Rational Expectations, Stable Beliefs, and Stable Matching[*]

Qingmin Liu[†]

August 5, 2018

Abstract

We propose a new criterion for stability of two-sided matching markets with asymmetric information. The criterion requires the Bayesian consistency of three probabilistic beliefs: exogenously given prior beliefs, off-path beliefs conditional on counterfactual pairwise blockings, and on-path beliefs for stable matchings in the absence of such blockings. The formulation provides a language for assessing matching outcomes with respect to their supporting beliefs and enables further belief-based refinements.

We also define criteria of match efficiency, rational expectations competitive equilibrium, and the core. Their contrast with pairwise stability manifests the role of information asymmetry in matching formation.

Contents

1   Introduction

This paper develops a Bayesian criterion of stability for two-sided matching markets with asymmetric information. Specifically, we study a market with transfers, 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 idea of asymmetric information in two-sided markets is longstanding and revolutionary (Akerlof 1970, Spence 1973, Rothschild and Stiglitz 1976 etc.), but the main analytical tools are competitive equilibrium and non-cooperative game theory, which are substantially different from stability.[1]

The present paper aims 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 solution concepts for *dynamic* non-cooperative games of asymmetric information elucidates the 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 outcomes in 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 the beliefs—both on and off the equilibrium path—are endogenously determined in equilibrium. This powerful insight of separating equilibrium beliefs from equilibrium strategies, so natural as to be taken for granted today, is instrumental in wide-ranging applications of dynamic games and it paves the way for subsequent development of belief-based equilibrium refinements. We may expect

---

[1]Two features distinguish equilibrium theories from stability. First, equilibrium theories are developed on the premise of individual optimization while holding fixed the behavior of all other actors. In many small 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. Second, 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 ad hoc assumptions on the non-cooperative game forms might seem reasonable in one context but may become unrealistic in other applications. The advantage of the cooperative concept of stability is that it focuses on payoff assumptions and abstracts away from details of strategic interactions.

[2]See, e.g., Nash (1953), Gul (1989), Perry and Reny (1994) for the exposition of this idea commonly known as "the Nash program." Crawford and Knoer (1981) and Blum, Roth, and Rothblum (1997) offer related discussions for two-sided matching markets. In fact, it is a common view that all solution concepts are shortcuts to capture some dynamic interactions, including the fundamental non-cooperative concept of Nash equilibrium (as argued by Nash himself and others).

an analog in a cooperative theory of stability with asymmetric information; otherwise there would be no hope for a new concept to capture relevant behaviors in 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 (cf. beliefs on the equilibrium path that are consistent with equilibrium strategies in sequential equilibrium) and a formulation of "off-path stable beliefs" upon "an off-path blocking" that deter the blocking could naturally enable the discussion of belief-based refinements (cf. off-equilibrium beliefs upon a unilateral deviation that deter the deviation). Both on-path and off-path beliefs should be endogenous. This idea seems obvious and natural, but, surprisingly, it has not been formally examined in matching problems in particular and cooperative games in general.

The necessity of endogenous on-path and off-path beliefs for a cooperative concept of stability should after all stem from the cooperative perspective of matching problems. In a complete-information model, two players from opposite sides of the market block a matching if both can strictly improve their payoffs by a rematch; a stable matching is a situation in which no such blocking opportunity exists. With asymmetric information, an uninformed player's uncertainty at each contingency should be described by a probabilistic belief. In a counterfactual mutually profitable blocking of a putative matching, an uninformed player's belief should be consistent with the assumption that the blocking is profitable for its informed opponent. This is an off-path event relative to the putative matching and the associated belief is an off-path belief. The putative matching being stable means that there is no mutually profitable blocking opportunity, and hence such a matching should result in an on-path belief. In short, if a blocking being mutually profitable reveals something and leads to belief updating, the non-existence of such a blocking should reveal something else and lead to belief updating as well.

