,h} = \beta_{i,\emptyset}$, so Bayesian updating is respected. Hence, $\beta_i$ is a well-defined CPS. By construction, the pair $(s_i, \beta_i)$ satisfies (P1-$n$)–(P4-$n$). □

As recorded below, for a fixed matching outcome $\mathbf{m}$ the set $\{s \in S : (\mathbf{m}, s) \in \mathfrak{R}^n\}$ has a product structure. The result follows immediately from Definition 1 and its proof is omitted.

LEMMA 2. *Fix $s \in S$ and $\mathbf{m} \in \mathbf{M}$. If $(\mathbf{m}, s_k) \in \mathfrak{R}^n_k$ for each $k$, then $(\mathbf{m}, s) \in \mathfrak{R}^n$.*

We conclude this subsection with a lemma on the composition of multiple strategies. Recall that $H_j$ denotes the set of histories at which firm $j$ has received at least one offer.

LEMMA 3. *Fix $n \geq 0$, $\mathbf{m} \in \mathbf{M}$, and $j \in J$. Consider a finite sequence*

$$(\mathbf{m}^1, s^1_j), \ldots, (\mathbf{m}^m, s^m_j) \quad \text{in } \mathfrak{R}^n_j$$

*such that $\mathcal{P}_j(\mathbf{m}^1) = \cdots = \mathcal{P}_j(\mathbf{m}^m)$. If a strategy $s_j$ is such that*

$$s_j(h) \in \{s^1_j(h), \ldots, s^m_j(h)\} \quad \text{for all } h \in H_j,$$

*then $(\mathbf{m}, s_j)$ belongs to $\mathfrak{R}^n_j$.*

PROOF. For every $r = 1, \ldots, m$, let $\beta^r_j$ be a consistent CPS such that $s^r_j$ and $\beta^r_j$ satisfy properties (P1-$n$)–(P4-$n$). For every $h \in H_j$, let $r(h) \in \{1, \ldots, m\}$ be such that $s_j(h) = s^{r(h)}_j(h)$. Define the CPS $\beta_j$ as $\beta_{j,h} = \beta^{r(h)}_{j,h}$ for every $h \in H_j$ and $\beta_{j,h} = \beta^1_{j,h}$ for every $h \in H \backslash H_j$. The CPS $\beta_j$ is well-defined. To see this, notice that the only history different from $\emptyset$ that is reached with positive probability under $\beta_{j,\emptyset} = \beta^1_{j,\emptyset}$ is the history $h^*$ in which all workers abstain from making offers. Because $h^* \notin H_j$, then $\beta_{j,h^*} = \beta^1_{j,h^*}$. Hence, the requirement that each $\beta_{j,h}$ is obtained by conditioning is respected. Since each $\beta^r_j$ is consistent, it follows that $\beta_j$ is consistent as well. In addition, because $\beta_{j,\emptyset} = \beta^1_{j,\emptyset}$ then $\beta_j$ satisfies (P2-$n$) and (P3-$n$). We now verify that (P4-$n$) holds. For every $m \in \{0, \ldots, n-1\}$ and every history $h$, if $(\mathbf{M} \times S_{-j}(h)) \cap \mathcal{P}_j(\mathbf{m}) \cap \mathfrak{R}^m_{-j} \neq \emptyset$ then $\beta^{r(h)}_{j,h}$ assigns probability 1 to $\mathfrak{R}^m_{-j}$, hence $\beta_{j,h}$ assigns probability 1 to $\mathfrak{R}^m_{-j}$ as well. Thus, property (P4-$n$) is satisfied. Finally, the action $s_j(h)$ is optimal with respect to $\beta^{r(h)}_{j,h} = \beta_{j,h}$ at every history $h \in H_j$. Hence, $s_j$ is optimal with respect to $\beta_j$. Therefore, we can conclude that $(\mathbf{m}, s_j) \in \mathfrak{R}^n_j$. □

### A.3 *Proof of Theorems 1 and 2*

Let $S_I = \prod_{i\in I} S_i$ and $S_J = \prod_{j\in J} S_j$. For every $n$, denote by $\mathfrak{R}^n_I$ the projection of $\mathfrak{R}^n$ on $\mathbf{M}\times S_I$ and by $\mathfrak{R}^n_J$ the projection on $\mathfrak{R}^n$ on $\mathbf{M}\times S_J$. Also, let $a_I = (a_i)_{i\in I}$ and for each $i$ denote by $a_{-i}$ the vector $(a_k)_{k\in I\setminus\{i\}}$.

LEMMA 4. *For every $\mathbf{m}\in\mathbf{M}$, $i\in I$, $n\geq 1$, and $s_i\in S_i$:*

1. *If $(\mathbf{m}, s_i)\in\mathfrak{R}^n_i$, then $(\mathbf{m}, a_I)\in\mathfrak{R}^{n-1}_I$;*
2. *If $(\mathbf{m}, a_I)\in\mathfrak{R}^{n-1}_I$, then $(\{\mathbf{m}\}\times S)\cap\mathfrak{R}^n\neq\emptyset$.*

PROOF. (1) Suppose $(\mathbf{m}, s_i)\in\mathfrak{R}^n_i$. Consider a worker $k\neq i$ (recall that $|I|\geq 2$ by assumption). Then, since $(\mathbf{m}, s_k)\in\mathfrak{R}^n_k$, there must exist a corresponding CPS $\beta_k$ such that $\beta_{k,\emptyset}(\mathfrak{R}^{n-1}_{-k}) = 1$, $\beta_{k,\emptyset}(\{\mathbf{m}\}\times S_{-k}) > 0$ and $\beta_{k,\emptyset}(A_i) = 1$, where $A_i = \{(\mathbf{m}, s_{-k}) : s_i = a_i\}$. Therefore,

$$\beta_{k,\emptyset}\big(\mathfrak{R}^{n-1}_{-k}\cap\big(\{\mathbf{m}\}\times S_{-k}\big)\cap A_i\big) > 0$$

So, in particular, $(\mathbf{m}, a_i)\in\mathfrak{R}^{n-1}_i$. Because $s_i$ and $i$ are arbitrary, it follows that $(\mathbf{m}, a_i)\in\mathfrak{R}^{n-1}_i$ for every $i\in I$. Hence, Lemma 2 implies $(\mathbf{m}, a_I)\in\mathfrak{R}^{n-1}_I$.

(2) Let $(\mathbf{m}, a_I)\in\mathfrak{R}^{n-1}_I$. Then $(\{\mathbf{m}\}\times\{a_I\}\times S_J)\cap\mathfrak{R}^{n-1}\neq\emptyset$. Thus, for each player $k$ we can find a probability $\rho_k\in\Delta(\mathfrak{R}_{-k})$ assigning probability 1 to $(\{\mathbf{m}\}\times\{a_{-k}\}\times S_J)\cap\mathfrak{R}^{n-1}_{-k}$.

The probability $\rho_k$ can then be extended to a consistent CPS $\beta_k$ such that $\beta_{k,\emptyset} = \rho_k$. To this end, define a vector $(\beta_{k,h})_{h\in H}$ as follows. Let $\beta_{k,\emptyset} = \rho_k$. As in the proof of Lemma 1, let $H^A_{-k}$ be the set of histories following no offers from workers $I\setminus\{k\}$. For every $h\in H^A_{-k}$, let $\beta_{k,h} = \beta_{k,\emptyset}$. Now consider all histories $h\notin H^A_{-k}$ such that $h\neq\emptyset$ and (4) holds for $m = n-1$. For every such history, define $\beta_{k,h}$ to assign probability 1 to $(\mathcal{P}_k(\mathbf{m})\times S_{-k}(h))\cap\mathfrak{R}^m_{-i}$. Proceeding inductively, decrease $m$ and repeat the argument to obtain a vector $\beta_k = (\beta_{k,h})_{h\in H}$. We need to verify that $\beta_k$ is a well-defined conditional probability system. Because $\beta_{k,\emptyset}$ assigns probability 1 all workers abstaining from making offers (except possibly for $k$), only histories in $H^A_{-k}$ are reached with strictly positive probability under $\beta_{k,\emptyset}$. For every such history $h$, we have $\beta_{k,h} = \beta_{k,\emptyset}$. Hence, $\beta_k$ is a well-defined conditional probability system. By construction, it is consistent. In addition, $\beta_k$ believes in no competing offers, and it is immediate to verify it satisfies properties (P3-$n$) and (P4-$n$). Any strategy $s_k$ that is optimal with respect to $\beta_k$ is such that the pair $(s_k, \beta_k)$ satisfies properties (P1-$n$)–(P4-$n$). A profile $s$ of such strategies satisfies $(\mathbf{m}, s)\in\mathfrak{R}^n$. Hence, $(\{\mathbf{m}\}\times S)\cap\mathfrak{R}^n\neq\emptyset$. □

The next two lemmas provide conditions that are sufficient and necessary for a matching outcome $\mathbf{m}$ to satisfy $(\mathbf{m}, a_I)\in\mathfrak{R}^n_I$. To ease notation, we denote by $[i, j, q]$ the second-stage history reached when all workers except for $i$ abstain from making offers and $i$ makes offer $(j, q)$.

