Stability and Bayesian Consistency in Two-Sided Markets*

Qingmin Liu†

April 15, 2020

Abstract

We propose a criterion of stability for two-sided markets with asymmetric information. A central idea is to formulate matching functions, off-path beliefs conditional on counterfactual pairwise deviations, and on-path beliefs in the absence of such deviations. A matching-belief configuration is stable if the matching is individually rational with respect to the system of on-path beliefs and is not blocked with respect to the system of off-path beliefs. The formulation provides a language for assessing matching outcomes with respect to their supporting beliefs and opens the door to further belief-based refinements. The main refinement analyzed in the paper requires the Bayesian consistency of on-path and off-path beliefs with prior beliefs. We also define concepts of Bayesian efficiency, the rational expectations competitive equilibrium, and the core. Their contrast with pairwise stability manifests the role of information asymmetry in matching formation.

*I thank Navin Kartik, Andy Newman, Andy Postlewaite, George Mailath, David Easley, Larry Blume, Pierre-Andre Chiappori, Michael Grubb, Heng Liu, Vince Crawford, Juan Ortner, Chiara Margaria, Paul Koh, Bumin Yenmez, Debraj Ray, Yeon-Koo Che, Ariel Rubinstein, Dilip Abreu, Hanming Fang, Qianfeng Tang, Mike Borns, Yu Fu Wong, anonymous referees, and audiences at various seminars and conferences. I am especially grateful to Bob Wilson and Jeff Ely whose insights and ideas have significantly improved the paper. The paper previously circulated under the title "Rational Expectations, Stable Beliefs and Stable Matching." This research is supported by NSF grant SES-1824328.

†Columbia University, qingmin.liu@columbia.edu.

Contents

1 Introduction

This paper develops a criterion of stability for two-sided markets with asymmetric information. Specifically, we study a market where agents on one side of the market are privately informed of their payoff-relevant attributes. Such a theory of stability is required on two grounds. On the one hand, the solution concept of stability studied by Gale and Shapley (1962) and Shapley and Shubik (1971) has been successful in analyzing matching applications, but the assumption of complete information is often restrictive. On the other hand, the theory of asymmetric information in two-sided markets is long-standing and revolutionary (e.g., Wilson 1967, Akerlof 1970, Spence 1973, Rothschild and Stiglitz 1976, etc.), but the main analytical tools are the competitive equilibrium and non-cooperative game theory, which differ from stability in fundamental respects.[1]

In a complete-information matching problem, two players from opposite sides of the market "block" a matching if both are better off by rematching with each other; a matching is called stable if it is individually rational and no such pairwise blocking opportunity exists. With asymmetric information, the players' incentive to block depends on their private information; hence, the presence of a blocking opportunity reveals information that should lead to a reassessment of the viability of the blocking opportunity. Likewise, the lack of any viable blocking opportunity, a situation that stability describes, should also reveal information. Therefore, the uncertainty in a stable matching is endogenous. In a Bayesian theory of stability, a player's uncertainty at each contingency should be described by a probabilistic belief. However, the circular nature of the inference problem makes this task difficult.

We aim to develop a Bayesian theory of stability and to avoid, if at all possible, compromising cooperative models of matching with ad hoc non-cooperative assumptions. Nevertheless, a comparison with the theory of dynamic non-cooperative games of asymmetric information elucidates aspects of beliefs that must be captured in a satisfactory Bayesian theory of stability. This comparison is warranted because it is a commonly held view that coalitional solution concepts are reduced-form ways of capturing equilibrium or steady-state

[1]Two features distinguish equilibrium theories from stability. First, equilibrium theories are often developed on the premise of individual optimization while holding fixed the behavior of all other actors. In many two-sided markets with pairwise relationships, pairwise blocking or optimization that jointly involves two players from opposite sides of the market is no less plausible than unilateral deviation or optimization. Secondly, non-cooperative games can be used to model coalition formation, but they often require complete specifications of the strategic interactions including actions available to each player, orders of moves, rules of information revelation, etc. In reality, however, researchers may not know the exact nature of the interactions among players. Some assumptions on non-cooperative game forms may seem reasonable in one context but may become unrealistic in another. The advantage of the cooperative concept of stability is that it focuses on payoff assumptions and abstracts away from details of strategic interactions. Indeed, many classic frameworks of adverse selection impose an endgame that precludes further transactions, a problem that is resolved in the framework of this paper.

outcomes of dynamic interactions.[2] The starting point of this comparison is, naturally, sequential equilibrium, a leading solution concept for dynamic games. As Kreps and Wilson (1982, p. 886) explain, a central principle of their theory is to describe equilibrium as a belief-strategy pair, where equilibrium beliefs—both on and off the equilibrium path—are determined in concert with equilibrium strategies. This powerful insight of separating equilibrium beliefs from equilibrium strategies, so convenient as to be taken for granted today, is instrumental in wide-ranging applications of dynamic games and it paved the way for subsequent development of belief-based equilibrium refinements. We may expect an analog in a cooperative theory of stability with asymmetric information; otherwise there would be little hope for the new theory to capture the outcomes of decentralized dynamic market interactions. Specifically, a notion of "on-path stable beliefs" that are consistent with "stable matching outcomes" should be prescribed for a stability concept, and a notion of "off-path stable beliefs" at "off-path" blocking opportunities that deter blocking should be formulated, which would open the door for belief-based refinements. This idea seems obvious and natural, but, surprisingly, it has not been formally examined in matching problems in particular and cooperative games in general.

By explicitly formulating the on-path belief at each matching outcome and the off-path belief at each blocking opportunity, it is possible to define stability through the consistency of the matching-belief configuration: the putative matching is individually rational with respect to the system of on-path beliefs, and it is immune to pairwise blocking with respect to the system of off-path beliefs. Matchings and beliefs that pass the consistency test are referred to as stable matchings and stable beliefs. Although we have borrowed the terminology of "on path" and "off path" from non-cooperative games, we do not impose a specific non-cooperative interpretation on "off-path" events. In defining stability, we simply test a putative matching against all counterfactual blocking opportunities, as is consistent with the complete-information theory of stability.

If we agree that endogenous on-path and off-path stable beliefs are qualitatively different from a prior belief that is an exogenous primitive of our model, the immediate next conceptual question is how to relate these three kinds of beliefs to each other. Since cooperative matching games do not specify strategies and game forms, we cannot apply Bayes' rule as in non-cooperative games. In particular, it would be a futile attempt to explicitly derive an on-path belief by updating the prior belief from a sequence of failed pairwise deviations. This direct approach does not cut through the Gordian knot of circular inference. Here is how we resolve

[2]See, e.g., Gul (1989) and Perry and Reny (1994) for the exposition of this idea known as the "Nash program." In fact, we may argue that many other solution concepts are shortcuts to capture some dynamic interactions, including the fundamental non-cooperative concept of Nash equilibrium.

this question. We make use of an idea of "outcome functions" similar to that in the literature on rational expectations equilibrium: players understand the stable relationship between the underlying uncertainties (players' types) and the observables (matching outcomes). Using Bayes' rule, an on-path stable belief is "updated" from the prior conditional on an observable outcome that is an output of the outcome function; an off-path stable belief associated with an off-path blocking opportunity is further "updated" from the on-path stable belief conditional on the off-path event that the blocking opportunity is mutually acceptable. This *Bayesian consistency* property *refines* the notion of stability, and closes the loop of the circular inference for defining on-path and off-path beliefs in a stable matching: there is no individual or pairwise deviation from a stable matching outcome given the supporting stable beliefs; stable beliefs and a prior belief are Bayesian consistent given the stability of the matching.

Although the concepts of stability and Bayesian consistency apply to both transferable utility and non-transferable utility problems, we study their implications in matching problems with transfers. With complete information, transfers and payoff distributions in a stable matching exhibit a large degree of flexibility, which, unsurprisingly, continues to hold under asymmetric information. However, it is well known that a stable matching with transfers must maximize the total surpluses (Shapley and Shubik 1971). This efficiency is a simple yet remarkable structural property of stability, and extending it to an asymmetric information environment is obviously worthwhile. The existence of stable beliefs allows us to evaluate match efficiency in the Bayesian sense. We define a criterion of Bayesian efficiency as maximization of the expected social surpluses with respect to on-path stable beliefs, the kind of beliefs that an uninformed planner would have. We give conditions under which all stable matchings that are supported by Bayesian consistent beliefs must be Bayesian efficient, and these conditions apply to familiar models of adverse selection.

Incomplete information is qualitatively different from complete information in ways that go beyond efficiency. Motivated by two concepts that have occupied significant places in economic theory, we define the competitive equilibrium and the core to further study the role of informational friction in matching formation. Our notion of rational expectations competitive equilibrium extends the notion of the complete-information competitive equilibrium (see, e.g., Koopmans and Beckmann 1957, Shapley and Shubik 1971, and Becker 1973). A competitive equilibrium specifies a price for any two players (including unmatched pairs) and postulates that the validity of a player's unilateral deviation is independent of the other player's willingness to match with the deviating player. By contrast, stability dispenses with both postulates of price-taking behavior and unilateral deviation. However, the two conceptually different notions are outcome-equivalent for matching under complete informa-

tion. The equivalence breaks down under informational asymmetry. The set of competitive equilibrium matchings and the set of stable matchings with Bayesian consistent beliefs must overlap but in general neither one contains the other. The reason is precisely that the two theories of market mechanisms, stability and competitive equilibrium, make different assumptions about how information is processed in deviations. But a competitive equilibrium is always Bayesian efficient conditional on the information being revealed, because of the nature of unilateral optimization.

Under complete information, pairwise stability is the same as the concept of the core; i.e., a matching is not blocked by any pair of players if and only if it is not blocked by any set of players. Under incomplete information, the set of stable matchings contains the core as a subset. The core can strictly refine stability even when (prior and posterior) type distributions across players are independent: a blocking cycle can be created using multiple pairs, where the formation of each blocking pair is conditional on the formation of others blocking pairs, and every player in this cycle makes inference from the incentives his respective partners both in the putative match and in the coalitional deviation.

The rest of the paper is organized as follows. Section 1.1 demonstrates several features of beliefs and stability using two examples. Section 1.2 discusses the related literature. Section 2 introduce the model and Section 3 defines the notion of stability. Section 4 studies the criterion of Bayesian efficiency. Section 5 compares stability and competitive equilibrium. Section 6 offers several extensions: the core, the notion of correlated stability, and the incentive compatibility. Section 7 concludes.

1.1 Examples

The first example shows that beliefs associated with stability, the on-path beliefs, are qualitatively different from prior beliefs, and they cannot be determined a priori independently of the stability of a matching. The second example demonstrates natural restrictions on beliefs across players.

Example 1. There is one worker whose type is drawn from $\{t_1, t_1'\}$ according to a commonly known prior distribution that assigns probability $q \in (0,1)$ to t_1 and $1-q$ to t_1'. There are two firms (firm 1 and firm 2). The matrix of matching values is given below:

	firm 1	firm 2
t_1	$\underline{-1}, 2$	$-3, \underline{5}$
t_1'	$0, 2$	$\underline{-4}, \underline{5}$

where, for instance, the vector $(-1, 2)$ in the matrix means that, before a transfer is made,

the matching of the type-t_1 worker and firm 1 gives the worker a payoff of -1 (e.g., cost of effort) and the firm a payoff of 2 (e.g., output). Assume that the payoff of an unmatched player is 0.

Suppose a matching is formed. We claim that it is impossible that firm 2 is matched with the worker of type t_1' for a reasonable notion of stability that captures an immunity to pairwise blocking. To see this, suppose to the contrary that the worker's type happens to be t_1' and this worker ends up matching with firm 2. Note that the salary that the worker of type t_1' receives from firm 2 cannot exceed 5 (because firm 2's matching value is 5) and hence the worker's payoff is at most $-4+p \leq 1$, where p is the salary. Since there is only one worker, firm 1 is unmatched. The worker of type t_1' can block the above matching outcome with the unmatched firm 1 with a salary of $p' = 1.5$: the worker obtains a payoff of $0 + p' = 1.5 > 1$ and firm 1 obtains a payoff of $2 - p' = 0.5$ regardless of the worker's type.

Similarly, it is impossible that firm 1 is matched with the worker of type t_1. Otherwise, this worker and the unmatched firm 2 could block the matching outcome with a salary of $p' = 4.5$: the worker would obtain a payoff of $-3+p' = 1.5$, which is higher than the maximal payoff he could obtain in a match with firm 1, and firm 2 would make a payoff of $5-p' = 0.5$ regardless of the worker's type.

Thus we conclude that, in a stable matching, the type-t_1 worker cannot be hired by firm 1 and the type-t_1' worker cannot be hired by firm 2: a firm's posterior belief associated with any stable matching (i.e., the on-path stable belief) must assign probability 1 to t_1 when the worker is matched with firm 2, and must assign probability 1 to t_1' when the worker is matched with firm 1. Therefore, there is a full separation of worker's types irrespective of the prior distribution.

The takeaway of the example is that beliefs in a stable matching should not simply be taken as prior beliefs and they cannot be fixed a priori. The on-path belief must always be determined together with the stability of the matching. ■

Example 2. Consider the following example with three firms and one worker whose type is either t_1 or t_1'.

	firm 1	firm 2	firm 3
t_1	$0, 2$	$0, 0$	$0, 5$
t_1'	$0, 2$	$0, 5$	$0, 0$

Can it be stable for the worker to match with firm 1 with some transfer? Suppose that firms 2 and 3 start with a common prior over $\{t_1, t_1'\}$ and make identical observations (including the fact that firm 1 and the worker are matched); then the two firms should share the same posterior belief, say $q \in [0, 1]$, on t_1. It is clear from the matrix of matching values that no matter what q is, one of the two firms can form a blocking pair with the worker. Therefore,

it is not a stable matching for firm 1 to hire the worker.

The argument above is intuitive, and one feels that a natural notion of stability should not predict otherwise. But we shall raise an issue that may have broader implications, and it should not be surprising for students of non-cooperative game theory. If the putative matching of firm 1 and the worker is under consideration, a new pair formed by firm 2 (or firm 3) and the worker is "off path". The above argument assumes that the off-path belief of the firm is the same as the on-path belief. This assumption is appealing in the context of this example, as there does not seem to be a compelling reason for the two firms to change their belief about t_1 or t_1' given that both types obtain a constant matching value of 0 and prefer to work for whichever firm for a higher wage. But we must be clear that this assumption is a refinement of off-path beliefs. If firm 2 and firm 3 are allowed to have heterogeneous posterior beliefs, and say, firm 2 thinks the worker's type is t_1 and firm 3 thinks the worker's type is t_1', i.e., each firm looks at its respective worst-case scenario, then it would be stable for the worker to match with firm 1.

The takeaway of this example is that the specification of off-path beliefs should not be completely arbitrary. Additional restrictions on beliefs based on our intuition about the game will strengthen the predictive power of the solution concept, as is already well known from the equilibrium refinement literature. The principal refinement we propose in this paper, when applied to this example, will yield the intuitive prediction we started with. ■

1.2 Related Literature

Given the wide range of applications of stable matching, attempts to relax the restriction on complete information are nothing new. Roth (1989) and Chakraborty, Citanna, and Ostrovsky (2010) study implementation of matching outcomes using existing non-cooperative concepts. Their response to incomplete information is natural, but the choice of game forms matters for the equilibrium outcomes and beliefs. By contrast, we propose the notion of stability as a test of a putative matching against all potential pairwise blocking opportunities, as in complete-information theory of stability. Unlike this early work, no game form is imposed and hence the opportunity of blocking is never restricted. This aspect is the central distinction between cooperative and non-cooperative game theory, as well as the main conceptual question for a cooperative theory of incomplete information, a topic pioneered by Wilson (1978). In the sequel, we explain how the present paper diverge from a broad literature in terms of problem formulation and methodology.

1.2.1 The Incomplete-Information Core

In his pathbreaking work, Wilson (1978) defines "coarse core" and "fine core" corresponding to two protocols of information aggregation within a blocking coalition. See Forges, Minelli, and Vohra (2002) for a survey of subsequent developments. Two other concepts stand out. Holmström and Myerson (1983) propose a notion of durability based on a voting game. Dutta and Vohra (2005) define the credible core, where the information that a deviating coalition conditions on is the information that makes the deviation profitable, such that the set of states that engage in a deviation is endogenously determined as a fixed point; Yenmez (2013) defines stability with a similar kind of fixed point.

This literature analyzes a situation in which the final outcomes are not observed, or at least the contracts already in place must be carried out. By contrast, we study a situation in which the outcomes are observed and players consider deviating from an outcome based on their updated information and inference.[3] Our framework is suitable for decentralized applications, such as marriage and labor markets, where players observe an actual market outcome and stability describes a situation in which there is no further coalitional deviation from this outcome. This critical distinction gives rise to the circular inference problem we previously summarized, which is absent from Wilson (1978) and the literature that follows. The new problem calls for a new approach to stability and beliefs. The following self-contained example illustrates the difference in approaches.