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 for "off-path" events. In defining stability, we simply test a putative matching against all counterfactual pairwise blockings, as is consistent with Gale and Shapley (1962) and Shapley and Shubik (1971). If we agree that endogenous on-path and off-path stable beliefs are qualitatively different from prior beliefs that are exogenously specified as primitives of our models, the immediate conceptual question is how to define stable beliefs and stability for a matching game (modeled by a payoff function and a prior belief). 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 as an outcome of updating the prior belief from a sequence of unsuccessful blockings, because, at

the very least, belief updating within each and every blocking coalition must be determined concurrently. Here is how we resolve this conceptual question. We make use of an idea of "outcome functions" similar to that in the literature on rational expectations equilibrium: players understand the equilibrium relationship between the underlying uncertainties (in our case players' types) and the observables (in our case matches and prices). Using Bayes' rule, on-path stable beliefs are "derived" from priors conditional on the observables, and off-path stable beliefs are further "derived" from on-path stable beliefs conditional on the event of off-path blocking. Only when the Bayesian consistency is in place do we impose conditions that the putative matching is individually rational (where players' payoffs from the matching are computed using on-path beliefs) and it is immune to pairwise blocking (where the payoff from a counterfactual blocking is computed using the off-path belief). This closes the loop of defining (on-path and off-path) stable beliefs and 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.

Given the wide range of potential applications, attempts to relax the restriction of complete information in matching models are nothing new. Roth (1989) and Chakraborty, Citanna, and Ostrovsky (2010) study implementation of matching outcomes without transfers using non-cooperative equilibrium concepts. Their approach to incomplete information is natural, although the choice of game forms matters for the equilibrium outcomes.[3] Dutta and Vohra (2005) and Yenmez (2013) study interim notions of blocking. These papers belong to a long-standing literature on the asymmetric information core initiated by Wilson (1978). Section 1.2 offers more details; the essence is that this literature only formulates beliefs within deviating coalitions (or off-path beliefs), and it often takes the exogenously given prior beliefs (or exogenous interim beliefs conditional on types) as on-path beliefs which are endogenously determined according to our argument.

Recognizing the intricacy of belief formation in matching problems, Liu, Mailath, Postlewaite, and Samuelson (2014) take a different, belief-free approach. They define stable matching as a fixed-point set that survives iterative elimination of matching outcomes that are blocked, a procedure similar to rationalizability. The fixed-point set indirectly captures inferences made from the non-existence of pairwise blocking. Chen and Hu (2017) generalize this stability notion further and construct a converging adaptive process to stability,

---

[3]Roth (1989) studies implementation in Bayesian Nash equilibria and dominant strategies for direct revelation games. His stability notion is the complete information one. Chakraborty, Citanna, and Ostrovsky (2010) consider an extension of the revelation game: colleges report their private information to a designer; based on their private observations about the designer's recommended match, all colleges simultaneously decide whether to make one round of offers to which students must respond. The authors define stability as the perfect Bayesian equilibrium of this extensive-form game wherein players follow the designer's recommendation.

settling a foundational question left open by the early paper; Pomatto (2015) provides an epistemic foundation by using a forward induction logic; Alston (2017) demonstrates that extreme pessimism is the only belief configuration that ensures the generic non-emptiness of the fixed-point set. Section 1.2.2 explains in detail the belief-free approach; the essence is that this approach, although offering strong restrictions in specific contexts, evades the central insight of decision-theoretic and game-theoretic analysis of uncertainty: the Bayesian formulation of posterior and prior beliefs across players.

With complete information, transfers and hence 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 match 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 the supporting stable beliefs. We give tight conditions on the primitives of matching games under which all stable matchings must be Bayesian efficient, and these conditions include familiar models of adverse selection as special cases.

Asymmetric information is qualitatively different from complete information in ways that go beyond match efficiency. Motivated by two concepts that have occupied significant places in economic theory, we define the core and competitive equilibrium for matching markets with asymmetric information. The findings highlight the role of informational friction in matching formation. With 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 asymmetric information, the set of stable matchings contains the core as a subset. A matching that is not blocked by any single pair of players may be blocked by multiple pairs together, even when (prior and posterior) type distributions across players are independent. Consider a scenario of two blocking pairs. In deciding whether to join a blocking, a player compares his on-path match and off-path match (who are another player's off-path match and on-path match, respectively). With one blocking pair, a player learns only from the incentive of his off-path match. In a blocking that involves two pairs jointly, a player also learns from the fact that his on-path match is willing to form a blocking pair with another player (who does a similar calculation for the two matches in agreeing to the blocking, thus closing the cycle of inference). Critically, the formation of each blocking pair is necessarily conditional on the formation of the other blocking pair (otherwise, pairwise stability is the same as the core).