LEMMA 5. *For every $n\geq 1$, $(\mathbf{m}, a_I)\in\mathfrak{R}^n_I$ if and only if $(\mathbf{m}, a_I)\in\mathfrak{R}^{n-1}_I$ and there exists a strategy profile $(s^*_j)_{j\in J}$ such that:*

1. $(\mathbf{m}, s_j^*) \in \mathfrak{R}_j^{n-1}$ *for every* $j$*; and*
2. $s_j^*(h) = r$ *for every* $j$ *and every history* $h = [i, j, q]$ *that satisfies*

$$\nu(\mathbf{w}_i, \mathbf{f}_j) + q > \nu(\mathbf{w}_i, \mathbf{f}_{\mu(i)}) + \mathbf{p}_{i,\mu(i)}.$$

Proof of Lemma 5. Let $(\mathbf{m}, a_I) \in \mathfrak{R}_I^n$. Consider an offer $(j, q)$ by worker $i$ such that $\nu(\mathbf{w}_i, \mathbf{f}_j) + q > \nu(\mathbf{w}_i, \mathbf{f}_{\mu(i)}) + \mathbf{p}_{i,\mu(i)}$ and fix $h = [i, j, q]$. We claim there must exist a strategy of firm $j$, which we denote by $s_j^{i,q}$, with the properties that $(\mathbf{m}, s_j) \in \mathfrak{R}_j^{n-1}$ and $s_j^{i,q}(h) = r$.

Suppose, as a way of contradiction, that such a strategy does not exists. Then offer $(j, q)$ is accepted (i.e., $s_j(h) = i$) by any strategy $s_j \in S_j$ such that $(\mathbf{m}, s_j) \in \mathfrak{R}_j^{n-1}$. Let $\beta_i$ be a CPS that satisfies the grain of truth assumption and such that $\beta_{i,\emptyset}(\mathfrak{R}_{-i}^{n-1}) = 1$. Then $\beta_{i,\emptyset}$ must attach strictly positive probability to the event where $s_j(h) = i$. Therefore, if $\beta_i$ is a consistent CPS that satisfies properties (P2-$n$) and (P3-$n$) then $a_i$ cannot be optimal with respect to $\beta_i$. This contradicts the assumption that $(\mathbf{m}, a_i) \in \mathfrak{R}_i^n$ and concludes the proof of the claim.

Given a firm $j$, define the set

$$D_j = \left\{ s_j^{i,q} : i \in I,\, q \in Q \text{ and } \nu(\mathbf{w}_i, \mathbf{f}_j) + q > \nu(\mathbf{w}_i, \mathbf{f}_{\mu(i)}) + \mathbf{p}_{i,\mu(i)} \right\}$$

We now compose the strategies in $D_j$ into a new strategy $s_j^*$ as follows: For every history $h = [i, j, q]$ such that $\nu(\mathbf{w}_i, \mathbf{f}_j) + q > \nu(\mathbf{w}_i, \mathbf{f}_{\mu(i)}) + \mathbf{p}_{i,\mu(i)}$, let $s_j^*(h) = s_j^{i,q}(h)$. For any other history $h \in H_j$, let $s_j^*(h) = s_j(h)$ for some strategy $s_j$ such that $(\mathbf{m}, s_j) \in \mathfrak{R}_j^{n-1}$. Because the set $D_j$ is finite, Lemma 3 implies $(\mathbf{m}, s_j^*) \in \mathfrak{R}_j^{n-1}$. This concludes the first part of the proof.

We now prove the converse implication. Because $(\mathbf{m}, a_I) \in \mathfrak{R}_I^{n-1}$, we know from Lemma 4 that $(\{\mathbf{m}\} \times S) \cap \mathfrak{R}^n \neq \emptyset$. We now show that $(\mathbf{m}, a_I) \in \mathfrak{R}_I^n$. Let $s_J^* = (s_j^*)_{j \in J}$ be a profile of strategies that satisfies conditions (1) and (2) in the statement. For every worker $i$, let $\rho_i \in \Delta(\mathfrak{R}_{-i})$ assign probability 1 to $(\mathbf{m}, a_{-i}, s_J^*)$. Because $(\mathbf{m}, a_I) \in \mathfrak{R}_I^{n-1}$ and $(\mathbf{m}, s_J^*) \in \mathfrak{R}_J^{n-1}$, then $(\mathbf{m}, a_I, s_J^*) \in \mathfrak{R}^{n-1}$ by Lemma 3. So, $\rho_i$ assigns probability 1 to $\mathfrak{R}_{-i}^{n-1}$. Strategy $a_i$ is optimal with respect to $\rho_i$. Using Lemma 1, we can define a consistent CPS $\beta_i$ such that $\beta_{i,\emptyset} = \rho_i$ and $a_i$ and $\beta_i$ satisfy properties (P1-$n$)–(P4-$n$). Hence, $(\mathbf{m}, a_i) \in \mathfrak{R}_i^n$. By repeating the construction for every $i \in I$, we obtain $(\mathbf{m}, a_I) \in \mathfrak{R}_I^n$. □

Lemma 6. *Let* $\mathbf{m} = (\mathbf{w}, \mathbf{f}, \mu, \mathbf{p})$*. For every* $n \geq 2$*,* $(\mathbf{m}, a_I) \in \mathfrak{R}_I^n$ *if and only if* $(\mathbf{m}, a_I) \in \mathfrak{R}_I^{n-1}$ *and there is no worker* $i$ *and strategy* $s_i = (j, q)$ *such that* $\nu(\mathbf{w}_i, \mathbf{f}_j) + q > \nu(\mathbf{w}_i, \mathbf{f}_{\mu(i)}) + \mathbf{p}_{i,\mu(i)}$ *and*

$$\phi\left(\mathbf{w}_i', \mathbf{f}_j\right) - q > \phi(\mathbf{w}_{\mu^{-1}(j)}, \mathbf{f}_j) - \mathbf{p}_{\mu^{-1}(j),j}$$

*for all and at least one profile* $\mathbf{w}' \in \mathbf{W}$ *such that*

$$\mathbf{w}'_{\mu^{-1}(j)} = \mathbf{w}_{\mu^{-1}(j)} \quad \textit{and} \quad \left((\mathbf{w}', \mathbf{f}, \mu, p), s_i, a_{-i}\right) \in \mathfrak{R}_I^{n-2}. \tag{10}$$

Proof of Lemma 6. We first prove the "only if" part. Suppose $(\mathbf{m}, a_I) \in \mathfrak{R}_I^n$ and $s_i = (j, q)$ and $\mathbf{w}'$ are such that $\nu(\mathbf{w}_i, \mathbf{f}_j) + q > \nu(\mathbf{w}_i, \mathbf{f}_{\mu(i)}) + \mathbf{p}_{i,\mu(i)}$ and

$$\mathbf{w}'_{\mu^{-1}(j)} = \mathbf{w}_{\mu^{-1}(j)} \quad \text{and} \quad \left((\mathbf{w}', \mathbf{f}, \mu, p), s_i, a_{-i}\right) \in \mathfrak{R}_I^{n-2}.$$

We show that $\phi(\mathbf{w}_i'', \mathbf{f}_j) - q \leq \phi(\mathbf{w}_{\mu^{-1}(j)}, \mathbf{f}_j) - \mathbf{p}_{\mu^{-1}(j),j}$ for some $\mathbf{w}''$ that satisfies (10).