Example 3. Consider one seller whose cost is 0 and one buyer whose valuation is either 1 or $\varepsilon \in (0, 1/2)$, with equal prior probability. The following allocation rule is in the credible core of Dutta and Vohra (2005), a refinement of the coarse core: the price is 1 and only the high-type buyer trades with the seller. The reasoning is as follows. A potential coalition that involves the low-type buyer must have a price $p \leq \varepsilon$. This low price will attract the high-type buyer as well, so the seller's belief about the worker is the same as prior belief, as required by Dutta and Vohra (2005), and the expected payoff from the deviation is $p \leq \varepsilon$. But the seller has no incentive to deviate because he gets an expected payoff of $1/2$ from selling only to the high type at price 1. So the allocation rule is not blocked. Accordingly, if the buyer is the low type, the allocation rule prescribes no trade, and nothing can be done when this is a realized outcome. This is not the situation we are considering in this paper. *Observing* a no-trade outcome, the seller would know the buyer's type is low and they would block the no-trade outcome at a price of $\frac{1}{2}\varepsilon$. ■

[3] Forges (1994) proposes a notion of posterior efficiency that conditions on the information revealed by an outcome of a mechanism. Green and Laffont (1987) study "posterior implementability" that utilizes information revealed by observable outcomes.

The previous example illustrates the main difference between our approach and the other literature. However, it is important to point out that our approach is similar to the seminal paper of Wilson (1978), and some of the work that follows it, in two critical respects: we define a solution concept for a class of coalitional incomplete-information games instead of implementing the solution concept, and we avoid making non-cooperative assumptions an indispensable component in the definition of a cooperative solution concept. We shall discuss the issue of incentive compatibility in Section 6.3.

1.2.2 The Belief-Free Approach

As in the present paper, Liu, Mailath, Postlewaite, and Samuelson (2014) depart from the previous literature by assuming observability of matching outcomes, which makes inference from the non-existence of pairwise blocking necessary. Without pinning down the beliefs that this inference induces, they take an approach in spirit resembles the rationalizability concept of Bernheim (1984) and Pearce (1984), notwithstanding the issues created by incomplete information and the observability of matching outcomes. They observe that matching outcomes that can be blocked under any beliefs of the uninformed firm that are consistent with a worker's incentive to deviate should never be considered as stable. Once these matching outcomes are removed from consideration, the support of admissible beliefs shrinks, which enables further rounds of elimination. Liu et al. (2014) define a concept based on this iterated elimination procedure. Since there is no randomization, blocking under "any beliefs" is mathematically equivalent to blocking under the "worst-case scenario." The key of this definition is to make *no* exogenous restrictions on beliefs over the tentatively surviving outcomes in the elimination process, similar to the process of iterated elimination of never-best responses in non-cooperative games. It would be logically inconsistent otherwise. Alston (2020) and Bikhchandani (2017) demonstrate the consequence of such exogenous restrictions on beliefs in the elimination process.

Notice that from the uniformed firms' perspective, the set of workers' type profiles is not a Cartesian product, and hence it is implicitly a partitional structure. This information is refined along the elimination process. Liu et al. (2014) show that this iterative elimination has an equivalent fixed-point characterization.[4] In subsequent papers that adopt their model,

[4]It should be pointed out that this fixed point is different from the fixed points in Dutta and Vohra (2005) and Yenmez (2013). The latter concerns the self-fulfilling set of types of a blocking coalition and this set is not needed if incomplete information is one-sided as is in Liu et al. (2014) and this paper. The approach of Liu et al. (2014) is also different from the coalitional rationalizability of normal-form games studied by Ambrus (2006), where there is neither inference from incomplete information and observable outcomes nor the idea of blocking; they are different even in the special case of complete information, where the notion of Liu et al. (2014) reduces to the familiar notion of the core.

Chen and Hu (2020) explicitly formulate the partitional information in the fixed-point definition and construct an adaptive learning process learning to stability, and Pomatto (2015) provides an epistemic formulation of blocking by using a forward-induction logic.

This approach, although imposing strong restrictions in specific contexts[5], evades the central question of economic analysis of uncertainty: the Bayesian formulation of prior and posterior beliefs. We should not be content with a belief-free solution for Bayesian matching games that are parameterized by prior beliefs. It should be noted that one cannot obtain an on-path belief in the present paper by imposing a prior belief on a stable set obtained from Liu et al. (2014) and subsequent reformulations; it is logically inconsistent: just as in equilibrium theory, the on-path beliefs and off-path beliefs are determined concurrently, and the belief-free notion is based on different assumptions on the off-path beliefs.

2 The Model

The model is based on job matching between firms and workers studied by Crawford and Knoer (1981). But "firms" and "workers" are just semantics, and the model applies more generally (e.g., men and women, sellers and buyers, etc.). In addition, the model reduces to a non-transferable utility model if transfers are restricted to be zero, where the concepts we shall develop remain valid *mutatis mutandis*. We prove results under the assumption of quasi-linearity in transfers.

2.1 Asymmetric Information

Let $I = \{1, ..., n\}$ be a set of workers, and $J = \{n+1, ..., n+m\}$ be a set of firms. Let T_i be a finite set of types for worker i. Worker i's type $t_i \in T_i$ is his private information. Denote by $t = (t_1, ..., t_n) \in T = \times_{i=1}^{n} T_i$ a profile of private types for the n workers. There is a common prior $\beta^0 \in \Delta(T)$ on workers' type profiles, and β^0 has a full support. Firm j's type is commonly known and is denoted by its index j. Similarly, each worker i can also have publicly observable, payoff-relevant attributes that are denoted by i.

2.2 Match and Payoff

Let $a_{ij}(t_i) \in \mathbb{R}$ and $b_{ij}(t_i) \in \mathbb{R}$ be the **matching values** worker i (with type t_i) and firm j receive, respectively, when they match.[6] To ease notation, for a profile of workers' types

[5] In addition to payoff assumptions, Liu et al. (2014) consider a situation where private information within a matched pair is revealed.

[6] The matching value is allowed to depend on players' observable attributes denoted by i and j. It thus includes as a special case $a_{ij}(t_i) = u(t_i, w_i, f_j)$ and $b_{ij}(t_i) = v(t_i, w_i, f_j)$, where w_i and f_j are worker i's and

$t = (t_i, t_{-i}) \in T$, we write $a_{ij}(t) := a_{ij}(t_i)$ and $b_{ij}(t) := b_{ij}(t_i)$ whenever there is no confusion. We normalize the matching values of unmatched players i and j to 0 and, with a slight abuse of notation, write them as $a_{ii}(t) = b_{jj}(t) = 0$. A **matching game** with asymmetric information is fully summarized by the matching value function $(a, b) : I \times J \times T \to \mathbb{R}^2$ and the common prior $\beta^0 \in \Delta(T)$.

A **match** is a one-to-one function $\mu : I \cup J \to I \cup J$ that pairs up workers and firms such that the following holds for all $i \in I$ and $j \in J$: (i) $\mu(i) \in J \cup \{i\}$, (ii) $\mu(j) \in I \cup \{j\}$, and (iii) $\mu(i) = j$ if and only if $\mu(j) = i$. Here $\mu(i) = i \in I$ means that worker i is unmatched; similarly for $\mu(j) = j \in J$.

Let $p_{ij} \in \mathbb{R}$ be the transfer that worker i receives from firm j. A **transfer scheme** associated with a match μ is a vector $\mathbf{p}$ that specifies a transfer $p_{i\mu(i)} \in \mathbb{R}$ for each $i \in I$ and a transfer $p_{\mu(j)j} \in \mathbb{R}$ for each $j \in J$, where $p_{ii} = p_{jj} = 0$. If worker i and firm j are matched together with a transfer p_{ij} when the profile of workers' types is t, worker i's and firm j's ex post payoffs are $a_{ij}(t) + p_{ij}$ and $b_{ij}(t) - p_{ij}$, respectively.

We shall refer to a match together with a transfer scheme $(\mu, \mathbf{p})$ as a **matching outcome**. We shall assume that a matching outcome is publicly observable.[7]

2.3 Matching Function

For every $t = (t_1, ..., t_n) \in T$, some matching outcome $(\mu, \mathbf{p})$ materializes. In a stable matching, players should correctly understand the relationship between the underlying uncertainties and the observable outcomes, which is described by a function $M : t \mapsto (\mu, \mathbf{p})$. We shall call the function M a **matching function** or simply a **matching** for the matching game with asymmetric information. Three remarks are immediately needed.

Remark 1. The function $M : t \mapsto (\mu, \mathbf{p})$ describes a stable relationship between underlying uncertainties and observables, and players agree on this relationship. This is similar to the classic rational expectations equilibrium approach to markets with incomplete information pioneered by Radner (1979), where an equilibrium relationship is described by a mapping from unobservable uncertainties to publicly observable price vectors. In our matching environment, it is natural to assume that the assignment μ is observable in addition to price vectors $\mathbf{p}$. Economic theorists have utilized a similar approach in other contexts, such as the formulation of conjectural equilibria and self-confirming equilibria (e.g., Rubinstein and Wolinsky 1994, Dekel, Fudenberg, and Levine 2004). ■

firm j's observable characteristics, respectively.

[7] The observability of matches and transfers are empirically relevant; see Salanié (2015) for a discussion of marriage models with transfers.

Remark 2. The matching function describes a deterministic relationship between private types and matching outcomes. In Section 6.2, we incorporate stochasticity through $M : (t, s) \mapsto (\mu, \mathbf{p})$, where $s = (s_{n+1}, ..., s_{n+m})$ is a profile of private signals observed by firms. The assumption of observable matching outcomes simplifies the analysis. If we are interested in the partial observability of matching outcomes, we can introduce a private signal profile $\omega = (\omega_1, ..., \omega_{n+m})$ regarding a matching outcome $(\mu, \mathbf{p})$, and consider a mapping $M : t \mapsto \omega$. It does not take sophisticated thinking to formalize this extension once we see the definition for the case of observable matching outcomes. The extension involves no additional conceptual innovation but necessitates more notation. ■

Remark 3. We may impose the following restriction on the matching function: M is measurable with respect to the privately informed players' matching values, i.e., if for some worker $i \in I$ and his two types $t_i, t_i' \in T_i$, $a_{ij}(t_i) = a_{ij}(t_i')$ for all $j \in J$, then $M(t_i, t_{-i}) = M(t_i', t_{-i})$ for all $t_{-i} \in T_{-i}$. If this condition is satisfied, we say the matching M is **measurable**. Measurability reflects the idea that an uninformed player's private information can be revealed only when it affects the player's own payoff. This restriction is not without loss of generality and one can think of situations where it is not appealing. ■

3 The Criterion of Stability

In this section, we first introduce the plain-vanilla version of stability that incorporates individual rationality under on-path beliefs and the absence of pairwise blocking under off-path beliefs. We then introduce a refinement based on Bayesian consistency between exogenous prior beliefs and endogenous beliefs.

3.1 On-Path Beliefs and Individual Rationality

Consider any matching outcome $(\mu, \mathbf{p})$ that may appear according to the matching function M, i.e., $(\mu, \mathbf{p}) \in M(T)$. Upon observing $(\mu, \mathbf{p})$, each firm $j \in J$ forms a belief $\beta^1_{(\mu,\mathbf{p},j)} \in \Delta(T)$ over the types of all workers. This is firm j's **on-path belief** associated with the matching outcome $(\mu, \mathbf{p})$. Firm j's expected payoff from the matching outcome $(\mu, \mathbf{p})$ is

$$\mathbb{E}_{\beta^1_{(\mu,\mathbf{p},j)}}[b_{\mu(j)j}] - p_{\mu(j)j} = \sum\nolimits_{t \in T} b_{\mu(j)j}(t)\beta^1_{(\mu,\mathbf{p},j)}(t) - p_{\mu(j)j}.$$

As we explained in the introduction, the on-path belief is the endogenous belief formed when there is no longer an opportunity to deviate from a matching. Does the introduction of $\beta^1_{(\mu,\mathbf{p},j)}$ capture this endogeneity in a stable matching? Not yet, but it will. So far, $\beta^1_{(\mu,\mathbf{p},j)}$

describes firm j's posterior beliefs when the putative matching is in place. The to-be-defined stability of M will discipline the on-path belief, as demonstrated already by the first example in Section 1.1. The stability of M and the on-path belief β^1 will be determined jointly rather than separately.

Definition 1. A matching M is **individually rational** with respect to the system of on-path beliefs $\beta^1 = (\beta^1_{(\mu,\mathbf{p},j)})_{(\mu,\mathbf{p})\in M(T), j\in J}$ if

(i) $a_{i\mu(i)}(t) + p_{i\mu(i)} \geq 0$ for all worker $i \in I$, type profile $t \in T$, and matching outcome $(\mu, \mathbf{p}) = M(t)$, and

(ii) $\mathbb{E}_{\beta^1_{(\mu,\mathbf{p},j)}}[b_{\mu(j)j}] - p_{\mu(j)j} \geq 0$ for all firm $j \in J$ and matching outcome $(\mu, \mathbf{p}) \in M(T)$.

3.2 Off-Path Beliefs and Pairwise Blocking

Stability requires that there be no pairwise blocking; i.e., pairwise blocking is a counterfactual, off-path possibility if M is stable. A (pairwise) **deviating coalition** from a matching outcome $(\mu, \mathbf{p}) \in M(T)$ consists of a worker $i \in I$, a firm $j \in J$, and a transfer $p \in \mathbb{R}$ such that $j \neq \mu(i)$. We shall refer to $(\mu, \mathbf{p}, i, j, p)$ as a **coalitional deviation**, indicating that the coalition (i, j, p) is for the matching outcome $(\mu, \mathbf{p})$. Let C_M be the set of coalitional deviations for M. Formally,

$$C_M = \{(\mu, \mathbf{p}, i, j, p) : (\mu, \mathbf{p}) \in M(T), i \in I, j \in J, j \neq \mu(i), p \in \mathbb{R}\}.$$

We say $(\mu, \mathbf{p}, i, j, p) \in C_M$ is a coalitional deviation at type profile t if $(\mu, \mathbf{p}) = M(t)$ for some $t \in T$.

We would like to formalize the following intuitive idea: a coalitional deviation $(\mu, \mathbf{p}, i, j, p)$ for M at some type profile t is *viable* if the deviation is mutually beneficial to worker i and firm j, i.e., if they prefer, in the expected utility sense, a rematch with each other at the transfer p to their respective matches under $(\mu, \mathbf{p})$; a matching M is *blocked* if some coalitional deviation at some t is viable. To compare the firm's expected payoffs, we need to specify players' beliefs conditional on this coalitional deviation.

Consider a coalitional deviation $(\mu, \mathbf{p}, i, j, p)$ at $t \in T$. Worker i benefits from the deviation at t if and only if

$$a_{ij}(t) + p > a_{i\mu(i)}(t) + p_{i\mu(i)}. \tag{3.1}$$

Suppose firm j's **off-path belief** conditional on this deviation is $\beta^2_{(\mu,\mathbf{p},i,j,p)} \in \Delta(T)$. Firm j's profit from participating in the coalitional deviation $(\mu, \mathbf{p}, i, j, p)$ is

$$\mathbb{E}_{\beta^2_{(\mu,\mathbf{p},i,j,p)}}[b_{ij}] - p = \sum\nolimits_{t\in T} b_{ij}(t)\beta^2_{(\mu,\mathbf{p},i,j,p)}(t) - p. \tag{3.2}$$

The firm will also revise its expected payoff from the putative matching, using the off-path belief, to

$$\mathbb{E}_{\beta^2_{(\mu,\mathbf{p},i,j,p)}}[b_{\mu(j)j}] - p_{\mu(j)j} = \sum\nolimits_{t\in T} b_{\mu(j)j}(t)\beta^2_{(\mu,\mathbf{p},i,j,p)}(t) - p_{\mu(j)j}. \tag{3.3}$$

Firm j benefits from the coalitional deviation $(\mu, \mathbf{p}, i, j, p)$ if

$$\mathbb{E}_{\beta^2_{(\mu,\mathbf{p},i,j,p)}}[b_{ij}] - p > \max\left\{0, \mathbb{E}_{\beta^2_{(\mu,\mathbf{p},i,j,p)}}[b_{\mu(j)j}] - p_{\mu(j)j}\right\}; \tag{3.4}$$

that is, the firm must anticipate a positive payoff that is larger than what it expects to obtain in the putative matching.