We propose a notion of rational expectations competitive equilibrium that extends the notion of complete-information competitive equilibrium of Koopmans and Beckmann (1957), Shapley and Shubik (1971), and Becker (1973). Unlike stability, a competitive equilibrium specifies a price for any two players and postulates that a player's unilateral deviation is deemed valid even though the other player is unwilling to be matched with the deviating player. However, the two conceptually different notions are outcome-equivalent under complete information. Informational asymmetry makes a difference and the equivalence breaks down. The set of competitive matchings and the set of stable matchings must overlap but in general neither one contains the other. The reason for this is precisely that the two theories of market mechanisms, stability and competitive equilibrium, make different assumptions about the role of information. Stability says that a buyer's deviation must be approved by a seller who sells to the buyer (i.e., a match in our language); but a rational expectations equilibrium excludes information revelation through a mutually agreed deviation, and this information can either facilitate or impede blocking. But a competitive equilibrium is always Bayesian efficient, precisely because of the nature of unilateral optimization. In a small two-sided market, the concept of pairwise stability looks more appealing.

The rest of the paper is organized as follows. Section 1.1 demonstrates several features of consistent beliefs using two examples. Section 1.2 explains why the ideas of the present paper are not studied in related literature. Sections 2 and 3 contain the definitions of stable belief and stable matching. Section 4 studies efficiency of stable matchings. Section 5 compares stability and rational expectations equilibrium. Section 6 studies the core and the stability notion with heterogeneous private beliefs. Section 7 concludes.

1.1   Examples

We offer two examples to illustrate the criterion of stability. The first one shows that beliefs associated with stability are qualitatively different from prior beliefs, and they cannot be determined a priori independently of the stability of a matching. The second one demonstrates the restrictions of 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	$-1, 2$	$-3, 5$
t_1'	$0, 2$	$-4, 5$

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, note that the salary p 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 from matching with firm 2 is at most $-4 + p \leq 1$. But the worker of type t'_1 could block it 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 it 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, type t_1 cannot be hired by firm 1 and type t'_1 cannot be hired by firm 2: firms' 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 types irrespective of the prior distribution.

The takeaway of the example is that beliefs in a stable matching that is immune to pairwise blocking should not simply be taken as prior beliefs and they cannot be determined a priori. The on-path belief must be consistently determined together with the stability of a 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 will form a blocking pair with the worker. Therefore, it is not a stable matching for firm 1 to hire the worker.

6

The argument above is intuitive, and a natural notion of stability may not predict otherwise. But the example is conceively simple. We shall raise two issues that may have broader implications, which should not be surprising for students of non-cooperative game theory. First, if the putative matching of firm 1 and the worker is under consideration, the reference to a new pair formed by firm 2 (or firm 3) and the worker is about an off-path event. The argument above assumes that the off-path beliefs of the two firms are identically q which is the same as the on-path belief. This is appealing in the context of this example. There does not seem to be a compelling reason for the 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 any firm for a higher wage. But the refinement through off-path beliefs is an interesting question in general. The takeaway of this example is that the specification of off-path beliefs should not be too arbitrary, and it should be based on our intuition about the game, as is already well known from the equilibrium refinement literature.

Secondly, if firm 2 and firm 3 are allowed to have heterogeneous posterior beliefs, in particular, say firm 2 thinks the worker's type is t_1 and firm 3 thinks the worker's type is t_1', then it would be stable for the worker to match with firm 1. This will be the case if we assume that each firm looks at its respective worst-case scenario. However, these heterogeneous posterior beliefs are questionable given that the two firms' priors and observations are identical. This is an implication of Bayesian consistency of posteriors (both on path and off path) with a common prior.