Since $(\mathbf{m}, a_I) \in \mathfrak{R}_I^n$, we can apply Lemma 5. Let $(s_j^*)_{j\in J} \in S_J$ be the corresponding profile of strategies. In particular, $(\mathbf{m}, s_j^*) \in \mathfrak{R}_j^{n-1}$ for every $j$. For each $j$, let $\beta_j^*$ be a consistent CPS such that $s_j^*$ and $\beta_j^*$ satisfy properties (P1-$(n-1)$)–(P4-$(n-1)$).

Let $h = [i, j, q]$. Since $(\mathbf{w}', \mathbf{f}, \mu, p) \in \mathcal{P}_j(\mathbf{m})$ we have $(\mathbf{M} \times S_{-j}(h)) \cap \mathcal{P}_j(\mathbf{m}) \cap \mathfrak{R}_{-j}^{n-2} \neq \emptyset$. So, $b_{j,h}^*$ must assign probability 1 to $\mathfrak{R}_{-j}^{n-2}$. Because $s_j^*(h) = r$ then, in order for $r$ to be optimal with respect to $b_{j,h}^*$, the latter must attach strictly positive probability to some profile $\mathbf{w}'' \in \mathbf{W}$ such that

$$\mathbf{w}''_{\mu^{-1}(j)} = \mathbf{w}_{\mu^{-1}(j)} \quad \text{and} \quad \phi\big(\mathbf{w}_i'', \mathbf{f}_j\big) - q \leq \phi(\mathbf{w}_{\mu^{-1}(j)}, \mathbf{f}_j) - \mathbf{p}_{\mu^{-1}(j),j}.$$

This concludes the first part of the proof.

We now prove the "if" part. Let $(\mathbf{m}, a_I) \in \mathfrak{R}_I^{n-1}$ and assume that the other conditions in the "if" part of the statement are satisfied. We now show that $(\mathbf{m}, a_I) \in \mathfrak{R}_I^n$. For every firm $j$, let $H_j^*$ be the set of histories of the form $h = [i, j, q]$ for some offer $s_i = (j, q)$ such that $\nu(\mathbf{w}_i, \mathbf{f}_j) + q > \nu(\mathbf{w}_i, \mathbf{f}_{\mu(i)}) + \mathbf{p}_{i,\mu(i)}$ and $(\mathbf{M} \times S_{-j}(h)) \cap \mathcal{P}_j(\mathbf{m}) \cap \mathfrak{R}_{-j}^{n-2} \neq \emptyset$.

For every $h \in H_j^*$ we can define, by assumption, a probability $\rho_{j,h} \in \Delta(\mathfrak{R}_{-j}^{n-2})$ whose marginal on $\mathbf{M}$ assigns probability 1 to an outcome $(\mathbf{w}^h, \mathbf{f}, \mu, \mathbf{p}) \in \mathcal{P}_j(\mathbf{m})$ where

$$\phi\big(\mathbf{w}_i^h, \mathbf{f}_j\big) - q \leq \phi(\mathbf{w}_{\mu^{-1}(j)}, \mathbf{f}_j) - \mathbf{p}_{\mu^{-1}(j),j}.$$

We now extend the vector $(\rho_h)_{h\in H_j^*}$ to a CPS. First, the vector is extended to the collection of all histories $h$ such that $(\mathbf{M} \times S_{-j}(h)) \cap \mathcal{P}_j(\mathbf{m}) \cap \mathfrak{R}_{-j}^{n-2} \neq \emptyset$. To this end, define the probability $\rho_{j,\emptyset}$ to satisfy $\operatorname{marg}_{\mathbf{M}\times S_I} \rho_{j,\emptyset}(\mathbf{m}, a_I) = 1$ and $\rho_{j,\emptyset}(\mathfrak{R}_{-j}^{n-2}) = 1$. This is possible since $(\mathbf{m}, a_I)$ belongs to $\mathfrak{R}_I^{n-1} \subseteq \mathfrak{R}_I^{n-2}$ by assumption. If $h$ is the history following no offers to any firm, let $\rho_{j,h} = \rho_{j,\emptyset}$. For any other history $h$ such that $h \notin H_j^*$ but $(\mathbf{M} \times S_{-j}(h)) \cap \mathcal{P}_j(\mathbf{m}) \cap \mathfrak{R}_{-j}^{n-2} \neq \emptyset$, let $\rho_{j,h}$ assign probability 1 to $(\mathbf{M} \times S_{-j}(h)) \cap \mathcal{P}_j(\mathbf{m}) \cap \mathfrak{R}_{-j}^{n-2}$.

The resulting vector of conditional probabilities can now be extended to a CPS. Recall that $(\mathbf{m}, a_I) \in \mathfrak{R}_I^{n-1}$. So, we can apply Lemma 5 and obtain a profile $s_J^* = (s_j^*)_{j\in J}$ of strategies that satisfy $(\mathbf{m}, s_j^*) \in \mathfrak{R}_j^{n-2}$ for every $j$ as well as condition (2) of that lemma. For each $j$, let $\beta_j^*$ be a consistent CPS such that $s_j^*$ and $\beta_j^*$ satisfy properties (P1-$(n-2)$)–(P4-$(n-2)$). Define a CPS $\beta_j$ such that

$$\beta_{j,h} = \rho_{j,h} \quad \text{if } h \text{ is such that } \big(\mathbf{M} \times S_{-j}(h)\big) \cap \mathcal{P}_j(\mathbf{m}) \cap \mathfrak{R}_{-j}^{n-2} \neq \emptyset \text{ and}$$
$$\beta_{j,h} = \beta_{j,h}^* \quad \text{otherwise}$$

(Battigalli (1997) applies a similar argument).[12] It is immediate to verify $\beta_j$ is consistent.

Now let $s_j$ be a strategy such that:

(i) $s_j(h) = r$ for every $h \in H_j^*$;

[12]As before, to verify that the CPS $\beta_j$ is well-defined, we need to verify that after every history $h$ that has positive probability under $\beta_{j,\emptyset}$, the conditional probability $\beta(\cdot|h)$ is obtained by conditioning. The only such history is the history $h$ following no offers to any firm, and in that case $\beta_{j,h} = \beta_{j,\emptyset}$.

(ii) $s_j(h)$ is a best response to $\beta_{j,h}$ for every $h \in H_j \backslash H_j^*$ such that $(\mathbf{M} \times S_{-j}(h)) \cap \mathcal{P}_j(\mathbf{m}) \cap \mathfrak{R}_{-j}^{n-2} \neq \emptyset$; and

(iii) $s_j(h) = s_j^*(h)$ for every other history $h \in H_j$.

We now verify that $(\mathbf{m}, s_j) \in \mathfrak{R}_j^{n-1}$. By definition, $s_j(h)$ is a best response to $\beta_{j,h}$ at every $h \in H_j$. So $s_j$ is optimal with respect to $\beta_j$. By the definition of $\rho_{j,\emptyset}$, $\beta_j$ also satisfies (P2-$(n-1)$) and (P3-$(n-1)$).

To verify (P4-$(n-1)$), let $m \in \{0, \ldots, n-2\}$ and $h$ be such that $(\mathbf{M} \times S_{-j}(h)) \cap \mathcal{P}_j(\mathbf{m}) \cap \mathfrak{R}_{-j}^m \neq \emptyset$. If $m = n-2$, then $\beta_{j,h}(\mathfrak{R}_{-j}^{n-2}) = \rho_{j,h}(\mathfrak{R}_j^{n-2}) = 1$. If $m < n-2$, then $\beta_{j,h}(\mathfrak{R}_{-j}^m) = \beta_{j,h}^*(\mathfrak{R}_{-j}^m) = 1$. So, (P4-$(n-1)$) is satisfied. We can therefore conclude that $(\mathbf{m}, s_j) \in \mathfrak{R}_j^{n-1}$.