Definition 2. A coalitional deviation $(\mu, \mathbf{p}, i, j, p) \in C_M$ for the matching M at $t \in T$ is **viable** with respect to an off-path belief $\beta^2_{(\mu,\mathbf{p},i,j,p)} \in \Delta(T)$ if both (3.1) and (3.4) hold. A matching M is **blocked** with respect to a system of off-path beliefs $\beta^2 = (\beta^2_{(\mu,\mathbf{p},i,j,p)})_{(\mu,\mathbf{p},i,j,p)\in C_M}$ if there exists some coalitional deviation $(\mu, \mathbf{p}, i, j, p)$ at some $t \in T$ that is viable with respect to $\beta^2_{(\mu,\mathbf{p},i,j,p)}$.

In the sequel, we make three further remarks to facilitate the reader's understanding.

Remark 4. Definition 2 describes when a coalitional deviation is mutually profitable for both parties involved. It is silent about how two players find each other and how they negotiate a transfer between them, a detail that is abstracted away in the cooperative model. The definition proposes a test for a putative matching against arbitrary counterfactual coalitional deviations, as in the existing complete-information theory. ■

Remark 5. The "max" operator in (3.4) is required for the definition of a viable coalitional deviation because the firm's expected payoff in the putative matching computed using the off-path belief $\beta^2_{(\mu,\mathbf{p},i,j,p)}$ may be negative, in which case firm j's payoff from the coalitional deviation being negative does not ensure firm j's participation in the deviation. It is tempting to argue that viability is too strong a requirement, and that the putative matching M should be viewed as defeated as long as

$$\mathbb{E}_{\beta^2_{(\mu,\mathbf{p},i,j,p)}}[b_{ij}] - p > \mathbb{E}^2_{\beta_{(\mu,\mathbf{p},i,j,p)}}[b_{\mu(j)j}] - p_{\mu(j)j}, \tag{3.5}$$

because firm j would reject its assignment $\mu(j)$ under $(\mu, \mathbf{p})$ if its updated on-path payoff (3.3) were strictly negative, regardless of whether or not firm j and worker i rematch with each other.

This argument is flawed. Note that (3.5) and (3.4) differ only when firm j's off-path payoff, (3.2), is negative (i.e., firm j will not hire the deviating worker i). In this case,

worker i's incentive to work for firm j reveals to firm j that it should reject the match outcome $(\mu, \mathbf{p})$, but worker i also understands that he will not be hired by firm j because, being more informed, he can replicate firm j's computation. Therefore, worker i will not benefit from the coalitional deviation, thus violating mutual profitability required for a valid coalitional deviation. One might still argue that firm j can still pay worker i for the purpose of soliciting information from him but will not hire him. If that is the case, all types of worker i would want to obtain the payment without actually switching to firm j and, consequently, no information would be revealed, thus defeating the purpose of making the payment in the first place.

We should carry this logic even further: in a viable coalitional deviation, firm j should assign positive probability only to those types of worker i who know they will be accepted by firm j. This is a refinement of firm j's off-path belief. It is not captured by Definition 2 but will be captured by our main refinement in Section 3.4. ■

Remark 6. We consider all counterfactual coalitional deviations for the putative matching M. But we do not consider further rounds of counterfactual deviations from the counterfactual coalitional deviations. This issue of "farsighted blocking," which makes blocking even harder (and hence leads to a coarser concept of stability), is not the focus of this paper; see, e.g., Mauleon, Vannetelbosch, and Vergote (2011) and Ray and Vohra (2015) for related discussions of von Neumann–Morgenstern stable set in complete-information problems. Incomplete information will open a new venue of research on farsightedness. Another issue is coalitions that jointly involve multiple pairs of workers and firms, which opens up more blocking opportunities. This will lead to the concept of the core in Section 6.1. ■

3.3 Stability

A **matching-belief configuration** (M, β^1, β^2) consists of a matching function M, a system of on-path beliefs $\beta^1 = (\beta^1_{(\mu,\mathbf{p},j)})_{(\mu,\mathbf{p})\in M(T), j\in J}$, and a system of off-path beliefs $\beta^2 = (\beta^2_{(\mu,\mathbf{p},i,j,p)})_{(\mu,\mathbf{p},i,j,p)\in C_M}$. We have all the ingredients needed for the definition of stability.

Definition 3. A matching-belief configuration (M, β^1, β^2) is **stable** if M is individually rational with respect to β^1 and is not blocked with respect to β^2. If (M, β^1, β^2) is stable, we say M is a **stable matching** and β^1 and β^2 are, respectively, **on-path stable beliefs** and **off-path stable beliefs** that support M.

Definition 3 formulates the consistency of matching-belief configuration.[8] If T is a singleton, the stability notion coincides with the familiar complete-information notion of stability.

[8]We only study one-sided incomplete information, but it is clear that extending this definition to two-sided incomplete information and more general environment is rather straightforward.

Example 1 in Section 1.1 illustrates the implication of stability, without appealing to any belief refinement. The example belongs to the class of games defined below, a generalization of assignment problems studied by Koopmans and Beckmann (1957) and Shapley and Shubik (1971).

Assumption 1. $b_{ij}(t_i) = b_{ij}(t_i')$ for any $t_i, t_i' \in T_i$, $i \in I$ and $j \in J$.

Assumption 1 says that the uninformed player j's matching value b_{ij} is independent of the informed player i's private types, although it can vary with their observable types that are denoted by i and j. One application of this setting is multiple-object auctions in which privately informed bidders (workers) acquire heterogeneous objects (jobs) and the object owners can have heterogeneous reservation values. A special case is $b_{ij} \equiv 0$, where the uninformed players care only about the transfers. We do not make any restrictions on a_{ij}.

Definition 4. A matching M is **full-information efficient** if for all $t \in T$ and $(\mu, \mathbf{p}) = M(t)$, the match μ maximizes $\sum_{i=1}^{n}(a_{i\mu'(i)}(t) + b_{i\mu'(i)}(t))$ over all matches $\mu' : I \cup J \to I \cup J$.

The following result is a basic test of the notion of stability, which conforms to our understanding from auction theory (e.g., the Vickrey–Clarke–Groves allocation mechanism).

Proposition 1. *Suppose that Assumption 1 holds. Then (M, β^1, β^2) is stable if and only if for any $t \in T$, $M(t)$ is complete-information stable when t is common knowledge; consequently, a stable matching M is full-information efficient.*

3.4 The Main Refinement: Bayesian Consistent Beliefs

The notion of stability can be permissive if no further restriction on beliefs is imposed.[9] Although there is no invincible argument for any refinement of beliefs, and off-path beliefs in particular, some restrictions are arguably "desirable" or "intuitive." We propose the following two principles and derive their implications:

(i) *a firm's belief should be updated using Bayes' rule from the prior belief conditional on what the firm observes and knows*;

(ii) *in a viable coalitional deviation, the deviating firm knows that the deviating worker benefits from the deviation.*

Since firms *know* the function M (by the rational expectations assumption), upon *observing* the matching outcome $(\mu, \mathbf{p})$, firms think the possible profiles of types lie in the

[9]For instance, an on-path belief assigns very small probability to a "disastrous" worker type, but all off-path beliefs assign probability 1 to the disastrous type without taking into account the type's willingness to deviate, which makes coalitional deviation unlikely and hence supports many possible matching outcomes.

set $M^{-1}(\mu, \mathbf{p}) = \{t \in T : M(t) = (\mu, \mathbf{p})\}$. Therefore, by the first principle, firm j's on-path belief is

$$\beta^1_{(\mu,\mathbf{p},j)}(\cdot) = \beta^0(\cdot|M^{-1}(\mu, \mathbf{p})). \tag{3.6}$$

As a result, firms share the same on-path belief because they have a common prior β^0, a common observation $(\mu, \mathbf{p})$, and a common understanding of M.

Consider a coalitional deviation $(\mu, \mathbf{p}, i, j, p)$ at t. Worker i (strictly) benefits from the deviation if and only if the type profile is in the following set:

$$D_{(\mu,\mathbf{p},i,j,p)} = \left\{t' \in T : a_{ij}(t') + p > a_{i\mu(i)}(t') + p_{i\mu(i)}\right\}.$$

By the second principle, for $(\mu, \mathbf{p}, i, j, p)$ to be viable, firm j knows that worker i's type is in $D_{(\mu,\mathbf{p},i,j,p)}$. Then, by the first principle, firm j's off-path belief $\beta^2_{(\mu,\mathbf{p},i,j,p)}$ is derived from the prior according to Bayes' rule conditional on what it observes and knows:

$$\beta^2_{(\mu,\mathbf{p},i,j,p)}(\cdot) = \beta^0(\cdot|M^{-1}(\mu, \mathbf{p}) \cap D_{(\mu,\mathbf{p},i,j,p)}) = \beta^1_{(\mu,\mathbf{p},j)}(\cdot|D_{(\mu,\mathbf{p},i,j,p)}). \tag{3.7}$$

When $M^{-1}(\mu, \mathbf{p}) \cap D_{(\mu,\mathbf{p},i,j,p)}$ is empty, Bayes' rule in (3.7) has no restriction and the off-path belief is arbitrary.

Definition 5. A system of on-path and off-path beliefs (β^1, β^2) associated with a matching function M is **Bayesian consistent** with the prior belief β^0 if (3.6) is satisfied for all $(\mu, \mathbf{p}) \in M(T)$ and (3.7) is satisfied for all $(\mu, \mathbf{p}, i, j, p) \in C_M$. If a matching-belief configuration (M, β^1, β^2) is stable and (β^1, β^2) is Bayesian consistent with the prior β^0, we say that (M, β^1, β^2) is **stable with Bayesian consistent beliefs** and M is a **stable matching** supported by consistent beliefs (β^1, β^2).

Given a matching function M and a prior β^0, the system of Bayesian consistent beliefs (β^1, β^2) is pinned down by Bayes' rule except when $M^{-1}(\mu, \mathbf{p}) \cap D_{(\mu,\mathbf{p},i,j,p)}$ in (3.7) is empty, in which case the coalitional deviation $(\mu, \mathbf{p}, i, j, p)$ is not viable due to worker i's lack of incentive to participate and hence the arbitrariness is inconsequential for the definition of blocking and stability. Under Bayesian consistent beliefs, the individual rationality of firm j amounts to

$$\mathbb{E}[b_{\mu(j)j}|M^{-1}(\mu, \mathbf{p})] - p_{\mu(j)j} \geq 0,$$

and firm j benefits from the coalitional deviation $(\mu, \mathbf{p}, i, j, p)$ if

$$\mathbb{E}[b_{ij}|M^{-1}(\mu, \mathbf{p}) \cap D_c] - p > \max\left\{0, \mathbb{E}[b_{\mu(j)j}|M^{-1}(\mu, \mathbf{p}) \cap D_c] - p_{\mu(j)j}\right\}.$$

Remark 7. The Bayesian consistency of on-path beliefs with prior beliefs is familiar in the

literature on rational expectations equilibrium pioneered by Radner (1979). The Bayesian consistency of off-path beliefs with prior beliefs is also natural.[10] A similar idea appears in Rothschild and Stiglitz (1976) where the off-equilibrium belief associated with an off-equilibrium contract is derived from the prior belief by conditioning on the set of types that find the contract attractive; this idea reappears in refinements of the sequential equilibrium such as the notion of the "credible updating rule" by Grossman and Perry (1986); Dutta and Vohra (2005) use a similar idea in thier concept of the credible core.

The refinement of Bayesian consistency does not utilize the information on how much different types of a worker benefit from the deviation. Section 3.5.1 offers refinements to incorporate this consideration. We can also impose additional restrictions on on-path beliefs; see Section 3.5.2. Section 6.2 incorporates correlated private on-path beliefs that are consistent with a common prior, which can be viewed as "correlated stability," as reminscient of correlated equilibrium. ■

The following is an existence result that respects all the restrictions we have made so far.

Proposition 2. *For any matching game* (a, b, β^0)*, there exists a stable matching-belief configuration* (M, β^1, β^2) *with Bayesian consistent beliefs* (β^1, β^2) *and a measurable*[11] *matching function* M*.*

As a first step of the proof, we merge all types of worker $i \in I$ that are payoff equivalent for him, and redefine each firm's matching value to respect measurability by taking the average of the firm's original matching values over these types weighted by their prior probabilities. The rest of the proof is similar to the existence proof of a rational expectations equilibrium: in the redefined matching game, take the matching M such that $M(t) = (\mu, \mathbf{p})$ is stable when t is commonly known. The matching M so defined satisfies Definition 3 even though it is not invertible. In contrast to a standard rational expectations equilibrium where general existence is difficult (see, e.g., Kreps 1977), general existence is straightforward for us. The key difference is not that between equilibrium and stability or the special structure of two-sided markets, but that the match μ (or "allocation") is publicly observable.[12]

As in the rational expectations equilibrium, a special class of stable matchings is fully revealing. We are naturally interested in the relationship between fully revealing incomplete-information stability and complete-information stability.

[10] The second principle we use to motivate off-path beliefs is reminiscent of Milgrom and Stokey (1982), where trade is assumed to occur if and only if there is common knowledge of gains from it. In the context of exchange, no trade is the same as no blocking. Therefore, it is immediate that the second principle is readily applicable beyond our setup.

[11] See Remark 3 for the definition of measurability.

[12] Jordan (1983) considers observable net trades in his formulation of rational expectations.

Definition 6. A matching function M is **fully revealing** if M is invertible.

The following result shows that a matching outcome of a fully revealing stable matching supported by Bayesian consistent beliefs must be stable when there is complete information about the type profile. This desirable result, however, relies on the refinement of both on-path and off-path beliefs.

Proposition 3. *If M is a fully revealing stable matching supported by consistent beliefs (β^1, β^2), then, for each $t \in T$, $M(t)$ is a complete-information stable matching when t is commonly known. Conversely, if $M(t)$ is a complete-information stable matching when t is commonly known for all $t \in T$, then M is a stable matching supported by Bayesian consistent beliefs; if in addition M is invertible, then M is a fully revealing stable matching.*

The intuition for the result is as follows. First, the Bayesian consistency of the on-path belief with the prior belief implies that firms' on-path beliefs in a fully revealing matching assign probability 1 to the true types. Hence, the individual rationality of incomplete-information stability is the same as that of complete-information stability. Secondly, unlike a complete-information problem where the type of a deviating worker is observed, here Bayesian consistency does not pin down the firm's off-path belief $\beta^2_{(\mu,\mathbf{p},i,j,p)}$ when $M^{-1}(\mu,\mathbf{p}) \cap D_{(\mu,\mathbf{p},i,j,p)}$ is empty. Nevertheless, the arbitrariness of the off-path belief does not support more stable outcomes than in the case of complete information, because it follows from the emptiness of $M^{-1}(\mu,\mathbf{p}) \cap D_{(\mu,\mathbf{p},i,j,p)}$ that worker i does not benefit from the coalitional deviation which is therefore not viable. Thirdly, when $M^{-1}(\mu,\mathbf{p}) \cap D_{(\mu,\mathbf{p},i,j,p)}$ is not empty, the fully revealing property of M implies that the set is a singleton, and hence it follows from the Bayesian consistency that the off-path belief $\beta^2_{(\mu,\mathbf{p},i,j,p)}$ is correct (indeed, what is needed is that $\beta^2_{(\mu,\mathbf{p},i,j,p)}$ assigns positive probability only to types in this set). Hence the blocking condition is the same as in the complete-information case.

3.5 Alternative Specifications of Beliefs

3.5.1 Off-Path Beliefs

Stability with Bayesian consistent beliefs is a leading refinement that we use for the rest of the paper. The literature on belief-based equilibrium refinements for non-cooperative games offers many ideas for off-path beliefs β^2. In spite of this connection, we reiterate that the cooperative notion of stability tests a matching against pairwise deviations, and is agnostic about how pairwise coalitions are formed and transfers are determined, as in the case of complete-information theory of stability. The refinements we offer below serve as

intuitive qualifications for viable coalitional deviations, and do not suggest specific ways of non-cooperative implementations.[13]

Support Restriction. The most obvious specification is to restrict the off-path belief to be $\beta^2_{(\mu,\mathbf{p},i,j,p)} \in \Delta(M^{-1}(\mu,\mathbf{p}) \cap D_{(\mu,\mathbf{p},i,j,p)})$. This support restriction is natural, but it is more permissive than Bayesian consistency which is a special case.

Pessimistic Belief. In addition to the support restriction, we can define β^2 to be the belief under which a coalitional deviation appears to be the least favorable to the firm. That is, for a coalitional deviation $(\mu,\mathbf{p},i,j,p)$, the corresponding off-path belief is such that

$$\beta^2_{(\mu,\mathbf{p},i,j,p)} \in \text{argmin}_{\pi \in \Delta(M^{-1}(\mu,\mathbf{p}) \cap D_{(\mu,\mathbf{p},i,j,p)})} \left(\mathbb{E}_\pi[b_{ij}] - p - \max\left\{0, \mathbb{E}_\pi[b_{\mu(j)j}] - p_{\mu(j)j}\right\}\right). \quad (3.8)$$

This off-path belief makes blocking more difficult than the Bayesian consistent belief and, consequently, it supports more stable matching outcomes.

Optimistic Belief. We can define β^2 to be the optimistic belief of the firm. That is, we replace "argmin" by "argmax" in (3.8). This off-path belief makes blocking easier and, consequently, leads to a strict refinement.

Dominance. We can consider the set of worker types that benefit the *most* from a coalitional deviation and require that the off-path belief assign positive probability only to these types.[14] Formally, consider a coalitional deviation $(\mu,\mathbf{p},i,j,p)$ and $t \in T$ with $M(t) = (\mu,\mathbf{p})$ such that $a_{ij}(t) + p > a_{i\mu(i)}(t) + p_{i\mu(i)}$. Define

$$B_{(\mu,\mathbf{p},i,j,p)} = \text{argmax}_{t' \in T}(a_{ij}(t') + p) - (a_{i\mu(i)}(t') + p_{i\mu(i)}).$$

Thus $B_{(\mu,\mathbf{p},i,j,p)}$ is the set of type profiles under which worker i benefits the most from deviating. The off-path belief assigns positive probability only to types in $B_{(\mu,\mathbf{p},i,j,p)}$:

$$\beta^2_{(\mu,\mathbf{p},i,j,p)}(\cdot) = \beta^0(\cdot | M^{-1}(\mu,\mathbf{p}) \cap B_{(\mu,\mathbf{p},i,j,p)}).$$

Whether these restrictions make blocking easier or more difficult depends on whether the firm's preference is aligned with the worker's.[15] If the matching value is such that a deviation

[13] Many belief-based refinements of sequential equilibria do not suggest how restrictions on beliefs arise in larger games, but, instead, they capture our intuition about how these games are expected to be played.

[14] This is in spirit related to the refinement idea of dominance. For example, the idea behind D1 (Cho and Kreps 1987) is to compare the sets of responses of the uninformed players upon deviating. In a matching game, a firm's "response" is simply the decision of whether to join the coalition, and hence we cannot directly apply the existing formulation of dominance. One approach is to consider the maximal set of prices p' that induce blocking by certain worker types. This leads us to the formulation presented here, because, due to transferability, the types that benefit the most from deviating have the maximal set of prices p' under which (i,j,p') may block.

[15] Traditional dominance-based refinements are developed for signaling games with aligned preferences.

is more attractive to the firm whenever it is more attractive to the worker, this off-path belief will make blocking easy.

Tremble-Based Refinements. We can make all off-path events on-path using trembles, and then consider their limits, as in sequential equilibrium. This approach can also be used to model beliefs implicitly through the likelihood of trembles across different types, as in proper equilibrium.

Further Ideas. There are still many ways to refine the notion of stability. For instance, one plausible scenario is that to decide whether to join a coalitional deviation $(\mu, \mathbf{p}, i, j, p)$, firm j makes the assumption that any other coalitional deviation $(i, j', p', \mu, \mathbf{p})$ with $j' \neq j$ is less attractive to worker i. The refinement literature teaches us that it would be a Sisyphean task to capture all reasonable ideas of refinements in a single definition, and selections of these different notions depend on the economic applications, which is better left for future research. What is essential is that the separation of on-path and off-path beliefs provides a framework for model refinements and enables a coherent discussion of the (im)plausibility of stable matching in a purely cooperative framework without mixing cooperative and non-cooperative elements together.

3.5.2 On-Path Beliefs

In our refinement, firms share a common on-path belief. In some applications, a worker's employer observes more about the worker's payoff-relevant attributes than other firms do.[16] It is plausible that firms observe more about their workers in the continuation of the employment relationship, leading to further market movements, even though this information is not known before the finalization of the initial job matches (and hence cannot be used to define the stability of the initial job allocations). This additional information can be used to define the stability of the market at this later stage. Our framework can accommodate this, which merely amounts to an additional restriction on the on-path beliefs.

Formally, for each $i \in I$, let $T_i = T_i^1 \times T_i^2$, where the set T_i^1 denotes the set of attributes directly observable to worker i's employer, and T_i^2 denotes types only observable to worker i. A type of worker i is $t_i = (t_i^1, t_i^2)$, and a profile of workers' types is $t = (t_i, t_{-i})$. Consider a putative matching $M : t \to (\mu, \mathbf{p})$. After observing $(\mu, \mathbf{p})$ and the observable attribute $t^2_{\mu(j)}$

[16] For instance, Liu et al. (2014) and the literature that follows make the assumption that firms know perfectly their own workers' types in a putative match. Thus, individual rationality is ex post and a deviating firm's payoff is also compared to its ex post payoff in a putative matching. This assumption of perfect observability circumvents the difficulty of defining payoffs in a stable matching even though information in the stable matching should be endogenous; in addition, the assumption gives strength to the permissive concept of Liu et al. (2014). The present paper does not make the assumption, and the conceptual difficulty of defining payoffs in a stable matching is resolved by the notion of on-path belief and its joint determination with stability.

of its assigned worker $\mu(j)$, firm j's *private* on-path belief at $t \in T$ is

$$\beta^1_{(\mu,\mathbf{p},j,t)}(\cdot) := \beta^0(\cdot|M^{-1}(\mu,\mathbf{p}) \cap (T^1_{\mu(j)} \times \{t^2_{\mu(j)}\} \times T_{-\mu(j)})) \in \Delta(T).$$

Firm j's *private* off-path belief at $t \in T$ at a coalitional deviation $(\mu, \mathbf{p}, i, j, p)$ is

$$\beta^2_{(\mu,\mathbf{p},i,j,p,t)}(\cdot) := \beta^0(\cdot|M^{-1}(\mu,\mathbf{p}) \cap (T^1_{\mu(j)} \times \{t^2_{\mu(j)}\} \times T_{-\mu(j)}) \cap D_{(\mu,\mathbf{p},i,j,p)}),$$

where $D_{(\mu,\mathbf{p},i,j,p)} = \left\{t' \in T : a_{ij}(t') + p > a_{i\mu(i)}(t') + p_{i\mu(i)}\right\}$.

By definition, $\beta^1_{(\mu,\mathbf{p},j,t)}$ and $\beta^2_{(\mu,\mathbf{p},i,j,p,t)}$ depend on t only through $t^2_{\mu(j)}$. Individual rationality, blocking, and stability can be defined with respect to $\beta^1 = (\beta^1_{(\mu,\mathbf{p},j,t)})$ and $\beta^2 = (\beta^2_{(\mu,\mathbf{p},i,j,p,t)})$ in the same way as in Definitions 1, 2, and 3, respectively.

For applications, the right assumption on what an uninformed firm can observe depends ultimately on the market situation we want to analyze. For instance, in the market for junior economists, matches are formed and the market clears before employers know perfectly the actual types of job candidates. If we are interested in the stability of markets at this stage, it is not reasonable for us to assume full information revelation within a matched pair because this extra information is not used to stabilize the market in the first place; instead, the relevant stability notion should be defined without the uninformed players' uncertainties about their ex post payoffs being exogenously assumed away, although belief updating through indirect inference must be a component of stability. The same is true for assignment problems where sellers do not directly observe the buyers' types when the market clears, although indirect inference can be made.

4 Structural Properties of Stability and Bayesian Consistency

We are interested in the (in)efficiency of matchings for the following reasons. First, Shapley and Shubik (1971) observe that, under complete information, *all* stable matching outcomes maximize the sum of individual payoffs, although indeterminacy of transfers and payoff distributions is generally inevitable. The full-information efficiency criterion introduced in Definition 4 is too demanding for incomplete-information problems. It is thus worthwhile to explore how and to what extent this robust allocative efficiency property identified by Shapley and Shubik (1971) extends to incomplete information. Secondly, our definition of stability with incomplete information is a joint requirement for beliefs and matching outcomes; hence, efficiency is a joint prediction of endogenous matches and information in addition to being

a welfare property. Indeed, we shall see that stability and competitive equilibrium, two outcome-equivalent concepts under complete information, impose very different restrictions.

We must clarify the environments in which the efficiency criterion applies. Take as an example the market for lemons. In traditional models of adverse selection, the market opens only once and further interactions are excluded by the exogenous restriction of an endgame, whereas in our model of stability, the option of trade (rematch) is always available although players may choose not to exercise it (see Example 3). Thus stability and the prior literature examine different scenarios of adverse selection. More concretely, if there is partial trade in a traditional model of a market for lemons, beliefs will be updated after partial trade. An updated belief is irrelevant in traditional models of adverse selection because the market will not reopen. If market interactions continue, the updated belief will open up further trading opportunities. An analysis of a discounted dynamic lemon market with unlimited trading opportunities is provided by Deneckere and Liang (2006) who show that gains from trade are eventually realized with probability 1, albeit slowly. That is, although allocative efficiency is achieved in the long run, inefficiency takes the form of delay over time. We propose stability as a shortcut to the limiting case in dynamic games, and efficiency of stable matchings concerns only the allocative efficiency of this limit. This does not mean that efficiency is easy to achieve. If there is a complete breakdown of trade as in Akerlof's (1970) original model, there will be no belief revision, and the inefficient no-trade outcome will be stable according to our definition.

4.1 Criterion of Bayesian Efficiency

The full-information efficiency introduced in Definition 4 is obviously too demanding for incomplete-information problems. We propose the following notion.

Definition 7. A matching M is **Bayesian efficient** if for all $(\mu, \mathbf{p}) \in M(T)$, the match μ maximizes

$$\mathbb{E}\left[\sum_{i=1}^{n}(a_{i\mu'(i)} + b_{i\mu'(i)})|M^{-1}(\mu, \mathbf{p})\right]$$

over all matchings $\mu' : I \cup J \to I \cup J$.

At the risk of belaboring the obvious, we make the following comments.

Remark 8. Bayesian efficiency differs from full-information efficiency in only one respect: it concerns surplus maximization conditional on each $M(t)$ instead of each $t \in T$. It then follows immediately that a full-information efficient matching M is Bayesian efficient. ■

Remark 9. In Definition 7, $\mathbb{E}\left[\sum_{i=1}^{n} a_{i\mu'(i)}|M^{-1}(\mu, \mathbf{p})\right]$ should not be interpreted as the workers' expected total surplus conditional on the information revealed by the matching outcome,

because the workers know their own types and ex post payoffs. This expectation is from the viewpoint of an *outside observer* whose probability distribution over T is the prior β^0 conditional on the publicly observable outcome $(\mu, \mathbf{p})$. It is, of course, also the workers' surplus computed from the firms' perspective, because firms' consistent on-path beliefs about the workers are *correct* in stable matching with Bayesian consistent beliefs. ■

Remark 10. Bayesian efficiency is motivated by a the following question: when we economists observe matching outcomes in the data, and correctly update the distribution over the underlying types (assuming that we are as uninformed as the firms in the model and the data are generated by a stable matching), can we conclude that the observed match must maximize, among all possible matches, the expected social surplus computed using the updated distribution? This question is precisely about the criterion of Bayesian efficiency. Similar notions of efficiency are proposed by Forges (1994) to take into account information revealed by outcomes of a mechanism. ■

Bayesian inefficiency can persist and no pairwise recontracting arrangement can correct it (this is in contrast to the competitive equilibrium notion studied in Section 5 where inefficiency can be corrected if deviation is unilaterally). The following is an example of Bayesian inefficient stable matching.

Example 4. There are two workers and one firm. Worker 1's type is t_1 or t_1' with equal probability. Worker 2's type is known to be t_2. Suppose that the matching values are as follows:

t_1	t_1'	t_2
$(-1, 5)$	$(1, -2)$	$(0, 1)$

Here $(-1, 5)$ means that by matching with the firm, worker 1 of type t_1 obtains a payoff of -1, and the firm obtains a payoff of 5.

Consider the matching M in which the firm hires worker 2 at a price of 0, regardless of worker 1's type. The total surplus is 1. This matching is stable (even with Bayesian consistent beliefs) for the following reason. Any coalitional deviation acceptable to the firm must involve worker 1 of type t_1, and thus the price p must be at least 1 to satisfy the individual rationality of type t_1. But this price will attract both types of worker 1. So the firm's expected payoff from the blocking with worker 1 will be $\frac{1}{2} \times 5 + \frac{1}{2} \times (-2) - p \leq 0.5$, which is less than its payoff in the matching with worker 2.

This matching is not Bayesian efficient. It is dominated by a match μ' in which the firm is matched with worker 1, which yields an expected total surplus of 1.5. But μ' cannot be part of a Bayesian stable matching with consistent beliefs for any price: the firm must pay at least 1 to worker 1 (by the individual rationality of type t_1) and thus its expected payoff

is at most 0.5; but the firm can block the matching with the unmatched worker 2 to obtain a larger payoff.

In this example, the firm needs to pay a high price to recruit worker 1 of type t_1 (who is more productive for the firm), but transfers between players are not counted toward the social surplus. Thus the source of social inefficiency is the usual conflict with individual incentives. ■

4.2 Bayesian Efficiency and Stability

We are interested in conditions under which *all* stable matchings of a given matching game (a, b, β^0) are Bayesian efficient. In particular, we shall make no assumptions on b_{ij} in order to include adverse selection problems as special cases.

Assumption 2. $a_{ij}(t_i) = a_{ij}(t_i')$ for any $t_i, t_i' \in T_i$, $i \in I$, and $j \in J$.

Assumption 2 says that the privately informed players do not directly care about their own types, which are payoff-relevant for the uninformed players (the informed players care about their types indirectly because they affect the matching outcomes). A special case that is of applied interest is when $a_{ij} \equiv 0$. This case captures a situation in which workers care only about the salaries they receive.

A weaker assumption is that all public and private attributes are directly payoff-relevant for the informed players, but $a_{ij}(t_i)$ is separable in t_i and j.

Assumption 3. $a_{ij}(t_i) = g(i, t_i) + h(i, j)$ for some functions $g : I \times T_i \to \mathbb{R}$ and $h : I \times J \to \mathbb{R}$.

A special case of Assumption 3 is familiar in many classic adverse-selection models such as signaling and screening: $a_{ij}(t_i) = g(i, t_i)$. This is to say, a worker does not value which firm he works for, but his own types may affect his reservation utilities or costs of effort, etc. This assumption allows $a_{ij}(t_i)$ to vary with the worker's private type t_i and the worker's identity i which summarizes all of his observable attributes, but the value is not allowed to vary with the firm's type, which is summarized in j.

The following result concerns Bayesian efficiency of stable matchings. Its proof is based on the duality theorem of linear programming. Unlike the case of complete information (e.g., Shapley and Shubik 1971), the proof is not immediate because surplus maximization and its dual are defined by on-path beliefs, but stability and blocking utilize off-path beliefs. We say workers are *fully matched* under M if $\mu(i) \neq i$ for all $i \in I$ and $(\mu, \mathbf{p}) \in M^{-1}(T)$.

Proposition 4. *A stable matching M supported by Bayesian consistent beliefs is Bayesian efficient if one of the following properties is satisfied:*

(i) *Assumption 2 holds.*

(ii) *Assumption 3 holds and workers are fully matched.*

We leave the full-match condition in the statement because (ii) can be reinterpreted as follows: if Assumption 3 holds, then constrained Bayesian efficiency obtains for all stable matching outcomes if it is restricted to matched agents. This condition excludes the no-trade outcome in Akerlof (1970). The full-match condition is ensured, for example, if workers are on the short side of the market and matching values are positive (or more generally there exists a price p such that $a_{ij}(t_i) + p > 0$ and $b_{ij}(t_i) - p > 0$ for all $t_i \in T_i$, $i \in I$ and $j \in J$). Assumptions on the short side of the market being fully matched are common; see Ashlagi, Kanoria, and Leshno (2017) and the references therein for discussions of unbalanced matching markets. Example 4 shows that the full-match restriction in condition (ii) is tight. It is easy to construct examples where Assumption 2 and Assumption 3 cannot be dispensed with.

A natural question is what happens when firms are on the short side of the market, and hence workers cannot be fully matched—in which case condition (ii) of Proposition 4 does not apply. Proposition 4 makes assumptions only on the payoffs (a, b) and Bayesian efficiency is obtained regardless of prior belief β^0. Since on-path beliefs play an important role in the definition of Bayesian efficiency, it is natural to think of restrictions on beliefs. If $\beta^0(t) = \prod_{i=1}^{n} \beta_i^0(t_i)$ for all $t = (t_1, ..., t_n) \in T$, where β_i^0 is the marginal of β^0 on T_i, we say that workers' types are **independent** under the prior β^0. In dynamic non-cooperative games in which types are independent under prior beliefs, it is common to assume that types remain independent *after any history* (see Fudenberg and Tirole 1991, p. 237). Naturally, we shall consider **independent** on-path beliefs after any observables; that is, workers' types are independent under $\beta^0(\cdot | M^{-1}(\mu, \mathbf{p}))$ for all $(\mu, \mathbf{p}) \in M(T)$.