We can now repeat this construction for every $j$. Consider the resulting profile $(s_j)_{j \in J} \in S_J$. Let $s_i = (j, q)$ be an offer such that $\nu(\mathbf{w}_i, \mathbf{f}_j) + q > \nu(\mathbf{w}_i, \mathbf{f}_{\mu(i)}) + \mathbf{p}_{i,\mu(i)}$, and let $h = [i, j, q]$. If $(\mathbf{m}, s_i, a_{-i}) \in \mathfrak{R}_I^{n-2}$, then $h \in H_j^*$ so $s_j(h) = r$ as required by (i) above. If $(\mathbf{m}, s_i, a_{-i}) \notin \mathfrak{R}_I^{n-2}$, then the intersection $(\mathbf{M} \times S_{-j}(h)) \cap \mathcal{P}_j(\mathbf{m}) \cap \mathfrak{R}_{-j}^{n-2}$ is empty; hence, $s_j(h) = s_j^*(h) = r$, as implied by (iii).

To conclude, the strategy profile $(s_j)_{j \in J}$ satisfies properties (1) and (2) in the statement of Lemma 5. Because $(\mathbf{m}, a_I) \in \mathfrak{R}_I^{n-1}$, then the same lemma implies $(\mathbf{m}, a_I) \in \mathfrak{R}_I^n$. □

LEMMA 7. *Let $(\mathbf{m}, a_I) \in \mathfrak{R}_I^n$. If the offer $s_i = (j, q)$ is such that*

$$\nu(\mathbf{w}_i, \mathbf{f}_j) + q > \nu(\mathbf{w}_i, \mathbf{f}_{\mu(i)}) + \mathbf{p}_{i,\mu(i)} \quad \textit{and} \tag{11}$$

$$\phi(\mathbf{w}_i, \mathbf{f}_j) - q \geq \phi(\mathbf{w}_{\mu^{-1}(j)}, \mathbf{f}_j) - \mathbf{p}_{\mu^{-1}(j),j}, \tag{12}$$

*then $(\mathbf{m}, s_i) \in \mathfrak{R}_i^n$.*

PROOF. Because $(\mathbf{m}, a_I) \in \mathfrak{R}_I^n$, we can apply Lemma 5. Let $(s_j^*)_{j \in J}$ be a profile that satisfies conditions (1) and (2) in the statement of that lemma. Fix a worker $i$ and an offer $s_i = (j, q)$ such that (11) and (12) hold. Let $h = [i, j, q]$. Define $s_j$ as $s_j(h) = i$ and $s_j(h') = s_j^*(h')$ for every $h' \in H_j$ different from $h$. So, the strategy $s_j$ accepts the offer $(j, q)$ and rejects any other offer that if accepted would improve $i$'s payoff strictly above the status quo.

We now claim that $(\mathbf{m}, s_i) \in \mathfrak{R}_i^m$ and $(\mathbf{m}, s_j) \in \mathfrak{R}_j^{m-1}$ for every $m \in \{1, \ldots, n\}$. The proof proceeds by induction on $m$. Given (11), the claim is easily seen to hold for $m = 1$. Suppose it is true for $m \in \{1, \ldots, n-1\}$. We now show that $(\mathbf{m}, s_i) \in \mathfrak{R}_i^{m+1}$ and $(\mathbf{m}, s_j) \in \mathfrak{R}_j^m$. Let $\beta_j^*$ be a consistent CPS such that $s_j^*$ and $\beta_j^*$ satisfy properties (P1-$(n-1)$)–(P4-$(n-1)$). Define a new CPS $\beta_j$ as follows: if $h = [i, j, q]$, then $\beta_{j,h}$ assigns probability 1 to

$$\left(\mathbf{m}, s_i, a_{-i}, \left(s_{\hat{j}}^*\right)_{\hat{j} \in J \backslash \{j\}}\right)$$

and if $h' \neq h$ then $\beta_{j,h'} = \beta_{j,h'}^*$. Notice that $\beta_j$ is a well-defined and consistent CPS.

Inequality (12) implies that $s_j(h)$ is optimal with respect to $\beta_{j,h}$. It follows that $s_j$ and $\beta_j$ satisfy (P1-$m$). Because $(\mathbf{m}, s_j^*) \in \mathfrak{R}_j^{n-1}$ then $\beta_j^*$ satisfies (P2-$(n-1)$) and (P3-$(n-1)$). Hence, $\beta_j^*$ satisfies (P2-$m$) and (P3-$m$). It follows then that $\beta_j$ also satisfies (P2-$m$) and (P3-$m$). To verify (P4-$m$), consider first the history $h = [i, j, q]$. Because $(\mathbf{m}, s_i) \in \mathfrak{R}_i^m \subseteq \mathfrak{R}_i^{m-1}$ by the inductive hypothesis and $(\mathbf{m}, a_I) \in \mathfrak{R}_I^n \subseteq \mathfrak{R}_I^{m-1}$ by assumption, then by Lemma 2 we have $(\mathbf{m}, s_i, a_{-i}) \in \mathfrak{R}_I^{m-1}$. Similarly, because $(\mathbf{m}, s_j) \in \mathfrak{R}_j^{m-1}$ and $(\mathbf{m}, s_{\hat{j}}^*) \in \mathfrak{R}_{\hat{j}}^{n-1}$ for every $\hat{j} \neq j$, we have

$$\left(\mathbf{m}, s_j, \left(s_{\hat{j}}^*\right)_{\hat{j} \in J - \{j\}}\right) \in \mathfrak{R}_J^{m-1}$$

Hence, using the fact that $(\mathbf{m}, a_I) \in \mathfrak{R}_I^{m-1}$, we obtain

$$\left(\mathbf{m}, s_i, a_{-i}, \left(s_{\hat{j}}^*\right)_{\hat{j} \in J \setminus \{j\}}\right) \in \mathfrak{R}_{-j}^{m-1}$$

so $(\mathbf{M} \times S_{-j}(h)) \cap \mathcal{P}_j(\mathbf{m}) \cap \mathfrak{R}_{-j}^{m-1} \neq \emptyset$; hence, $\beta_{j,h}$ assigns probability 1 to $\mathfrak{R}_{-j}^{m-1}$. It follows from the definition of $\beta_j^*$ and the fact that $(\mathbf{m}, s_j^*) \in \mathfrak{R}_j^{n-1}$ that property (P4-$(m)$) is verified with respect to any other history $h' \neq h$. We can conclude that $(\mathbf{m}, s_j) \in \mathfrak{R}_j^m$.

Let $\rho_i \in \Delta(\mathfrak{R}_{-i})$ assign probability 1 to

$$\left(\mathbf{m}, a_{-i}, s_j, \left(s_{\hat{j}}^*\right)_{\hat{j} \in J \setminus \{j\}}\right) \tag{13}$$

By the inductive hypothesis, $(\mathbf{m}, s_i) \in \mathfrak{R}_I^m$. As shown above, $(\mathbf{m}, s_j) \in \mathfrak{R}_j^m$ hence,

$$\left(\mathbf{m}, s_j, \left(s_{\hat{j}}^*\right)_{\hat{j} \in J \setminus \{j\}}\right) \in \mathfrak{R}_J^m$$

By assumption $(\mathbf{m}, a_I) \in \mathfrak{R}_I^m$. It follows that (13) belongs to $\mathfrak{R}_{-i}^m$. Hence, $\rho_i(\mathfrak{R}_{-i}^m) = 1$. Moreover, $s_i = (j, q)$ is optimal with respect to the probability $\rho_i$. By applying Lemma 1, we can define a consistent CPS $\beta_i$ such that $\beta_{i,\emptyset} = \rho_i$ and such that $s_i$ and $\beta_i$ satisfy (P1-$(m+1)$)–(P4-$(m+1)$). Therefore, $(\mathbf{m}, s_i) \in \mathfrak{R}_i^{m+1}$. This concludes the proof of the inductive step. We conclude that $(\mathbf{m}, s_i) \in \mathfrak{R}_I^n$. □

If $(i, j, q)$ is a combination that satisfies (5)–(9) in the definition of $\Lambda^n$, then we say the outcome $(\mathbf{w}, \mathbf{f}, \mu, \mathbf{p})$ is *$n$-blocked* by $(i, j, q)$. The next two lemmas are the main steps in the proof of Theorem 2.