Proposition 5. *A stable matching M with Bayesian consistent beliefs is Bayesian efficient if Assumption 3 holds, a_{ij} and b_{ij} are co-monotonic[17] for all $i \in I$ and $j \in J$, and the on-path beliefs are independent.*

The order over T_i with respect to which a_{ij} and b_{ij} are co-monotonic may vary (i, j). Thus this condition is weak. Co-monotonicity is an intuitive property in the special case where t_i is a real variable that ranks the worker's ability according to the total order of "greater than or equal to," a_{ij} can be interpreted as worker i's disutility from work, and b_{ij} can be interpreted as the output. The monotonicity of a_{ij} and b_{ij} says that the worker's disutility is decreasing in his ability and his output is increasing in his ability. Note, however,

[17] Two functions a_{ij} and b_{ij} are co-monotonic if $(a_{ij}(t_i) - a_{ij}(t_i'))(b_{ij}(t_i) - b_{ij}(t_i')) \geq 0$ for any $t_i, t_i' \in T_i$.

that in a lemon market (Akerlof 1970), co-monotonicity is not satisfied. The independence assumption in Proposition 5 cannot be relaxed, as the following example illustrates.

Example 5. Consider a market with two workers and one firm. The matching values of each worker and the firm are co-monotonic, and are as follows:

t_1	t_1'	t_2	t_2'
$(0.5, 5)$	$(1, 6)$	$(-2, 4)$	$(-1.9, 12)$

Suppose that $\beta^0(t_1, t_2) = \beta^0(t_1', t_2') = \frac{1}{2}$. Thus, the workers' types are not independent.

Consider a matching M in which the firm hires worker 2 at a price of 2 regardless of the workers' types. In this case, the Bayesian consistent on-path belief is the same as the prior belief β^0. This matching is not Bayesian efficient: it generates an expected total surplus of $\frac{1}{2} \times (-2 + 4) + \frac{1}{2} \times (-1.9 + 12) = 6.05$, while the matching in which the firm hires worker 1 generates an expected total surplus of $\frac{1}{2} \times (0.5 + 5) + \frac{1}{2} \times (1 + 6) = 6.25$.

But the matching M is stable with Bayesian consistent beliefs. The firm's expected payoff in this matching is $\frac{1}{2} \times 4 + \frac{1}{2} \times 12 - 2 = 6$. Consider a deviating coalition that involves the firm and worker 1 with a price p. No price p is such that only the type t_1 of worker 1 joins the coalition. If the price p is such that both types of worker 1 join the coalition, i.e., $p > -0.5$, then the firm's expected payoff is $\frac{1}{2} \times 5 + \frac{1}{2} \times 6 - p < 6$. In this case the firm rejects the coalition. If the price p is such that only the type t_1' of worker 1 joins the coalition, then the firm's payoff cannot be higher than 7, the total surplus produced by the pair. But because the two workers' types are correlated, when worker 1's type is t_1', worker 2's type must be t_2', and the firm infers that its payoff from M by matching with worker 2 is $12 - 2 = 10$. Therefore, the firm rejects the coalition with worker 1 in this case as well. ■

5 Competitive Equilibrium

5.1 Motivation and Definition

For complete information matching and assignment problems, Koopmans and Beckmann (1957) and Shapley and Shubik (1971) construct the following notion of competitive equilibrium. Each partnership $(i, j) \in I \times J$ is viewed as one unit of an indivisible commodity, and there is a price p_{ij} associated with each commodity, irrespective of whether i and j are matched or not in equilibrium. Let $\mathbf{p} = (p_{ij})_{i \in I, j \in J}$ denote the price matrix. We also define $p_{ii} = p_{jj} = 0$ for all $i \in I$ and $j \in J$. In a competitive equilibrium $(\mu, \mathbf{p})$, each individual player is maximizing in the sense that he does not profit from staying alone, or from switch-

ing to any other player on the opposite side of the market at the competitive price specified by $\mathbf{p}$ (i.e., demand the commodity (i, j) at a price p_{ij}).

The matching mechanism described by a competitive equilibrium has two critical differences from stable matching. First, a player's acceptability to the other player is not taken into account in defining a profitable deviation; that is, deviation is **unilateral**. Second, if a player deviates to another player, the price between them is determined by the competitive equilibrium price $\mathbf{p}$; that is, players are **price takers**. In spite of these disparities, Shapley and Shubik (1971, pp. 114–118) point out that competitive equilibrium and stability are equivalent in their model of complete information. We shall study how the assumptions of **unilateral deviation** and **price-taking behavior** manifest under incomplete information.

A natural notion of a competitive equilibrium in an economy with uncertainty and without state-contingent contracts is the rational expectations equilibrium of Radner (1979).[18] We now construct such a notion for two-sided matching markets.

A competitive matching is a function $M : t \mapsto (\mu, \mathbf{p})$, where $\mathbf{p} = (p_{ij})_{i \in I, j \in J}$. We may impose the same **measurability** condition on M as in Remark 3. Both the match μ and the commodity prices $\mathbf{p}$ are publicly observable. Upon observing $(\mu, \mathbf{p})$, players will update their prior belief to the on-path belief $\beta^0(\cdot | M^{-1}(\mu, \mathbf{p}))$, where $M^{-1}(\mu, \mathbf{p}) = \{t \in T : M(t) = (\mu, \mathbf{p})\}$.

Definition 8. A matching $M : t \mapsto (\mu, \mathbf{p})$ is a (rational expectations) **competitive equilibrium** if the following conditions hold for all $t \in T$ and $(\mu, \mathbf{p}) = M(t)$:

(i) $a_{i\mu(i)}(t) + p_{i\mu(i)} \geq a_{ij}(t) + p_{ij}$ for all $i \in I$ and $j \in J \cup \{i\}$;

(ii) $\mathbb{E}[b_{\mu(j)j} | M^{-1}(\mu, \mathbf{p})] - p_{\mu(j)j} \geq \mathbb{E}[b_{ij} | M^{-1}(\mu, \mathbf{p})] - p_{ij}$ for all $j \in J$ and $i \in I \cup \{j\}$.

A competitive equilibrium satisfies individual rationality: take $j = i$ in (i) and $i = j$ in (ii). Notice also that only the on-path belief $\beta^0(\cdot | M^{-1}(\mu, \mathbf{p}))$ is utilized in the definition, because only unilateral deviation is involved. When T is a singleton, this definition reduces to the familiar notion of competitive equilibrium under complete information.

Proposition 6. *A (measurable) competitive equilibrium matching exists for each matching game* (a, b, β^0).

As in Proposition 2, general existence is straightforward because the matching outcomes are publicly observable. In a two-sided market, the assumptions behind stability look more appealing than those behind competitive equilibrium. We should emphasize that, despite its previous usage, we define competitive equilibrium for the purpose of comparison rather than as a competing concept.

[18] Complete state-contingent contracts bring the problem back to complete information. The Arrow–Debreu formulation of competitive equilibrium under uncertainty is not a suitable solution concept for our purposes.

5.2 Stability and Competitive Equilibrium

Stability and competitive equilibrium are two different ways of looking at a matching problem. A stable matching outcome $(\mu, \mathbf{p}^s)$ does not specify a price for an unmatched pair (i, j), where $\mu(i) \neq j$, while the price matrix $\mathbf{p}^c$ for a competitive matching outcome does specify a price for every pair (i, j). The observability of the price matrix $\mathbf{p}^c$ may seem to suggest that prices in a competitive equilibrium matching $t \mapsto (\mu, \mathbf{p}^c)$ reveal more information than prices in a stable matching $t \mapsto (\mu, \mathbf{p}^s)$ do. This intuition is *incorrect*, because it focuses literally on on-path beliefs but ignores the fact that stability makes restrictions directly on off-path beliefs and hence indirectly on on-path beliefs. The difference between stability and competitive equilibrium thus has to stem from the *incentives* and *information* embedded in their definitions.

Definition 9. A stable matching M^s **extends** to a competitive matching M^c if for each $t \in T$, the matching outcomes $M^s(t) = (\mu^s, \mathbf{p}^s)$ and $M^c(t) = (\mu^c, \mathbf{p}^c)$ share the same match, $\mu^s = \mu^c = \mu$, and $\mathbf{p}^s$ and $\mathbf{p}^c$ agree on the matched pair $(i, \mu(i))$ for all $i \in I$. In this case, we say that M^c is an **extension** of M^s.

We present an example in which a competitive equilibrium matching cannot be an extension of a stable matching supported by Bayesian consistent beliefs.

Example 6. Consider a market with two workers and one firm. Worker 1's type is known to be t_1. Worker 2's type is t_2 or t_2' with equal probability. The matching values are as follows:

t_1	t_2	t_2'
$(1, 5)$	$(1, -4)$	$(2, 4)$

The following matching is a competitive equilibrium: the firm hires worker 1 with a price of $p_{11} = 0$, and worker 2 is unmatched regardless of his type; the price for the firm to hire worker 2 is $p_{21} = -3$. By deviating to worker 2, the firm's expected payoff is $\frac{1}{2} \times (-4) + \frac{1}{2} \times 4 - (-3) = 3$. Hence the firm does not deviate. By working for the firm, type t_2 obtains a payoff of $1 - 3 = -2$ and type t_2' obtains a payoff of $2 - 3 = -1$. Hence neither type of worker 2 deviates. Therefore, this matching is a competitive equilibrium.

The matching outcome of this competitive equilibrium cannot be stable with Bayesian consistent beliefs. Worker 2 with type t_2' and the firm could block with a price of -1.5. Type t_2 will earn a negative payoff from this match and type t_2' will earn a positive payoff. The firm will infer the worker's type correctly and hire him to obtain a payoff of 5.5. ■

The above example demonstrates that flexible off-path prices allow for more information revelation, so one would conjecture that stability refines competitive equilibrium. This

is again *incorrect*. The key is that having more information does not necessarily facilitate blocking when the rematch is ex post undesirable; unilateral deviation may still be possible in a competitive environment with less information revelation. We confirm this point by providing an example where a stable matching cannot be extended to a competitive equilibrium.

Example 7. Consider a market with two workers and one firm, where worker 1's type is known to be t_1, and worker 2's type is t_2 or t_2' with equal probability. The matching values are as follows.

t_1	t_2	t_2'
$(1,5)$	$(2,1)$	$(1,6)$

The following is a stable matching supported by Bayesian consistent beliefs: the firm hires worker 1 for a price of 0, and worker 2 is unmatched regardless of his type (with a payoff of 0). The firm's payoff is 5. We now argue that the firm cannot block the matching with worker 2 for any price p. If $p \leq -2$, neither type of worker 2 deviates; if $p \in (-2,-1]$, only type t_2 deviates, and the firm's payoff from rematching with t_2 is $1-p \leq 3$; if $p > -1$, both types of worker 2 deviate, and the firm's expected payoff from the deviation is $\frac{1}{2}\times 1+\frac{1}{2}\times 6-p < 4.5$. Therefore, the firm does not deviate.

This stable matching cannot be extended to a competitive equilibrium for any prespecified price between the firm and worker 2. If $p > -2$, one or both types of worker 2 deviate. If $p \leq -2$, the firm's expected payoff from deviating to worker 2 is $\frac{1}{2}\times 1+\frac{1}{2}\times 6-p \geq 5.5$; hence the firm deviates. ■

Although stable matchings and competitive matchings are generally not the same, they must overlap. We summarize our finding in the following result.

Proposition 7. (i) *For any matching game (a,b,β^0), there exists a stable matching M^s with Bayesian consistent beliefs that can be extended to a competitive matching. There exists a matching game (a,b,β^0) with a stable matching M^s with Bayesian consistent beliefs that cannot be extended to a competitive matching.* (ii) *For any matching game (a,b,β^0), there exists a competitive matching M^c that is an extension of a stable matching with Bayesian consistent beliefs. There exists a matching game (a,b,β^0) with a competitive matching M^c that is not an extension of a stable matching with Bayesian consistent beliefs.*

5.3 Bayesian Efficiency of Competitive Equilibrium

Given a competitive equilibrium matching $M : t \mapsto (\mu, \mathbf{p})$, the notions of full-information efficiency and Bayesian efficiency can be reproduced verbatim from Definition 4 and Definition 7, respectively, by taking into account $\mathbf{p} = (p_{ij})_{i\in I, j\in J}$. Remarks 8–10 apply here.

Recall that a stable matching is not guaranteed to be Bayesian efficient. By contrast, a competitive equilibrium matching is always Bayesian efficient. This result is reminiscent of the first fundamental theorem of welfare economics. The logic is as follows: if there is overall inefficiency conditional on the information revealed in a matching, at least some player is inefficiently matched, and this player can correct this inefficiency by a unilateral rematch, *under the same information*. The contrast with stability is notable: the new information generated from a blocking pair can prevent the inefficiency from being corrected. We would like to reiterate that the efficiency is about allocative efficiency, but it does not take into account what it takes to achieve it, as is similar to Bayesian efficiency of stable matching.

Proposition 8. *A competitive equilibrium matching $M : t \mapsto (\mu, \mathbf{p})$ is Bayesian efficient. If Assumption 1 holds, then a competitive equilibrium matching M is full-information efficient and $M(t)$ is a complete-information competitive equilibrium matching when t is common knowledge for all $t \in T$.*

We should emphasize that the result does not imply that a competitive equilibrium has a better welfare property than a stable matching, because the amount of information that is revealed may be different and Bayesian efficiency is defined relative to information.

6 Extensions

6.1 The Core

Pairwise deviations are natural in two-sided markets. Conceptually, it is useful to consider deviations by a coalition of multiple pairs of firms and workers. Given a matching $M : t \mapsto (\mu, \mathbf{p})$, suppose that $(\mu, \mathbf{p}) = M(t)$ is a matching outcome at $t \in T$. Each firm j should have an on-path belief $\beta^1_{(\mu,\mathbf{p},j)}$ associated with this outcome. Consider the following blocking possibility: a subset of workers $I' \subset I$ and a subset of firms $J' \subset J$ walk away from $(\mu, \mathbf{p})$ and rematch among themselves according to $\mu' : I' \cup J' \to I' \cup J'$ and a transfer scheme $\mathbf{p}' = (p'_{i\mu'(i)})_{i \in I'}$ associated with the match μ', where μ' is not the same as μ restricted to $I' \cup J'$.[19] We call μ' a **rematch** relative to μ. We write this coalitional deviation by $c = (\mu, \mathbf{p}, I', J', \mu', \mathbf{p}')$. Each firm $j \in J'$ should have an off-path belief $\beta^1_{(c,j)}$ associated with this deviating coalition. Let us denote a matching-belief configuration by (M, β^1, β^2) where β^1 is the system of on-path beliefs and β^2 is the system of off-path beliefs.

Individual rationality of a matching M with respect to the system of on-path belief β^1 is defined as in Definition 1. The blocking condition is defined below.

[19] We have assumed that a player receives transfers only from his matched partner. A relaxation is straightforward.

Definition 10. A matching M is **blocked** with respect to a system of off-path beliefs β^2 if there does exist a coalitional deviation $c = (\mu, \mathbf{p}, I', J', \mu', \mathbf{p}')$, where $(\mu, \mathbf{p}) = M(t)$ for some $t \in T$, $I' \subset I$, $J' \subset J$, $\mu' : I' \cup J' \to I' \cup J'$ is a rematch, and $\mathbf{p}' = (p'_{i\mu'(i)})_{i \in I'}$ is a transfer scheme associated with the rematch μ', such that

(i) $a_{i\mu'(i)}(t) + p'_{i\mu'(i)} > a_{i\mu(i)}(t) + p_{i\mu(i)}$ for all $i \in I'$, and

(ii) $\mathbb{E}_{\beta^2_{(c,j)}}\left[b_{\mu'(j)j}\right] - p'_{\mu'(j)j} > \max\left\{0, \mathbb{E}_{\beta^2_{(c,j)}}\left[b_{\mu(j)j}\right] - p_{\mu(j)j}\right\}$ for all $j \in J'$.

Condition (ii) needs a remark. The formulation implicitly excludes the possibility that $\mu'(j) = j$ for some $j \in J'$; i.e., j joins the coalitional deviation but stays unmatched in μ', because otherwise the left-hand side of condition (ii) becomes 0, thus violating the condition. This exclusion is without loss of generality because an unmatched firm j does not contribute any information or value to the coalitional deviation.