LEMMA 8. *For every $n \geq 0$ and $\mathbf{m} \in \mathbf{M}$, if $(\mathbf{m}, a_I) \in \mathfrak{R}_I^{3n}$ then $\mathbf{m} \in \Lambda^n$. Hence, $\mathfrak{S}^{3n} \subseteq \Lambda^n$.*

PROOF. The proof proceeds by induction. The result is vacuously true when $n = 0$. Now assume the result is true for $n \geq 0$. Let $\mathbf{m} = (\mathbf{w}, \mathbf{f}, \mu, \mathbf{p}) \in \mathbf{M} \setminus \Lambda^{n+1}$. We now show that $(\mathbf{m}, a_I) \notin \mathfrak{R}^{3n+3}$. Assume that $\mathbf{m} \in \Lambda^n \setminus \Lambda^{n+1}$. This assumption is without loss of generality since, if $\mathbf{m} \notin \Lambda^n$ then $(\mathbf{m}, a_I) \notin \mathfrak{R}_I^{3n}$ by the inductive hypothesis. By the definition of $Q$ (see Section A.1), it follows that we can find a tuple $(i, j, q)$ where $q \in Q$ and that $(n+1)$-blocks $\mathbf{m}$.

So, $(i, j, q)$ satisfies $\nu(\mathbf{w}_i, \mathbf{f}_j) + q > \nu(\mathbf{w}_i, \mathbf{f}_{\mu(i)}) + \mathbf{p}_{i,\mu(i)}$ and $\phi(\mathbf{w}'_i, \mathbf{f}_j) - q > \phi(\mathbf{w}_{\mu^{-1}(j)}, \mathbf{f}_j) - \mathbf{p}_{\mu^{-1}(j),j}$ for all $\mathbf{w}' \in \mathbf{W}$ such that

$$\begin{aligned}(\mathbf{w}', \mathbf{f}, \mu, \mathbf{p}) &\in \Lambda^n,\\ \mathbf{w}'_{\mu^{-1}(j)} &= \mathbf{w}_{\mu^{-1}(j)}, \quad \text{and}\\ \nu(\mathbf{w}'_i, \mathbf{f}_j) + q &> \nu(\mathbf{w}'_i, \mathbf{f}_{\mu(i)}) + \mathbf{p}_{i,\mu(i)}.\end{aligned}$$

Because $(\mathbf{w}, \mathbf{f}, \mu, \mathbf{p}) \in \Lambda^n$, it follows that $\phi(\mathbf{w}_i, \mathbf{f}_j) - q > \phi(\mathbf{w}_{\mu^{-1}(j)}, \mathbf{f}_j) - \mathbf{p}_{\mu^{-1}(j),j}$.

We now show that Lemma 6 implies $(\mathbf{m}, a_I) \notin \mathfrak{R}_I^{3n+3}$. Assume, by way of contradiction, that $(\mathbf{m}, a_I) \in \mathfrak{R}_I^{3n+3}$ and consider the offer $s_i = (j, q)$. Because $\nu(\mathbf{w}_i, \mathbf{f}_j) + q > \nu(\mathbf{w}_i, \mathbf{f}_{\mu(i)}) + \mathbf{p}_{i,\mu(i)}$ and $\phi(\mathbf{w}_i, \mathbf{f}_j) - q > \phi(\mathbf{w}_{\mu^{-1}(j)}, \mathbf{f}_j) - \mathbf{p}_{\mu^{-1}(j),j}$, Lemma 7 implies $(\mathbf{m}, s_i, a_{-i}) \in \mathfrak{R}_I^{3n+3}$. Hence, $(\mathbf{m}, s_i, a_{-i}) \in \mathfrak{R}_I^{3n+1}$. Consider now any profile $\mathbf{w}'$ that, as $\mathbf{w}$, satisfies

$$\begin{aligned}\mathbf{w}'_{\mu^{-1}(j)} &= \mathbf{w}_{\mu^{-1}(j)}, \quad \text{and}\\ \big((\mathbf{w}', \mathbf{f}, \mu, \mathbf{p}), s_i, a_{-i}\big) &\in \mathfrak{R}_I^{3n+1}.\end{aligned}$$

We now show that $\mathbf{w}'$ must satisfy $\phi(\mathbf{w}'_i, \mathbf{f}_j) - q > \phi(\mathbf{w}_{\mu^{-1}(j)}, \mathbf{f}_j) - \mathbf{p}_{\mu^{-1}(j),j}$. By Lemma 6, this will imply $(\mathbf{m}, a_I) \notin \mathfrak{R}_I^{3n+3}$.

Let $\mathbf{m}' = (\mathbf{w}', \mathbf{f}, \mu, \mathbf{p})$. Because $(\mathbf{m}', s_i, a_{-i}) \in \mathfrak{R}_I^{3n+1}$, Lemma 4 implies $(\mathbf{m}', a_I) \in \mathfrak{R}_I^{3n}$. By the inductive hypothesis, we conclude that $\mathbf{m}' \in \Lambda^n$. By assumption, $\mathbf{w}'_{\mu^{-1}(j)} = \mathbf{w}_{\mu^{-1}(j)}$. In addition, because $s_i$ is optimal then $\nu(\mathbf{w}'_i, \mathbf{f}_j) + q > \nu(\mathbf{w}'_i, \mathbf{f}_{\mu(i)}) + \mathbf{p}_{i,\mu(i)}$. Therefore, since $(i, j, q)$ $(n+1)$-blocks $\mathbf{m}$, we conclude that $\phi(\mathbf{w}'_i, \mathbf{f}_j) - q > \phi(\mathbf{w}_{\mu^{-1}(j)}, \mathbf{f}_j) - \mathbf{p}_{\mu^{-1}(j),j}$. This concludes the proof of the result. □

LEMMA 9. *For every $n \geq 1$ and every $\mathbf{m} \in \mathbf{M}$, if $\mathbf{m} \in \Lambda^n$ then $(\mathbf{m}, a_I) \in \mathfrak{R}_I^{1+2n}$. Hence, $\Lambda^n \subseteq \mathfrak{S}^{1+2n}$.*

PROOF. Any outcome $\mathbf{m} \in \mathbf{M}$ satisfies $(\mathbf{m}, a_I) \in \mathfrak{R}_I^1$. This follows from the fact that $a_i$ is optimal under the belief that all offers are rejected. Therefore, $\Lambda^0 = \mathbf{M} = \mathfrak{R}^1$.

Proceeding inductively, assume the result is true for $n \geq 0$. Let $\mathbf{m} = (\mathbf{w}, \mathbf{f}, \mu, \mathbf{p})$ be such that $(\mathbf{m}, a_I) \notin \mathfrak{R}_I^{1+2n+2}$. We show that $\mathbf{m} \notin \Lambda^{n+1}$. It is without loss of generality to assume $\mathbf{m} \in \Lambda^n$ and $(\mathbf{m}, a_I) \in \mathfrak{R}_I^{1+2n}$ (if $(\mathbf{m}, a_I) \notin \mathfrak{R}_I^{1+2n}$ then $\mathbf{m} \notin \Lambda^n$ by the inductive hypothesis). So, $(\mathbf{m}, a_I) \in \mathfrak{R}_I^m \setminus \mathfrak{R}_I^{m+1}$, where $m \in \{1+2n, 1+2n+1\}$.

By Lemma 6, there exists an offer $s_i = (j, q)$ such that

$$\nu(\mathbf{w}_i, \mathbf{f}_j) + q > \nu(\mathbf{w}_i, \mathbf{f}_{\mu(i)}) + \mathbf{p}_{i,\mu(i)}$$

and

$$\phi(\mathbf{w}'_i, \mathbf{f}_j) - q > \phi(\mathbf{w}_{\mu^{-1}(j)}, \mathbf{f}_j) - \mathbf{p}_{\mu^{-1}(j),j}$$

for every and at least one profile $\mathbf{w}'$ such that

$$\mathbf{w}'_{\mu^{-1}(j)} = \mathbf{w}_{\mu^{-1}(j)} \quad \text{and} \quad (\mathbf{m}', s_i, a_{-i}) \in \mathfrak{R}_I^{m-1} \tag{14}$$

where $\mathbf{m}' = (\mathbf{w}', \mathbf{f}, \mu, \mathbf{p})$.