Definition 11. A matching-belief configuration (M, β^1, β^2) is in the **core** if M is individually rational with respect to the system of on-path beliefs β^1 and is not blocked with respect to the system of off-path beliefs β^2. We also say M is a **core matching** supported by (β^1, β^2) if (M, β^1, β^2) is a stable configuration.

The refinement of Bayesian consistent beliefs also has a counterpart: beliefs are updated from the prior conditional on players' observations and the information revealed by their incentive to participate in the coalitional deviation.

Definition 12. A system of on-path and off-path beliefs (β^1, β^2) associated with a matching function M is **Bayesian consistent** with the prior belief β^0 if $\beta^1_{(\mu,\mathbf{p},j)} = \beta^0(\cdot|M^{-1}(\mu, \mathbf{p}))$ for each $j \in J$ and $(\mu, \mathbf{p}) \in M(T)$, and $\beta^2_{(c,j)} = \beta^0(\cdot|M^{-1}(\mu, \mathbf{p}) \cap D_c)$, where

$$D_c = \left\{t' \in T : a_{i\mu'(i)}(t') + p'_{i\mu'(i)} > a_{i\mu(i)}(t') + p_{i\mu(i)} \text{ for all } i \in I'\right\}$$

for each deviating coalition $c = (\mu, \mathbf{p}, I', J', \mu', \mathbf{p}')$ and $j \in J'$. If (M, β^1, β^2) is in the core and (β^1, β^2) is Bayesian consistent with the prior β^0, we say that M is a **core matching supported by Bayesian consistent beliefs**.

It should be noted that D is the set of workers' types $(t_1, ..., t_n)$ with which all workers in I' find the rematch profitable. It does not take into account the incentives of firms in the set $J' \setminus \{j\}$ because these firms are uninformed and their incentives to block reveal no information unknown to firm j (firm j can replicate their calculation).

In complete-information matching games, the core and stability coincide, but they differ under incomplete information.

Proposition 9. *If M is a core matching supported by Bayesian consistent beliefs, then it is a stable matching supported by Bayesian consistent beliefs; however, a stable matching M supported by consistent beliefs is not necessarily a core matching supported by consistent beliefs.*

One direction is straightforward. Individual rationality is the same for stability and the core. A pairwise coalition $(\mu, \mathbf{p}, i, j, p)$ is a special coalition $(\mu, \mathbf{p}, I', J', \mu', \mathbf{p}')$ with $I' = \{i\}, J' = \{j\}, \mu'(i) = j$, and $p'_{ij} = p$. Specifically, if (M, β^1, β^2) is in the core, then it is not blocked by any coalition including a pairwise coalition; hence, $(M, \beta^1, \overline{\beta}^2)$ is stable, where $\overline{\beta}^2$ is a restriction of β^2 to pairwise coalitions. This property does not rely on belief refinements. The following example demonstrates the subtle reason that the core is a strict refinement of stability even when β^0 is independent: a blocking by a larger coalition can be found when a pairwise blocking does not exist. The example has a pair of a firm and a worker who are matched together in the given matching, but both deviate to rematch with other players. It is precisely its own worker's incentive to join the coalitional deviation that reveals to the firm that its payoff from the putative matching is actually lower than it has thought, which incentivizes the firm to rematch with the other worker; meanwhile, the deviation of the firm's own worker is made possible precisely for the same reason: the other firm accepts him because the other worker's deviation reveals information. This existence of this four-player cycle refines stability.

Example 8. Consider two workers and two firms. Suppose that $\beta^0 = \beta^0_1 \times \beta^0_2$, where $\beta^0_1(t_1) = \beta^0_1(t'_1) = \beta^0_2(t_2) = \beta^0_2(t'_2) = \frac{1}{2}$. The matrix of matching values is as follows:

	firm 1	firm 2
t_1	$0, -1$	$1, 1$
t'_1	$1, 7$	$-2, 0$
t_2	$1, 1$	$0, -1$
t'_2	$-2, 0$	$1, 7$

It is readily verified that a_{ij} and b_{ij} are co-monotonic.

Consider the following matching M: regardless of their types, worker i is assigned to firm $j = i$, and the salaries of both workers are 0. In this matching, the expected payoffs for both firms are $\frac{1}{2} \times (-1) + \frac{1}{2} \times 7 = 3$. The matching M is stable with Bayesian consistent beliefs for the following reason. Let us consider pairwise deviation by worker i and firm $j = 3 - i$. For the firm to join the deviation, its expected payoff from the deviation must be more than 3, but the total surplus from a match with worker i cannot exceed 2 regardless of the worker's type.

But M is not in the core with Bayesian consistent beliefs. A viable coalitional deviation involves a rematch of both firms and both workers when their types are t_i with a transfer of 0. Given that each worker $i = 1, 2$ finds it profitable to deviate to firm $j = 3 - i$ with a price of 0, both firms infer that worker $i = 1, 2$ must have type t_i instead of t_i'. With this information, firm $j = i$ knows that its payoff in the matching M is actually -1. For this reason, firm i is willing to accept worker $3 - i$.

The refinement of the off-path beliefs is used only in that its support should be the set of types that benefit from the deviations. ■

The following is an immediate corollary of Propositions 1, 4, and 9.

Corollary 1. *Suppose that (M, β^1, β^2) is in the core. Then M is full-information efficient if Assumption 1 holds. Suppose further that (β^1, β^2) is Bayesian consistent. Then M is Bayesian efficient if one of the following properties is satisfied:*

(i) *Assumption 2 holds.*

(ii) *Assumption 3 holds and workers are fully matched.*

(iii) *Assumption 3 holds, a_{ij} and b_{ij} are co-monotonic in t_i for all $i \in I$ and $j \in J$, and on-path beliefs are independent.*

6.2 Correlated Stability and Stochastic Matching Functions

Modeling the firms' private observations and their private beliefs is a natural question. We have considered deterministic matching functions so far. Naturally, we are interested in stochastic matching functions. The two tasks can be accomplished together. This idea is an analog of the correlated equilibrium.

For each $j \in J$, let S_j be the finite set of *payoff-irrelevant* signals. We denote by $s = (s_{n+1}, ..., s_{n+m})$ the profile of signals of the m firms, and write $S = \times_{j \in J} S_j$. We do not need to introduce private signals for workers, because this amounts to a reinterpretation of workers' types $t = (t_1, ..., t_n)$. Assume that there is a common prior belief $\beta^0 \in \Delta(T \times S)$.[20] A matching (with private signals) is a function $M : (t, s) \mapsto (\mu, \mathbf{p})$. It is readily seen that the formulation proposed here includes a stochastic mapping as a special case where s is a public signal. In what follows, we will skip the plain-vanilla version of stability, and sketch the formulation of stability with Bayesian consistent beliefs.

Each firm j observes its own signal $s_j \in S_j$, but is uncertain about workers' types $t = (t_1, ..., t_n)$ and other firms' signals $s_{-j} = (s_{n+1}, ..., s_{j-1}, s_{j+1}, ..., s_{n+m})$. Similarly, each worker $i \in I$ observes its type t_i but is unaware of t_{-i} and s. Each firm $j \in J$, upon observing $s_j \in S_j$,

[20] Modeling heterogeneous priors is straightforward.

updates its belief to the conditional probability measure $\beta^0(\cdot|T \times \{s_j\} \times S_{-j}) \in \Delta(T \times S)$, which we shall denote simply by $\beta^0_{s_j}(\cdot)$. To further ease notation, we adopt the following harmless convention: for a non-empty subset $E \subset T \times S$, we write $\beta^0_{s_j}(t|E) := \beta^0_{s_j}(\{t\} \times S|E)$, and for a function $f : T \to \mathbb{R}$, we write $\mathbb{E}_{\beta^0_{s_j}}[f|E] := \sum_{t \in T} f(t)\beta^0_{s_j}(t|E)$.

Each firm $j \in J$, after observing its private signal s_j and the matching outcome $(\mu, \mathbf{p})$, holds a Bayesian consistent private on-path belief over workers' types $\beta^0_{s_j}(\cdot|M^{-1}(\mu, \mathbf{p})) \in \Delta(T)$. In a deviating coalition $(\mu, \mathbf{p}, i, j, p)$, firm j, which receives a private signal s_j, holds a Bayesian consistent private off-path belief $\beta^0_{s_j}(\cdot|M^{-1}(\mu, \mathbf{p}) \cap (D_{(\mu,\mathbf{p},i,j,p)} \times S))$, where $D_{(\mu,\mathbf{p},i,j,p)} = \left\{t' : a_{ij}(t') + p > a_{i\mu(i)}(t') + p_{i\mu(i)}\right\}$ is the set of types such that worker i benefits from the coalitional deviation.

With the on-path and off-path beliefs in place, notions of individual rationality, blocking, and stability of the matching $M : (t, s) \mapsto (\mu, \mathbf{p})$ can be defined in the same way as in Definitions 1, 2, and 3, respectively.

6.3 Incentive Compatibility

Although we have argued that stability is a reduced-form way of capturing the outcome of dynamic decentralized interactions, Bayesian incentive compatibility of a stable matching function $M : t \longmapsto (\mu, \mathbf{p})$ implies a one-shot implementation of a stable matching and serves as a desirable selection among stable matchings. However, in general, Bayesian incentive compatibility cannot be achieved.

Example 9. Consider a one-worker and one-firm problem. The worker privately knows the cost of his production (i.e., the negative of the worker's matching value), which takes the value of either 0 or 1 with equal prior probability. The firm's matching value is $L \in (0, 1)$ if the worker's cost is 0 and $H > 1$ if the worker's cost is 1. We assume that $\frac{1}{2}(L + H) < 1$; i.e., the firm's prior average matching value is less than the high cost. We claim that, in all stable matchings, the low-cost worker must be employed. This is true because, otherwise, the worker and the firm can block the matching with a salary of, say $\frac{1}{2}L$, whereby the low-cost worker is better off and the firm is better off regardless of its belief about the worker's type. Given that the low-cost worker must be matched, the high-cost worker cannot stay unmatched in a stable matching with Bayesian consistent beliefs; otherwise, the firm will assign probability 1 to the unmatched worker's cost being high, and the worker and the firm can block the matching with a salary of, say $\frac{1}{2}(1 + H) > 1$, whereby the high-cost worker is better off and the firm, knowing the worker's type, is also better off.

We have established the claim that, in a stable matching M that supported is by Bayesian consistent beliefs, both types of the worker must be hired. Given this, Bayesian incentive

compatibility of M requires that the salaries for both worker types be the same. However, the highest price the firm is willing to offer is $\frac{1}{2}(L+H)$, which the high-cost worker will reject. Thus, no stable matching can be Bayesian incentive compatible in this example. ■

The conflict between Bayesian incentive compatibility and stability is not surprising. The direct-revelation game associated with incentive compatibility is sometimes too restrictive for our purposes. For instance, we could allow for more general dynamic mechanisms, which is consistent with our motivation that stability is a reduced-form way of capturing the equilibrium outcome of decentralized interactions.[21] Indeed, Deneckere and Liang (2006, Proposition 2) cover this example and show that allocative efficiency is achieved in a sequential equilibrium of a firm-offer bargaining game with two separating prices and delayed trading for the high-cost type. This fully revealing outcome is stable by Proposition 3.

Characterizing the joint implications of Bayesian incentive compatibility and stability needs to remain an open question for now, but we do have a positive result under Assumption 1. In fact, dominant-strategy incentive compatibility can be obtained. The argument proceeds in two steps. First, under Assumption 1, the preferences of firms are independent of workers' private types, the matching function M that specifies a worker-optimal stable matching for each type profile t is dominant-strategy incentive compatible for workers. This claim follows from a result for complete information problems: when the firms' preference is fixed, the worker-optimal complete-information stable matching is strategyproof for the workers (e.g., Demange 1982 and Leonard 1983) and can be implemented by the VCG mechanism. Secondly, it follows from Proposition 1 that M so defined is incomplete-information stable.

Another special case where Bayesian incentive compatibility is easy to satisfy is "fully non-revealing" matching. If there exists an outcome $(\mu, \mathbf{p})$ such that it is complete-information stable matching for all $t \in T$, then $M \equiv (\mu, \mathbf{p})$ is stable by Proposition 3. The existence of such $(\mu, \mathbf{p})$ is not generally ensured and it depends on the value function (a, b).

7 Concluding Discussion

The main conceptual contribution of the paper is to propose a criterion of stability for two-sided markets with asymmetric information, with a formulation of Bayesian consistency of prior beliefs, on-path stable beliefs, and off-path stable beliefs. This criterion lays the foundation for further developments. It has immediate implications for empirical analysis of

[21] The literature of frictional search may offer useful insights in this direction; see, e.g., Lauermann (2013) and the references therein.

matching; see, e.g., Chiappori (2017). Although existing empirical work allows certain characteristics of players to be unobservable to the analysts, players themselves are assumed to have complete information, and hence the solution concept is complete-information stability.

We do not pretend that the results developed in this paper are immediately applicable to practical market design questions. However, providing a logically coherent Bayesian theory of stability is a necessary step toward understanding how players respond to information and incentives in both decentralized and centralized matching environments. The idea developed in this paper can easily be extended to markets with networked structures, more general coalitional games, or incomplete-information modeled by Harsanyi type spaces. The research agenda we propose here, which we can call the "Kreps–Wilson program," aims to develop cooperative concepts and their refinements under incomplete information using the insights from non-cooperative games.

A Appendix

A.1 Proof of Proposition 1

It follows from a similar construction as in Proposition 2 that if $M(t)$ is a complete-information stable matching when t is common knowledge, then M is stable. We now show the converse under Assumption 1. By the individual rationality of M, for any $t \in T$ with $M(t) = (\mu, \mathbf{p})$, we have

$$a_{i\mu(i)}(t) + p_{i\mu(i)} \geq 0 \text{ for all } i \in I \tag{A.1}$$

and $\mathbb{E}_{\beta^1_{(\mu,\mathbf{p},j)}}\left[b_{\mu(j)j}\right] - p_{\mu(j)j} \geq 0$ for all $j \in J$. By Assumption 1, $b_{\mu(j)j}(t)$ is independent of t, and hence $\mathbb{E}_{\beta^1_{(\mu,\mathbf{p},j)}}\left[b_{\mu(j)j}\right] = b_{\mu(j)j}(t)$. Thus,

$$b_{\mu(j)j}(t) - p_{\mu(j)j} \geq 0 \text{ for all } j \in J. \tag{A.2}$$

Hence, (A.1) and (A.2) imply that $(\mu, \mathbf{p})$ is individually rational when there is complete information about t.

Consider any coalitional deviation $c = (\mu, \mathbf{p}, i, j, p)$ to M at t. Since M is stable, c is not viable. If $a_{ij}(t) + p \leq a_{i\mu(i)}(t) + p_{i\mu(i)}$, then c is not viable even if t is common knowledge. If $a_{ij}(t) + p > a_{i\mu(i)}(t) + p_{i\mu(i)}$, then

$$\mathbb{E}_{\beta^2_c}\left[b_{ij}\right] - p \leq \max\left\{0, \mathbb{E}_{\beta^2_c}\left[b_{\mu(j)j}\right] - p_{\mu(j)j}\right\}. \tag{A.3}$$

By Assumption 1, $\mathbb{E}_{\beta_c^2}[b_{ij}] = b_{ij}(t)$ and $\mathbb{E}_{\beta_c^2}\left[b_{\mu(j)j}\right] = b_{\mu(j)j}(t)$. Inequality (A.3) can be rewritten as

$$b_{ij}(t) - p \leq \max\left\{0, b_{\mu(j)j}(t) - p_{\mu(j)j}\right\} = b_{\mu(j)j}(t) - p_{\mu(j)j},$$

where the last equality follows from (A.2). Therefore, c is not a viable coalitional deviation if there is complete information about t. We have thus proved that $M(t) = (\mu, \mathbf{p})$ is complete-information stable at $t \in T$. A stable matching under complete information maximizes the sum of surpluses, and hence the stable matching M is full-information efficient. ■

A.2 Proof of Proposition 2

Consider a matching game (a, b, β^0). If two types t_i and t_i' of worker i are indistinguishable, we write $t_i \sim t_i'$. We write $t \sim t'$ if $t_i \sim t_i'$ for each $i \in I$. For each $t \in T$, let $E(t) = \{t' : t' \sim t\}$ be the type profiles in the same equivalent class of t, and let $T^* = \{E(t) : t \in T\}$ be the collection of indistinguishable classes. For each $t \in T$, $i \in I$, and $j \in J$, define

$$a_{ij}^*(E(t)) = a_{ij}(t); \tag{A.4}$$

$$b_{ij}^*(E(t)) = \frac{1}{\beta^0(E(t))} \sum_{t' \in E(t)} b_{ij}(t')\beta^0(t'). \tag{A.5}$$

For each $t \in T$, pick any stable matching $(\mu, \mathbf{p})$ for the complete information matching game where the matching values are defined by $(a_{ij}^*(E(t)), b_{ij}^*(E(t)))_{i \in I, j \in J}$. If $t' \in E(t)$, we pick the same $(\mu, \mathbf{p})$ for t'. The existence of $(\mu, \mathbf{p})$ is ensured by Shapley and Shubik (1971) and Crawford and Knoer (1981). We claim that the matching function $M : t \mapsto (\mu, \mathbf{p})$ defined in this way is stable with Bayesian consistent beliefs.