We now show that $(i, j, q)$ $(n+1)$-blocks $\mathbf{m}$. To reach this conclusion, we need to show that every profile $\mathbf{w}''$ that, as $\mathbf{w}$, satisfies

$$\begin{aligned} &\mathbf{m}'' = \big(\mathbf{w}'', \mathbf{f}, \mu, \mathbf{p}\big) \in \Lambda^n, \qquad \mathbf{w}''_{\mu^{-1}(j)} = \mathbf{w}_{\mu^{-1}(j)} \quad \text{and} \\ &\nu\big(\mathbf{w}''_i, \mathbf{f}_j\big) + q > \nu\big(\mathbf{w}''_i, \mathbf{f}_{\mu(i)}\big) + \mathbf{p}_{i,\mu(i)} \end{aligned} \tag{15}$$

has the property that $\phi(\mathbf{w}''_i, \mathbf{f}_j) - q > \phi(\mathbf{w}_{\mu^{-1}(j)}, \mathbf{f}_j) - \mathbf{p}_{\mu^{-1}(j),j}$.

The next step in the proof is to show that any $\mathbf{m}''$ that satisfies (15) must also satisfy $(\mathbf{m}'', s_i, a_{-i}) \in \mathfrak{R}_I^{m-1}$. By (14), this will imply $\phi(\mathbf{w}''_i, \mathbf{f}_j) - q > \phi(\mathbf{w}_{\mu^{-1}(j)}, \mathbf{f}_j) - \mathbf{p}_{\mu^{-1}(j),j}$, establishing that $(i, j, q)$ $(n+1)$-blocks $\mathbf{m}$.

To this end, fix an outcome $\mathbf{m}''$ that satisfies (15). By the inductive hypothesis, we know that $(\mathbf{m}'', a_I) \in \mathfrak{R}_I^{1+2n}$. Because $m \leq 1 + 2n + 1$, then $\mathfrak{R}_I^{1+2n+1} \subseteq \mathfrak{R}_I^m$ so $\mathfrak{R}_I^{1+2n} \subseteq \mathfrak{R}_I^{m-1}$. Thus, $(\mathbf{m}'', a_I) \in \mathfrak{R}_I^{m-1}$. We now show that $(\mathbf{m}'', s_i, a_{-i}) \in \mathfrak{R}_I^{m-1}$. This conclusion is reached in three steps.

First, fix a matching outcome $\mathbf{m}'$ that satisfies (14). Because $(\mathbf{m}', s_i) \in \mathfrak{R}_i^{m-1}$, there must exist a pair $(\mathbf{m}, \tilde{s}_J) \in \mathfrak{R}_J^{m-2}$ such that $\mathbf{m} \in \mathcal{P}_i(\mathbf{m}')$ and $\tilde{s}_j$ accepts the offer $s_i = (j, q)$, that is, $\tilde{s}_j([i, j, q]) = i$. If not, then $s_i$ could not be optimal with respect to a consistent CPS that assigns probability 1 to $\mathfrak{R}_{-i}^{m-2}$.

Second, because $(\mathbf{m}'', a_I) \in \mathfrak{R}_I^{m-1}$ we can apply Lemma 5 and obtain a profile $(\mathbf{m}'', s_J^*)$ in $\mathfrak{R}_J^{m-2}$ with the property that every offer $(\hat{\jmath}, \hat{q})$ by player $i$ such that $\nu(\mathbf{w}''_i, \mathbf{f}_{\hat{\jmath}}) + \hat{q} > \nu(\mathbf{w}''_i, \mathbf{f}_{\mu(i)}) + \mathbf{p}_{i,\mu(i)}$ is rejected by $s_{\hat{\jmath}}^*$. That is, $s_{\hat{\jmath}}^*([i, \hat{\jmath}, \hat{q}]) = r$.

Third, consider a probability $\rho_i \in \Delta(\mathfrak{R}_{-i})$ that has support $(\mathbf{m}'', a_{-i}, s_J^*)$ and $(\mathbf{m}, a_{-i}, \tilde{s}_j)$. By construction, it satisfies $\rho_i(\mathfrak{R}_{-i}^{m-2}) = 1$ and the strategy $s_i = (j, q)$ is the unique best reply to $\rho_i$. By Lemma 1, $\rho_i$ can be extended to a consistent CPS $\beta_i$ such that $\beta_{i,\emptyset} = \rho_i$ and the pair $(s_i, \beta_i)$ satisfies (P1-$(m-1)$)–(P4-$(m-1)$) with respect to the outcome $\mathbf{m}''$. Hence, $(\mathbf{m}'', s_i, a_{-i}) \in \mathfrak{R}_I^{m-1}$.

Therefore, $\mathbf{m}''$ satisfies (14). Thus, $\phi(\mathbf{w}''_i, \mathbf{f}_j) - q > \phi(\mathbf{w}_{\mu^{-1}(j)}, \mathbf{f}_j) - \mathbf{p}_{\mu^{-1}(j),j}$. Because this is true for every $\mathbf{m}''$ that satisfies (15), we conclude that $(i, j, q)$ $(n+1)$-blocks the outcome $(\mathbf{w}, \mathbf{f}, \mu, \mathbf{p})$. So, $(\mathbf{w}, \mathbf{f}, \mu, \mathbf{p}) \notin \Lambda^{n+1}$. □

Proof of Theorem 2. Lemmas 8 and 9 show that $\Lambda^n \subseteq \mathfrak{S}^{1+2n}$ and $\mathfrak{S}^{3n} \subseteq \Lambda^n$. □

Proof of Theorem 1. Consider a market $\mathbf{M}$. Since $\mathbf{M} \times S$ is finite there exists $N$ large enough such that or $\mathfrak{R}^\infty = \mathfrak{R}^N$. Similarly, there exists $n$ such that $\Lambda^\infty = \Lambda^n$. Therefore, $\Lambda^n = \Lambda^\ell$ for every $\ell \geq n$. Let $\mathbf{m} \in \Lambda^\infty$. By Theorem 2, taking $\ell \geq N$ we obtain $\Lambda^\ell \subseteq \Lambda^N \subseteq \mathfrak{S}^{1+2N} \subseteq \mathfrak{S}^N = \mathfrak{S}^\infty$. Hence, $\mathbf{m} \in \mathfrak{S}^\infty$. Conversely, let $\mathbf{m} \in \mathfrak{S}^\infty$. If $\ell \geq N$, then $\mathbf{m} \in \mathfrak{S}^\ell = \mathfrak{S}^{3\ell}$ and Theorem 2 implies $\mathbf{m} \in \mathfrak{S}^{3\ell} \subseteq \Lambda^\ell$. Since $\ell$ is arbitrary, then $\mathbf{m} \in \Lambda^\infty$. □

### A.4 *Proofs of other results*

Proof of Theorem 3. Let $\mathbf{m}$ be stable under forward induction. So, $(\mathbf{m}, a_I) \in \mathfrak{R}_I^\infty$. Let $n$ be such that $\mathfrak{R}^{n-1} = \mathfrak{R}^\infty$. Since $(\mathbf{m}, a_I) \in \mathfrak{R}_I^n$, by Lemma 6 there exists a strategy profile $(s_j^*)$ such that $(\mathbf{m}, s_j^*) \in \mathfrak{R}_j^{n-1} = \mathfrak{R}_j^\infty$ for every $j$ and $s_j^*(h) = r$ for every $j$ and every history $h = [i, j, q]$ that satisfies $\nu(\mathbf{w}_i, \mathbf{f}_j) + q > \nu(\mathbf{w}_i, \mathbf{f}_{\mu(i)}) + \mathbf{p}_{i,\mu(i)}$. Lemma 2 implies $(\mathbf{m}, a_I, (s_j^*)_{j \in J}) \in \mathfrak{R}^{n-1} = \mathfrak{R}^\infty$. □

Proof of Theorem 4. Suppose $\mathbf{m}$ is not strictly stable. Let $(\mathbf{w}, s_i) \in \mathfrak{R}_i^\infty$ where $s_i = (j, q)$. We can choose $n \geq 0$ large enough so that $\mathfrak{R}^\infty = \mathfrak{R}^n = \mathfrak{R}^{n-2}$. There must exist a strategy $s_j$ such that $(\mathbf{w}, s_j) \in \mathfrak{R}_j^\infty$ and $s_j$ accepts the offer $(j, q)$. Let $\beta_j$ be a CSP such that $s_j$ and $\beta_j$ satisfy (P1-$n$)–(P4-$n$). By (P4-$n$), it must be that $\beta_{j,h}(\mathfrak{R}_{-j}^\infty) = 1$. Hence, there is a profile $\mathbf{w}' \in \mathbf{W}$ in the support of $\beta_{j,h}$ such that $\phi(\mathbf{w}_i', \mathbf{f}_j) - q \geq \phi(\mathbf{w}_{\mu^{-1}(j)}, \mathbf{f}_j) - \mathbf{p}_{\mu^{-1}(j),j}$, $\mathbf{w}'_{\mu^{-1}(j)} = \mathbf{w}_{\mu^{-1}(j)}$, and $(\mathbf{w}', s_i) \in \mathfrak{R}_i^\infty$. By Lemma 2, $(\mathbf{w}', a_I) \in \mathfrak{R}_I^\infty$. Hence, $(\mathbf{w}', \mathbf{f}, \mu, \mathbf{p}) \in \Lambda^\infty$. This concludes the proof.