Individual rationality. For each $t \in T$ and $(\mu, \mathbf{p}) = M(t)$,

$$a_{i\mu(i)}(t) + p_{i\mu(i)} = a_{i\mu(i)}^*(E(t)) + p_{i\mu(i)} \geq 0,$$

where the first equality follows from (A.4) and the inequality follows from the individual rationality of $(\mu, \mathbf{p})$. In addition, $M^{-1}(\mu, \mathbf{p})$ can be written as the union of disjoint equivalent classes $E_1, E_2, ..., E_k$. Therefore,

$$\mathbb{E}\left[b_{\mu(j)j} | M^{-1}(\mu, \mathbf{p})\right] = \frac{1}{\beta^0(M^{-1}(\mu, \mathbf{p}))} \sum_{t' \in M^{-1}(\mu, \mathbf{p})} b_{ij}(t')\beta^0(t') \tag{A.6}$$

$$= \frac{1}{\beta^0(\cup_{\ell=1}^k E_\ell)} \sum_{\ell=1}^k \beta^0(E_\ell) \left(\frac{1}{\beta^0(E_\ell)} \sum_{t' \in E_\ell} b_{ij}(t')\beta^0(t') \right) \tag{A.7}$$

$$= \frac{1}{\beta^0(\cup_{\ell=1}^k E_\ell)} \sum_{\ell=1}^k \beta^0(E_\ell) b_{ij}^*(E_\ell), \tag{A.8}$$

where the last equality follows from (A.5). Hence,

$$\begin{aligned}\mathbb{E}\left[b_{\mu(j)j}|M^{-1}(\mu,\mathbf{p})\right]-p_{\mu(j)j} &\geq \frac{1}{\beta^0(\cup_{\ell=1}^k E_\ell)}\sum_{\ell=1}^k \beta^0(E_\ell)\left(b^*_{\mu(j)j}(E_\ell)-p_{\mu(j)j}\right)\\ &\geq 0,\end{aligned}$$

where the last inequality follows from firm j's individual rationality in $(\mu,\mathbf{p})$.

No blocking. Consider a coalitional deviation $c=(\mu,\mathbf{p},i,j,p)$ at $t\in T$ such that $(\mu,\mathbf{p})=M(t)$. Suppose $a_{ij}(t)+p>a_{i\mu(i)}(t)+p_{i\mu(i)}$ (otherwise, the deviation is not viable). Let $D_c=\left\{t'\in T: a_{ij}(t')+p>a_{i\mu(i)}(t')+p_{i\mu(i)}\right\}$. If $t'\in D_c$, then $E(t')\subset D_c$. Therefore, $M^{-1}(\mu,\mathbf{p})\cap D_c$ can be written as a union of equivalent classes $F_1,...,F_h$.

Following the same arguments as in (A.6), (A.7), and (A.8), we have

$$\begin{aligned}\mathbb{E}\left[b_{ij}|M^{-1}(\mu,\mathbf{p})\cap D_c\right]-p &= \frac{1}{\beta^0(\cup_{\ell=1}^h F_\ell)}\sum_{\ell=1}^h \beta^0(F_\ell)\left(b^*_{ij}(F_\ell)-p\right) \text{ and}\\ \mathbb{E}\left[b_{\mu(j)j}|M^{-1}(\mu,\mathbf{p})\cap D_c\right]-p_{\mu(j)j} &= \frac{1}{\beta^0(\cup_{\ell=1}^h F_\ell)}\sum_{\ell=1}^h \beta^0(F_\ell)\left(b^*_{\mu(j)j}(F_\ell)-p_{\mu(j)j}\right).\end{aligned}$$

It follows from the complete-information stability of $(\mu,\mathbf{p})$ that

$$b^*_{ij}(F_\ell)-p\leq b^*_{\mu(j)j}(F_\ell)-p_{\mu(j)j},$$

where the right-hand side is positive by worker j's individual rationality. Hence,

$$\mathbb{E}\left[b_{ij}|M^{-1}(\mu,\mathbf{p})\cap D_c\right]-p\leq \mathbb{E}\left[b_{\mu(j)j}|M^{-1}(\mu,\mathbf{p})\cap D_c\right]-p_{\mu(j)j}.$$

This implies that firm j will not join the coalitional deviation c under impartial beliefs. Hence c is not a viable coalitional deviation for M. ∎

A.3 Proof of Proposition 3

To show the first claim, take any $t\in T$ and let $(\mu,\mathbf{p})=M(t)$. Since M is fully revealing, $\beta^1_{(\mu,\mathbf{p},j)}(t)=\beta^0(t|M^{-1}(\mu,\mathbf{p}))=1$, and hence the individual rationality of $(\mu,\mathbf{p})$ follows from the stability of M. Suppose to the contrary that $(\mu,\mathbf{p})$ is not stable when there is complete information about t. Then there exists $(i,j,p)\in I\times J\times\mathbb{R}$ such that $a_{ij}(t)+p>a_{i\mu(i)}(t)+p_{i\mu(i)}$ and $b_{ij}(t)-p>b_{\mu(j)j}(t)-p_{j\mu(j)}$. Consider $D_{(\mu,\mathbf{p},i,j,p)}=\{t': a_{ij}(t')+p>a_{i\mu(i)}(t')+p_{i\mu(i)}\}$. Since $\beta^1_{(\mu,\mathbf{p},j)}(t)=1$ and $t\in D_{(\mu,\mathbf{p},i,j,p)}$, it follows that $\beta^2_{(\mu,\mathbf{p},i,j,p)}(t)=1$. Thus $\mathbb{E}_{\beta^2_{(\mu,\mathbf{p},i,j,p)}}[b_{ij}(t)]-p>\mathbb{E}_{\beta^2_{(\mu,\mathbf{p},i,j,p)}}[b_{\mu(j)j}(t)]-p_{j\mu(j)}$. Therefore, the coalitional deviation $(\mu,\mathbf{p},i,j,p)$ for M is viable,

a contradiction. It should be noted that the power of Bayesian consistency of $\left(\beta^1, \beta^2\right)$ is not fully used in the argument; it is sufficient that the support of $\beta^2_{(\mu,\mathbf{p},i,j,p)}$ is restricted to $M^{-1}(\mu, \mathbf{p})$.

The proof of the second claim proceeds exactly the same way as the proof of Proposition 2 by work with an equivalence relation $\widetilde{E}(t) = \{t\}$. ∎

A.4 Proof of Propositions 4 and 5

A.4.1 Duality of Bayesian Efficiency

Consider a stable matching M, and any matching outcome $(\mu, \mathbf{p}) \in M(T)$. Since $\beta^1_{(\mu,\mathbf{p},j)} = \beta^0(t|M^{-1}(\mu, \mathbf{p}))$ is independent of j, we abuse the notation to write $\beta^1_{(\mu,\mathbf{p})} = \beta^0(\cdot|M^{-1}(\mu, \mathbf{p}))$. To show that M is Bayesian efficient, it is equivalent to show that μ maximizes

$$\sum_{i\in I}\left(\sum_{t\in T}\beta^0(t|M^{-1}(\mu,\mathbf{p}))a_{i\mu'(i)}(t) + \sum_{t\in T}\beta^0(t|M^{-1}(\mu,\mathbf{p}))b_{i\mu'(i)}(t)\right) \tag{A.9}$$

over all matches $\mu' : I \cup J \to I \cup J$, where $b_{ii} := 0$.

Primal. We introduce a vector of non-negative real variables $x = (x_{ij})_{i\in I, j\in J}$. Consider a problem that maximizes

$$V(x) := \sum_{i\in I}\sum_{j\in J} x_{ij}\left(\sum_{t\in T}\beta^0(t|M^{-1}(\mu,\mathbf{p}))(t)a_{ij}(t) + \sum_{t\in T}\beta^0(t|M^{-1}(\mu,\mathbf{p}))b_{ij}(t)\right)$$

subject to

$$\begin{aligned}&\textstyle\sum_{j\in J} x_{ij} \le 1;\\ &\textstyle\sum_{i\in I} x_{ij} \le 1;\\ &x_{ij} \ge 0,\ i \in I,\ j \in J.\end{aligned}$$

It is well known that this linear programming problem has an optimal solution x^* with all $x^*_{ij} = 0$ or 1. Such (x^*_{ij}) can be equivalently written as a match μ^*: $\mu^*(i) = j$ if and only if $x^*_{ij} = 1$, and the objective function of the linear program can be viewed as the sum of surpluses weighted by the probability measure $\beta^1_{(\mu,\mathbf{p})}$. Therefore, Bayesian efficiency of M is ensured if the match μ is an optimal solution to the linear programming problem.

Dual. The dual of this linear programming problem is to choose real variables $u = (u_i)_{i\in I}$ and $v = (v_j)_{j\in J}$ to minimize

$$U(u, v) := \sum_{i\in I} u_i + \sum_{j\in J} v_j$$

subject to

$$\begin{aligned} u_i + v_j &\geq \textstyle\sum_{t\in T} \beta^0(t|M^{-1}(\mu,\mathbf{p}))a_{ij}(t) + \sum_{t\in T} \beta^0(t|M^{-1}(\mu,\mathbf{p}))b_{ij}(t),\ i\in I,\ j\in J, \\ u_i &\geq 0,\ i\in I; \\ v_j &\geq 0,\ j\in J. \end{aligned} \tag{A.10}$$

Denote the optimal value of the dual by $U_{\min}$ and the optimal value of the primal by $V_{\max}$. By the strong duality theorem, $V_{\max} = U_{\min}$.

If there is complete information, the duality analysis is well known: the dual problem links the stable matching, and the strong duality theorem says that a stable matching is (full-information) efficient. With asymmetric information, the linkage of the dual to a stable matching is not immediate because the system of off-path beliefs β^2 is used to define stability whereas the system of the on-path beliefs β^1 appears in the dual problem. The impartial belief that links β^2 with β^1 through conditionality is critical here.

A.4.2 Proof of Propositions 4 and 5

Define $u^* = (u_1^*, ..., u_n^*)$, $v^* = (v_1^*, ..., v_m^*)$, and $x^* = (x_{ij}^*)_{i\in I, j\in J}$ as follows:

$$\begin{aligned} u_i^* &= \textstyle\sum_{t\in T} \beta^0(t|M^{-1}(\mu,\mathbf{p}))a_{i\mu(i)}(t) + p_{i\mu(i)}; \\ v_j^* &= \textstyle\sum_{t\in T} \beta^0(t|M^{-1}(\mu,\mathbf{p}))b_{\mu(j)j}(t) - p_{\mu(j)j}; \\ x_{ij}^* &= \begin{cases} 1 & \text{if } \mu(i) = j \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

By definition, x^* is feasible for the primal problem. We need to show that x^* is the optimal solution to the primal problem under certain conditions. We proceed in two steps.

Step 1. We shall establish the following claim: if (u^*, v^*) is a feasible solution to the dual problem, then x^* is an optimal solution to the primal problem, and consequently the match μ maximizes (A.9).

To prove this claim, note that

$$U(u^*, v^*) \geq U_{\min} = V_{\max} \geq V(x^*),$$

where the first relation follows from the assumption that (u^*, v^*) is a feasible solution to the dual problem, the second relation follows from the strong duality theorem, and the third relation follows because x^* is a feasible solution to the primal problem.

Note also that $V(x^*) = U(u^*, v^*)$ because each of them is the total expected payoff from

$(\mu, \mathbf{p})$ with belief $\beta^1_{(\mu,\mathbf{p})}$. Therefore,

$$U(u^*, v^*) = U_{\min} = V_{\max} = V(x^*).$$

This proves that x^* is an optimal solution to the primal problem.

Step 2. We shall show that (u^*, v^*) is a feasible solution to the dual problem, if the conditions in Propositions 4 and 5 are satisfied.

By definition, (u^*, v^*) is non-negative. It remains to show that (u^*, v^*) satisfies the constraint (A.10) in the dual problem. We claim that for any t in the support of $\beta^0(\cdot|M^{-1}(\mu, \mathbf{p}))$, and any $i \in I$ and $j \in J$,

$$\begin{aligned} & a_{i\mu(i)}(t) + p_{i\mu(i)} + \textstyle\sum_{t\in T} \beta^0(t|M^{-1}(\mu, \mathbf{p}))b_{\mu(j)j}(t) - p_{\mu(j)j} \qquad (A.11) \\ \geq\ & a_{ij}(t) + \textstyle\sum_{t\in T} \beta^0(t|M^{-1}(\mu, \mathbf{p}))b_{ij}(t). \end{aligned}$$

The claim is trivially true if $\mu(i) = j$. To prove this claim, suppose by way of contradiction that (A.11) does not hold for some $\bar{t}$ in the support of $\beta^0(\cdot|M^{-1}(\mu, \mathbf{p}))$ and some pair $(i, j) \in I \times J$, $\mu(i) \neq j$. Then, there exists $p \in \mathbb{R}$ such that

$$a_{i\mu(i)}(\bar{t}) + p_{i\mu(i)} < a_{ij}(\bar{t}) + p, \text{ and} \qquad (A.12)$$

$$\textstyle\sum_{t\in T} \beta^0(t|M^{-1}(\mu, \mathbf{p}))b_{\mu(j)j}(t) - p_{\mu(j)j} < \sum_{t\in T} \beta^0(t|M^{-1}(\mu, \mathbf{p}))b_{ij}(t) - p. \qquad (A.13)$$

Inequality (A.12) captures worker i's incentive to form a coalitional deviation with firm j. Consider the coalitional deviation $c = (\mu, \mathbf{p}, i, j, p)$, and the set

$$D_c := \left\{t \in T : a_{i\mu(i)}(t) + p_{i\mu(i)} < a_{ij}(t) + p\right\}.$$

By (A.12), D_c is a non-empty set that contains $\bar{t}$. Firms' common off-path belief is given by $\beta^0(\cdot|M^{-1}(\mu, \mathbf{p}) \cap D_c)$.

Under condition (i) of Proposition 4, i.e, Assumption 2, $a_{i\mu(i)}(t)$ and $a_{ij}(t)$ are independent of t. Therefore, $D_c = \left\{t \in T : p_{i\mu(i)} < p\right\}$ if $\mu(i) \in J$, and $D_c = \left\{t \in T : p_{i\mu(i)} < h(i, j) + p\right\}$ if $\mu(i) = i$, where $h(i, j) = a_{ij}(t)$ and $a_{ii}(t) = 0$. In either case, $D_c = T$ since $\bar{t} \in D_c$.

Under condition (ii) of Proposition 4, $\mu(i) \neq i$, and by Assumption 3, $a_{i\mu(i)}(t) = a_{ij}(t) = g(i, t) + h(i, j)$. Hence,

$$D_c = \left\{t \in T : h(i, \mu(i)) + p_{i\mu(i)} < h(i, j) + p\right\}.$$

Again, $D_c = T$ since $\bar{t} \in D_c$.

Under either condition (i) or condition (ii) of Proposition 4, $\beta^0(D_c|M^{-1}(\mu,\mathbf{p}))=1$. If we replace the on-path belief $\beta^0(t|M^{-1}(\mu,\mathbf{p}))$ by the off-path belief $\beta^0(t|M^{-1}(\mu,\mathbf{p})\cap D_c)$ in (A.13), the inequality is unchanged. Therefore, (A.13) implies that firm j is willing to deviate with worker i. That is, $(\mu,\mathbf{p},i,j,p)$ is a successful blocking, a contradiction.

Suppose that the conditions of Proposition 5 hold and that $\mu(i)=i$ (the case of $\mu(i)\neq i$ has already been covered by the proof of Proposition 4 under condition (ii)). Then

$$D_c=\left\{t\in T: p_{i\mu(i)}<a_{ij}(t)+p\right\}.$$

Since $a_{ij}(\cdot)$ and $b_{ij}(\cdot)$ are co-monotonic, there exists some linear order on T_i that is specific to the pair (i,j), such that both $a_{ij}(t_i)$ and $b_{ij}(t_i)$ are non-decreasing in t_i (note that since a_{ij} and b_{ij} depend only on t_i, the linear order naturally extends to an order on T). Therefore, D_c contains all t's such that t_i is larger than a cutoff according to the linear order. It follows from the monotonicity of $b_{ij}(t)$ in t_i that

$$\textstyle\sum_{t\in T}\beta^0(t|M^{-1}(\mu,\mathbf{p}))b_{ij}(t)-p\leq\sum_{t\in T}\beta^0(t|M^{-1}(\mu,\mathbf{p})\cap D_c)b_{ij}(t)-p. \tag{A.14}$$

Since $\mu(i)=i\neq j$, we have $\mu(j)\neq i$. It follows from the independence of $\beta^0(\cdot|M^{-1}(\mu,\mathbf{p}))$ that

$$\beta^0(\{t_{\mu(j)}\}\times T_{-\mu(j)}|M^{-1}(\mu,\mathbf{p}))=\beta^0(\{t_{\mu(j)}\}\times T_{-\mu(j)}|M^{-1}(\mu,\mathbf{p})\cap D_c).$$

Hence

$$\textstyle\sum_{t\in T}\beta^0(t|M^{-1}(\mu,\mathbf{p}))b_{\mu(j)j}(t)-p_{\mu(j)j}=\sum_{t\in T}\beta^0(t|M^{-1}(\mu,\mathbf{p})\cap D_c)b_{\mu(j)j}(t)-p_{\mu(j)j}. \tag{A.15}$$

It follows from (A.15) and (A.13) that

$$\textstyle\sum_{t\in T}\beta^0(t|M^{-1}(\mu,\mathbf{p})\cap D_c)b_{\mu(j)j}(t)-p_{\mu(j)j}<\sum_{t\in T}\beta^0(t|M^{-1}(\mu,\mathbf{p}))b_{ij}(t)-p. \tag{A.16}$$

By (A.14) and (A.16),

$$\textstyle\sum_{t\in T}\beta^0(t|M^{-1}(\mu,\mathbf{p})\cap D_c)b_{\mu(j)j}(t)-p_{\mu(j)j}<\sum_{t\in T}\beta^0(t|M^{-1}(\mu,\mathbf{p})\cap D_c)b_{ij}(t)-p.$$

That is, firm j is willing to deviate with worker i. Thus, the coalitional deviation $(\mu,\mathbf{p},i,j,p)$ is not viable, a contradiction. This establishes the claim that (A.11) holds.

Multiplying both sides of (A.11) by $\beta^0(t|M^{-1}(\mu,\mathbf{p}))$ and summing over t, we obtain

$$u_i^*+v_j^*\geq\textstyle\sum_{t\in T}\beta^0(t|M^{-1}(\mu,\mathbf{p}))a_{ij}(t)+\sum_{t\in T}\beta^0(t|M^{-1}(\mu,\mathbf{p}))b_{ij}(t).$$

That is, (u^*, v^*) satisfies (A.10). Thus, (u^*, v^*) is a feasible solution to the dual problem. ■

A.5 Proof of Proposition 6

For each $t \in T$, pick any complete-information competitive equilibrium matching $(\mu, \mathbf{p})$ associated with the complete-information matching game $(a^*_{ij}(E(t)), b^*_{ij}(E(t)))_{i\in I, j\in J}$ as in (A.4) and (A.5). We claim that the matching $M : t \mapsto (\mu, \mathbf{p})$ is a (rational expectations) competitive equilibrium. Since $(\mu, \mathbf{p})$ is a competitive equilibrium of the complete-information matching game $(a^*_{ij}(E(t)), b^*_{ij}(E(t)))_{i\in I, j\in J}$,

$$a^*_{i\mu(i)}(E(t)) + p_{i\mu(i)} \geq a^*_{ij}(E(t)) + p_{ij}.$$

Thus, by (A.4),

$$a_{i\mu(i)}(t) + p_{i\mu(i)} \geq a_{ij}(t) + p_{ij}.$$

That is, condition (i) of Definition 8 is satisfied. In addition, $M^{-1}(\mu, \mathbf{p})$ can be written as the union of equivalent classes $G_1, ..., G_c$. Following the argument as in (A.6)–(A.8), for all $j \in J$ and $i \in I \cup \{j\}$, we have

$$\begin{aligned}
\mathbb{E}\left[b_{\mu(j)j} | M^{-1}(\mu, \mathbf{p})\right] - p_{\mu(j)j} &= \frac{1}{\beta^0(\cup_{\ell=1}^k G_\ell)} \sum_{\ell=1}^k \beta^0(G_\ell) \left(b^*_{\mu(j)j}(G_\ell) - p_{\mu(j)j}\right); \\
\mathbb{E}\left[b_{ij} | M^{-1}(\mu, \mathbf{p})\right] - p_{ij} &= \frac{1}{\beta^0(\cup_{\ell=1}^k G_\ell)} \sum_{\ell=1}^k \beta^0(G_\ell) \left(b^*_{ij}(G_\ell) - p_{ij}\right).
\end{aligned}$$

By the definition of $(\mu, \mathbf{p})$, $b^*_{\mu(j)j}(G_\ell) - p_{\mu(j)j} \geq b^*_{ij}(G_\ell) - p_{ij}$. Therefore,

$$\mathbb{E}\left[b_{\mu(j)j} | M^{-1}(\mu, \mathbf{p})\right] - p_{\mu(j)j} \geq \mathbb{E}\left[b_{ij} | M^{-1}(\mu, \mathbf{p})\right] - p_{ij}.$$

That is, condition (ii) in Definition 8 is satisfied. ■

A.6 Proof of Proposition 7

By the proofs of Propositions 2 and 6, $M^s : t \mapsto (\mu, \mathbf{p}^s)$ and $M^c : t \mapsto (\mu, \mathbf{p}^c)$ are stable and competitive equilibrium respectively, when $(\mu, \mathbf{p}^s)$ is a complete-information stable matching at t and $(\mu, \mathbf{p}^c)$ is a complete-information competitive equilibrium extension of $(\mu, \mathbf{p}^s)$. This proves the first halves of (i) and (ii). The second halves are shown by Examples 6 and 7, respectively. ■

A.7 Proof of Proposition 8

Suppose to the contrary that a competitive equilibrium matching M is not Bayesian efficient. Then for some $(\mu, \mathbf{p}) \in M(T)$ there exists a match $\mu' : I \cup J \to I \cup J$ such that

$$\mathbb{E}\left[\sum_{i=1}^{n}(a_{i\mu(i)} + b_{i\mu(i)})|M^{-1}(\mu, \mathbf{p})\right] < \mathbb{E}\left[\sum_{i=1}^{n}(a_{i\mu'(i)} + b_{i\mu'(i)})|M^{-1}(\mu, \mathbf{p})\right]. \tag{A.17}$$

Since M is a competitive equilibrium,

$$\mathbb{E}\left[a_{i\mu(i)} + p_{i\mu(i)}|M^{-1}(\mu, \mathbf{p})\right] \geq \mathbb{E}\left[a_{i\mu'(i)} + p_{i\mu'(i)}|M^{-1}(\mu, \mathbf{p})\right] \tag{A.18}$$

for all $i \in I$, and

$$\mathbb{E}\left[b_{\mu(j)j} - p_{\mu(j)j}|M^{-1}(\mu, \mathbf{p})\right] \geq \mathbb{E}\left[b_{\mu'(j)j} - p_{\mu'(j)j}|M^{-1}(\mu, \mathbf{p})\right] \tag{A.19}$$

for all $j \in J$. Observe that, since $p_{ii} = p_{jj} = 0$,

$$\sum_{i=1}^{n} p_{i\mu(i)} = \sum_{j=1}^{m} p_{\mu(j)j} \text{ and } \sum_{i=1}^{n} p_{i\mu'(i)} = \sum_{j=1}^{m} p_{\mu'(j)j}.$$

Hence, summing (A.18) over $i \in I$ and (A.19) over $j \in J$, we have

$$\mathbb{E}\left[\sum_{i=1}^{n} a_{i\mu(i)} + \sum_{j=1}^{m} b_{\mu(j)j}|M^{-1}(\mu, \mathbf{p})\right] \geq \mathbb{E}\left[\sum_{i=1}^{n} a_{i\mu'(i)} + \sum_{j=1}^{m} b_{\mu'(j)j}|M^{-1}(\mu, \mathbf{p})\right],$$

which, since $a_{ii} \equiv 0 \equiv b_{jj}$, is equivalent to

$$\mathbb{E}_{\beta^1_{(\mu,\mathbf{p})}}\left[\sum_{i=1}^{n}\left(a_{i\mu(i)} + b_{i\mu(i)}\right)\right] \geq \mathbb{E}_{\beta^1_{(\mu,\mathbf{p})}}\left[\sum_{i=1}^{n}\left(a_{i\mu'(i)} + b_{i\mu'(i)}\right)\right]. \tag{A.20}$$

But (A.20) and (A.17) contradict each other.

We now prove the claim about full-information efficiency. It follows from the construction of Proposition 6 that if $M(t)$ is a competitive equilibrium matching when t is common knowledge, then M is a competitive equilibrium. Suppose that M is a competitive equilibrium. Then, by definition, for all $t \in T$ and $(\mu, \mathbf{p}) = M(t)$, $a_{i\mu(i)}(t) + p_{i\mu(i)} \geq a_{ij}(t) + p_{ij}$ for all $i \in I$ and $j \in J \cup \{i\}$, and $\mathbb{E}[b_{\mu(j)j}|M^{-1}(\mu, \mathbf{p})] - p_{\mu(j)j} \geq \mathbb{E}[b_{ij}|M^{-1}(\mu, \mathbf{p})] - p_{ij}$ for all $j \in J$ and $i \in I \cup \{j\}$. By Assumption 1, the last inequality is equivalent to $b_{\mu(j)j} - p_{\mu(j)j} \geq b_{ij} - p_{ij}$ for all $j \in J$ and $i \in I \cup \{j\}$. Thus, $(\mu, \mathbf{p})$ is a complete-information competitive equilibrium (and maximizes the sum of surpluses) when the type profile is t. Hence M is full-information efficient. ∎

References

[1] Akerlof, George. 1970. "The market for lemons." *Quarterly Journal of Economics* 84(3): 488–500.

[2] Alston, Max. 2020. "On the non-existence of stable matches with incomplete information." *Games and Economic Behavior* 120: 336–344.

[3] Ambrus, Attila. 2006. "Coalitional rationalizability." *The Quarterly Journal of Economics* 121(3): 903–926.

[4] Ashlagi, Itai, Yash Kanoria, and Jacob D. Leshno. 2017. "Unbalanced random matching markets: The stark effect of competition." *Journal of Political Economy* 125(1): 69–98.

[5] Becker, Gary S. 1973. "A theory of marriage: Part I." *Journal of Political Economy* 81(4): 813–846.

[6] Bernheim, B. Douglas. 1984. "Rationalizable strategic behavior."*Econometrica* 52(4): 1007–1028.

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

[8] Chakraborty, Archishman, Alessandro Citanna, and Michael Ostrovsky. 2010. "Two-sided matching with interdependent values." *Journal of Economic Theory* 145(1): 85–105.

[9] Chen, Yi-Chun and Gaoji Hu. 2020. "Learning by matching." *Theoretical Economics* 15: 29–56.

[10] Chiappori, Pierre-André. 2017. *Matching with Transfers: The Economics of Love and Marriage.* Princeton University Press.

[11] Cho, In-Koo and David M. Kreps. 1987. "Signaling games and stable equilibria." *The Quarterly Journal of Economics* 102(2): 179–221.

[12] Crawford, Vincent P. and Elsie Marie Knoer. 1981. "Job matching with heterogeneous firms and workers." *Econometrica* 49(2): 437–450.

[13] Dekel, Eddie, Drew Fudenberg, and David K. Levine. 2004. "Learning to play Bayesian games." *Games and Economic Behavior* 46(2): 282-303.

[14] Demange, Gabrielle. 1982. "Strategyproofness in the assignment market game." Working Paper, Labratorie d'Econometrie de l'Ecole Polytechnique, Paris.

[15] Deneckere, Raymond and Meng-Yu Liang. 2006. "Bargaining with interdependent values." *Econometrica* 74(5): 1309–1364.

[16] Dutta, Bhaskar and Rajiv Vohra. 2005. "Incomplete information, credibility and the core." *Mathematical Social Sciences* 50(2): 148–165.

[17] Forges, Françoise. 1994. "Posterior efficiency." *Games and Economic Behavior* 6(2): 238–261.

[18] Forges, Françoise, Enrico Minelli, and Rajiv Vohra. 2002. "Incentives and the core of an exchange economy: a survey." *Journal of Mathematical Economics* 38(1): 1–41.

[19] Fudenberg, Drew and Jean Tirole. 1991. "Perfect Bayesian equilibrium and sequential equilibrium." *Journal of Economic Theory* 53(2): 236–260.

[20] Gale, David and Lloyd S. Shapley. 1962. "College admissions and the stability of marriage." *The American Mathematical Monthly* 69(1): 9–15.

[21] Green, Jerry R. and Jean-Jacques Laffont. 1987. "Posterior implementability in a two-person decision problem. "*Econometrica* 55(1): 69–94.

[22] Grossman, Sanford J. and Motty Perry. 1986. "Perfect sequential equilibrium." *Journal of Economic Theory* 39(1): 97–119.

[23] Gul, Faruk. 1989. "Bargaining foundations of Shapley value." *Econometrica* 57(1): 81–95.

[24] Holmström, Bengt and Roger Myerson. 1993. "Efficient and durable decision rules with incomplete information." *Econometrica* 51(6): 1799–1819.

[25] Jordan, James S. 1983. "On the efficient markets hypothesis." *Econometrica* 51(5): 1325–43.

[26] Koopmans, Tjalling C. and Martin Beckmann. 1957. "Assignment problems and the location of economic activities." *Econometrica* 25(1): 53–76.

[27] Kreps, David M. 1977. "A note on 'fulfilled expectations' equilibria." *Journal of Economic Theory* 14(1): 32–43.

[28] Kreps, David M. and Robert Wilson. 1982. "Sequential equilibria." *Econometrica* 50(4): 863–894.

[29] Lauermann, Stephan. 2013. "Dynamic matching and bargaining games: A general approach."*American Economic Review* 103(2): 663–689.

[30] Liu, Qingmin, George J. Mailath, Andrew Postlewaite, and Larry Samuelson. 2014. "Stable matching with incomplete information." *Econometrica* 82(2): 541–587.

[31] Leonard, Herman B. 1983. "Elicitation of honest preferences for the assignment of individuals to positions." *Journal of Political Economy* 91(3): 461–479.

[32] Mauleon, Ana, Vincent J. Vannetelbosch, and Wouter Vergote. 2011. "Von Neumann–Morgenstern farsightedly stable sets in two-sided matching." *Theoretical Economics* 6(3): 499–521.

[33] Milgrom, Paul and Nancy Stokey. 1982. "Information, trade and common knowledge." *Journal of Economic Theory* 26(1): 17–27.

[34] Pearce, David G. 1982. "Rationalizable strategic behavior and the problem of perfection." *Econometrica* 52(4): 1029–1050.

[35] Perry, Motty and Philip J. Reny. 1994. "A noncooperative view of coalition formation and the core." *Econometrica* 62(4): 795–817.

[36] Pomatto, Luciano. 2015. "Stable matching under forward-induction reasoning." *Working Paper*, MEDS, Kellogg School of Management, Northwestern University.

[37] Radner, Roy. 1979. "Rational expectations equilibrium: Generic existence and the information revealed by prices." *Econometrica* 47(3): 655–678.

[38] Ray, Debraj and Rajiv Vohra. 2015. "The Farsighted Stable Set." *Econometrica* 83(3): 977–1011.

[39] Roth, Alvin E. 1989. "Two-sided matching with incomplete information about others' preferences." *Games and Economic Behavior* 1(2): 191–209.

[40] Rothschild, Michael and Joseph Stiglitz. 1976. "Equilibrium in competitive insurance markets: An essay on the economics of imperfect information." *The Quarterly Journal of Economics* 90(4): 629–649.

[41] Rubinstein, Ariel and Asher Wolinsky. 1994. "Rationalizable conjectural equilibrium: Between Nash and rationalizability." *Games and Economic Behavior*, 6(2): 299–311.

[42] Salanié, Bernard. 2015. "Identification in Separable Matching with Observed Transfers." Working Paper, Columbia University.

[43] Shapley, Lloyd S. and Martin Shubik. 1971. "The assignment game I: The core." *International Journal of Game Theory* 1(1): 111–130.

[44] Spence, Michael. 1973. "Job market signaling." *The Quarterly Journal of Economics* 87(3): 355–374.

[45] Yenmez, M. Bumin. 2013. "Incentive-compatible matching mechanisms: Consistency with various stability notions." *American Economic Journal: Microeconomics* 5(4): 120–141.

[46] Wilson, Robert B. 1967. "Competitive bidding with asymmetric information." *Management Science* 13(11): 816–820.

[47] Wilson, Robert B. 1978. "Information, efficiency, and the core of an economy." *Econometrica* 46(4): 807–816.