We now show the "if" part of the proof. Suppose we can find a tuple $(i, j, q)$ and a profile $\mathbf{w}' \in \mathbf{W}$ such that $\nu(\mathbf{w}_i, \mathbf{f}_j) + q > \nu(\mathbf{w}_i, \mathbf{f}_{\mu(i)}) + q$, $\mathbf{w}'_{\mu^{-1}(j)} = \mathbf{w}_{\mu^{-1}(j)}$, and

$$\phi(\mathbf{w}_i', \mathbf{f}_j) - q \geq \phi(\mathbf{w}_{\mu^{-1}(j)}, \mathbf{f}_j) - \mathbf{p}_{\mu^{-1}(j),j}, \tag{16}$$

$$\nu(\mathbf{w}_i', \mathbf{f}_j) + q > \nu(\mathbf{w}_i', \mathbf{f}_{\mu(i)}) + \mathbf{p}_{i,\mu(i)}, \quad \text{and} \tag{17}$$

$$(\mathbf{w}', \mathbf{f}, \mu, \mathbf{p}) \in \Lambda^\infty.$$

Because $(\mathbf{w}', \mathbf{f}, \mu, \mathbf{p}) \in \Lambda^\infty$, then $(\mathbf{w}', a_I) \in \mathfrak{R}_I^\infty$. Let $s_i = (j, q)$. Then (16), (17), and Lemma 7 imply $(\mathbf{w}', s_i) \in \mathfrak{R}_i^\infty$. We now show that $(\mathbf{w}, s_i) \in \mathfrak{R}_i^\infty$, concluding that (1) must be violated. The proof is similar to the proof of Lemma 7. Because $(\mathbf{w}', s_i) \in \mathfrak{R}_i^\infty$, there must exist a strategy $s_j$ such that $s_j$ accepts the offer $(j, q)$ and $(\mathbf{w}', s_j) \in \mathfrak{R}_j^\infty$. Because $(\mathbf{w}, a_I) \in \mathfrak{R}_I^\infty$, by Lemma 5 we can find a strategy profile $(s_j^*)_{j \in J}$ such that $(\mathbf{w}, s_j^*) \in \mathfrak{R}_j^\infty$ for every $j$ and such that any offer $(\hat{\jmath}, \hat{q})$ by worker $i$ that, if accepted, would improve worker $i$'s payoff above the default allocation, is rejected by strategy $s_j^*$. Now define a new strategy $s_j'$ as follows. At the history $h$ corresponding to the offer $(j, q)$ from worker $i$, let $s_j'(h) = s_j(h) = i$. At every other history $h$, $s_j'(h) = s_j^*(h)$. By Lemma 3, $(\mathbf{w}, s_j') \in \mathfrak{R}_j^\infty$. Let $\beta_i'$ be a conditional probability system such that $\beta'_{i,\emptyset}$ is concentrated on

$$\big(\mathbf{w}, a_{-i}, s_j', (s_{\hat{\jmath}}^*)_{\hat{\jmath} \in J - \{j\}}\big).$$

Under $\beta_i'$, the offer $s_i = (j, q)$ is a strict best response. It is immediate to check that $s_i$ and $\beta_i'$ satisfy (P1-$n$)–(P4-$n$), where $\mathfrak{R}^n = \mathfrak{R}^\infty$. Hence, $(\mathbf{w}, s_i) \in \mathfrak{R}_i^\infty$. Thus, $m$ is not strictly stable. □

## Appendix B: Alternative characterization

The purpose of this section is to establish an equivalence between Definition 1 and the definition of $\Delta$-rationalizability put forward in Battigalli and Siniscalchi (2003), adapted to the present framework. Our approach follows Battigalli (1997) and Battigalli and Prestipino (2013).

Definition 4. Let $\Sigma^0 = \mathbf{M} \times S$. Inductively, for every $n \geq 1$ define $\Sigma^n$ to be set of pairs $(\mathbf{m}, s) \in \Sigma^{n-1}$ such that for each player $k$ there exists a consistent conditional probability system $\beta_k$ so that the following properties hold:

**(P1′-$n$)** $s_k$ is optimal under $\beta_k$;
**(P2′-$n$)** $\beta_k$ believes in no competing offers;
**(P3′-$n$)** $\beta_k(\Sigma_{-k}^{n-1} | \emptyset) = 1$; and

**(P4′-$n$)** for all $h \in H$,

$$\text{if } \big(\mathcal{P}_k(\mathbf{m}) \times S_{-k}(h)\big) \cap \Sigma_{-k}^{n-1} \neq \emptyset \text{ then } \beta_k\big(\Sigma_{-k}^{n-1}|h\big) = 1. \tag{18}$$

This elimination procedure differs from Definition 1: At each step $n$, only pairs $(\mathbf{m}, s)$ that have survived $n-1$ steps are considered. In addition, property P4′-$n$ requires to keep track only of the last step of the elimination procedure, unlike P4-$n$, which requires to keep track of all previous steps. The next result shows that Definitions 1 and 4 are equivalent.

Proposition 1. *For every market* $\mathbf{M}$ *and* $n \geq 1$, $\mathfrak{R}^n = \Sigma^n$.

Proof. A simple argument by induction shows that $\mathfrak{R}^n \subseteq \mathfrak{R}^{n-1}$ for all $n$. By definition, $\mathfrak{R}^0 = \Sigma^0$. Assume $\mathfrak{R}^m = \Sigma^m$ for all $m \leq n-1$. We now show that $\mathfrak{R}^n \subseteq \Sigma^n$.

If $\mathfrak{R}^n = \emptyset$, the result is obvious. Assume $(\mathbf{m}, s_k) \in \mathfrak{R}_k^n$ and let $\beta_k$ a corresponding consistent conditional probability system such that conditions P1-$n$ to P4-$n$ hold. Since $(\mathbf{m}, s_k) \in \mathfrak{R}_k^n \subseteq \mathfrak{R}_k^{n-1}$, then $(\mathbf{m}, s_k) \in \Sigma_k^{n-1}$. It is moreover immediate to verify conditions P1′-n and P2′-n hold. Condition P3-$n$ and the inductive hypothesis imply

$$\beta_k\big(\mathfrak{R}_{-k}^{n-1}|\emptyset\big) = \beta_k\big(\Sigma_{-k}^{n-1}|\emptyset\big) = 1.$$

Finally, if $h$ satisfies (18), then P4-$n$ implies

$$\beta_k\big(\mathfrak{R}_{-k}^{n-1}|h\big) = \beta_k\big(\Sigma_{-k}^{n-1}|h\big) = 1,$$

hence P4′-$n$ holds. It follows that $(\mathbf{m}, s_k) \in \Sigma_k^n$. Hence, $\mathfrak{R}^n \subseteq \Sigma^n$.

We now show that $\Sigma^n \subseteq \mathfrak{R}^n$. Let $(\mathbf{m}, s_k) \in \Sigma_k^n$ and let $\beta_k$ be an associated consistent conditional probability system such that conditions P1′-$n$ to P4′-$n$ hold. Since $(\mathbf{m}, s_k) \in \Sigma_k^{n-1} = \mathfrak{R}_k^{n-1}$, then there exists a conditional probability system $\tilde{\beta}_k$ that satisfies P1-$(n-1)$ to P4-$(n-1)$. Define a new conditional probability system $\beta_k'$ as $\beta_k'(\cdot|h) = \beta_k(\cdot|h)$ if $h = \emptyset$ or if $h$ satisfies

$$\big(\mathcal{P}_k(\mathbf{m}) \times S_{-k}(h)\big) \cap \Sigma_{-k}^{n-1} = \big(\mathcal{P}_k(\mathbf{m}) \times S_{-k}(h)\big) \cap \mathfrak{R}_{-k}^{n-1} \neq \emptyset$$

(where the equality follows from the inductive hypothesis) and $\beta_k'(\cdot|h) = \tilde{\beta}_k(\cdot|h)$ for every other history $h$.

It is immediate to check that $\beta_k'$ is a well-defined conditional probability system: because $\beta_k$ and $\tilde{\beta}_k$ believe in no competing offers, then the property that $\beta_k'(\cdot|h)$ is derived from $\beta_k'(\cdot|\emptyset)$ by conditioning is trivially satisfied. Moreover, it is consistent. It is also easy to see that $s_k$ is optimal under $\beta_k'$ and $\beta_k'$ believes in no competing offers. Because $\beta_k$ satisfies P3′-n and $\mathfrak{R}_{-k}^{n-1} = \Sigma_{-k}^{n-1}$, then $\beta_k'$ satisfies P3-$n$. It remains to verify P4-$n$ holds. Let $h \in H$ and $m \in \{0, \dots, n-1\}$ be such that

$$\big(\mathcal{P}_k(\mathbf{m}) \times S_{-k}(h)\big) \cap \mathfrak{R}_{-k}^{m} \neq \emptyset. \tag{19}$$

If $m = n-1$, then $\beta'(\mathfrak{R}_k^{n-1}|h) = \beta(\mathfrak{R}_k^{n-1}|h) = \beta(\Sigma_k^{n-1}|h) = 1$. If instead

$$\big(\mathcal{P}_k(\mathbf{n}-\mathbf{1}) \times S_{-k}(h)\big) \cap \mathfrak{R}_{-k}^{n-1} = \emptyset$$

then $m < n-1$; hence, $\beta'(\mathfrak{R}_k^m|h) = \tilde{\beta}(\mathfrak{R}_k^m|h) = 1$. It follows that $\beta'_k$ satisfies P4-$n$. Hence, $(\mathbf{m}, s_k) \in \mathfrak{R}_k^n$, concluding the proof. □

## References


Banks, Jeffrey, S. and Joel Sobel (1987), "Equilibrium selection in signaling games." Econometrica, 55. [1622, 1623]

Battigalli, Pierpaolo (1996), "Strategic rationality orderings and the best rationalization principle." Games and Economic Behavior, 13. [1622]

Battigalli, Pierpaolo (1997), "On rationalizability in extensive games." Journal of Economic Theory, 74. [1640, 1645]

Battigalli, Pierpaolo (2003), "Rationalizability in infinite, dynamic games with incomplete information." Research In Economics, 57. [1620, 1631]

Battigalli, Pierpaolo and Amanda Friedenberg (2013), "Forward induction reasoning revisited." Theoretical Economics, 7. [1622]

Battigalli, Pierpaolo and Andrea Prestipino (2013), "Transparent restrictions on beliefs and forward-induction reasoning in games with asymmetric information." Advances in Theoretical Economics, 13. [1622, 1632, 1645]

Battigalli, Pierpaolo and Marciano Siniscalchi (2002), "Strong belief and forward induction reasoning." Journal of Economic Theory, 106. [1622, 1632]

Battigalli, Pierpaolo and Marciano Siniscalchi (2003), "Rationalization and incomplete information." Advances in Theoretical Economics, 3. [1620, 1622, 1629, 1631, 1645]

Bikhchandani, Sushil (2014), "Two-sided matching with incomplete information." mimeo.

Bikhchandani, Sushil (2017), "Stability with one-sided incomplete information." *Journal of Economic Theory*, 168, 372–399. [1622]

Chade, Hector (2006), "Matching with noise and the acceptance curse." Journal of Economic Theory, 129. [1621]

Chade, Hector, Gregory Lewis, and Lones Smith (2014), "Student portfolios and the college admission problem." The Review of Economic Studies, 81. [1621]

Chakraborty, Archishman, Alessandro Citanna, and Michael Ostrovsky (2010), "Two-sided matching with interdependent values." Journal of Economic Theory, 145. [1621]

Chen, Yi-Chun and Gaoji Hu (2019), "Learning by matching." Theoretical Economics. (forthcoming). [1622]

Cho, In-Koo and David M. Kreps (1987), "Signaling games and stable equilibria." The Quarterly Journal of Economics, 52. [1622, 1623]

Crawford, Vincent P. and Elsie M. Knoer (1981), "Job matching with heterogeneous firms and workers." Econometrica, 49. [1623]

de Clippel, Geoffroy (2007), "The type-agent core for exchange economies with asymmetric information." Journal of Economic Theory, 135. [1621]

Dutta, Bhaskar and Rajiv Vohra (2005), "Incomplete information, credibility and the core." Mathematical Social Sciences, 50. [1621]

Ehlers, Lars and Jordi Massó (2007), "Incomplete information and singleton cores in matching markets." Journal of Economic Theory, 1. [1621]

Govindan, Srihari and Robert Wilson (2009), "On forward induction." Econometrica, 77. [1623]

Hoppe, Heidrun C., Benny Moldovanu, and Aner Sela (2009), "The theory of assortative matching based on costly signals." Review of Economic Studies, 76. [1621]

Kohlberg, Elon (1981), "Some problems with the concept of perfect equilibrium." Rapp. Rep. NBER Conf. Theory Gen. Econ. Equilibr. K. Dunz. N. Singh, Univ. Calif. Berkeley. [1623]

Kohlberg, Elon and Jean-Francois Mertens (1986), "On the strategic stability of equilibria." Econometrica, 54. [1623]

Liu, Qingmin (2020), "Stability and Bayesian consistency in two-sided markets." American Economic Review, 110. [1622]

Liu, Qingmin, George J. Mailath, Andrew Postlewaite, and Larry Samuelson (2014), "Stable matching with incomplete information." Econometrica, 82. [1621, 1622, 1623, 1632, 1633]

Mailath, George J., Andrew Postlewaite, and Larry Samuelson (2013), "Pricing and investments in matching markets." *Theoretical Economics*, 8, 535–590. [1623]

Man, Priscilla T. Y. (2012), "Forward induction equilibrium." Games and Economic Behavior, 75. [1623]

Myerson, Roger B. (1997), Game theory: analysis of conflict. Harvard university press. [1625]

Myerson, Roger (2007), "Virtual utility and the core for games with incomplete information." Journal of Economic Theory, 136. [1621]

Pearce, David G. (1984), "Rationalizable strategic behavior and the problem of perfection." Econometrica, 52. [1620, 1622]

Peivandi, Ahmad (2013), "Participation and unbiased pricing in CDS settlement mechanisms." mimeo. [1621]

Reny, Philip J. (1992), "Backward induction, normal form perfection and explicable equilibria." Econometrica, 60. [1623]

Roth, Alvin E. (1989), "Two-sided matching with incomplete information about others' preferences." Games and Economic Behavior, 1. [1621]

Serrano, Roberto and Rajiv Vohra (2007), "Information transmission in coalitional voting games." Journal of Economic Theory, 134. [1621]

Shapley, Lloyd S. and Martin Shubik (1971), "The assignment game I: The core." International Journal of Game Theory, 1. [1623, 1624]

Sobel, Joel, Lars Stole, and Inigo Zapater (1990), "Fixed-equilibrium rationalizability in signaling games." Journal of Economic Theory, 52. [1622]

Van Damme, Eric (1989), "Stable equilibria and forward induction." Journal of Economic Theory, 48. [1623]

Vohra, Rajiv (1999), "Incomplete information, incentive compatibility, and the core." Journal of Economic Theory, 86. [1621]

Wilson, Robert (1978), "Information, efficiency, and the core of an economy." Econometrica, 46. [1621, 1624]

---

Co-editor Simon Board handled this manuscript.

Manuscript received 18 July, 2019; final version accepted 26 September, 2021; available online 26 October, 2021.