1.2   Related Literature

1.2.1   Relation to the Asymmetric-Information Core

Although related work on matching is sparse, there is a large literature on the core in incomplete information environments initiated by Wilson (1978). Wilson defines two concepts for the incomplete-information economy, "coarse core" and "fine core," corresponding to two exogenously given protocols of information-based trading within a blocking coalition in the interim stage. The literature after Wilson (1978) focuses on endogenous information aggregation within a blocking coalition, either by taking a mechanism design approach (a one-shot game) to blocking or by incorporating some other non-cooperative elements into the otherwise cooperative framework; see, e.g., Holmström and Myerson (1983), Glycopantis and Yannelis (2004), Dutta and Vohra (2005), and Myerson (2007), among many others. In particular, Dutta and Vohra (2005) define an interim notion of the core and apply it to matching problems: the information that a deviating coalition should condition on is precisely the information that makes the deviation profitable; thus, the set of types that engage

in a deviation is endogenously determined as a fixed point. Yenmez (2013) develops an interim notion of stability explicitly for matching mechanisms, where types of two blocking players have a similar fixed-point feature.[4]

Our paper departs from this literature in fundamental ways. While the literature since the seminal work of Wilson (1978) has always had to deal with beliefs of players in an blocking coalition, it does not consider learning in the absence of blocking and hence the idea of endogenous on-path beliefs is simply not presented, not to mention the Bayesian consistency of on-path beliefs, off-path beliefs, and prior beliefs for stability. Equally important, we try to avoid making non-cooperative elements an indispensable component in the definition of a cooperative solution concept of a cooperative matching game.[5] The prior literature also assumes that allocations are unobservable; in our framework, players observe and make inferences from matching outcomes (or more generally signals about them). The relationship of observables and beliefs about unobservables is a key device in defining stability.

We would like to stress that the present paper does not study matching mechanisms.[6] Our approach treats matching problems as a class of cooperative games, and defines a solution concept directly, but we do not study dynamic processes that implement the solution concept. This approach is consistent with that of various solution concepts in economics, such as the core, competitive equilibrium, Nash equilibrium, sequential equilibrium, and so on.

1.2.2 Relation to the Belief-Free Approach

We would like to emphasize the difference from the belief-free approach of Liu, Mailath, Postlewaite, and Samuelson (2014) and subsequent papers. To take a non-cooperative game as an analog, the belief-free approach is similar to "rationalizability" while the approach of the present paper is similar to "equilibrium." The restriction of an equilibrium notion relative to rationalizability comes from the consistency of beliefs: players' posterior beliefs are correct and consistent with priors and across different players through Bayes' rule. The separation of off-path and on-path beliefs enables belief-based refinements and makes it possible to

---

[4]It should be pointed out that the fixed points in these papers are different from the fixed point in Liu, Mailath, Postlewaite, and Samuelson (2014): the latter fixed point is the set of stable matching outcomes that are immune to blockings, while the former concerns the self-fulfilling set of types of players in a blocking coalition, and this fixed point does not even appear if asymmetric information is one-sided as is considered in Liu, Mailath, Postlewaite, and Samuelson (2014) and this paper.

[5]Crawford (1985) emphasizes the restrictions of ad hoc non-cooperative modeling assumptions in the context of collective decision making studied by Holmström and Myerson (1983).

[6]It is a priori unclear what kinds of mechanisms are most appropriate without further information about the strategic interactions among players. As shown by Roth (1989), a standard direct revelation mechanism is not promising. The mechanism approach is plagued with the non-existence problem especially in the context of matching with interdependent values that we consider. We may even be so bold as to argue that a one-shot mechanism is not suited to handle consistent stable beliefs and hence the non-existence is not completely unexpected: consistent equilibrium beliefs are warranted for equilibrium existence in non-cooperative games.

evaluate Bayesian properties of stable outcomes. It is obvious that we cannot be content with a belief-free solution for Bayesian matching games.

The strength of the otherwise permissive belief-free concept of Liu, Mailath, Postlewaite, and Samuelson (2014) relies on a critical assumption that a firm knows perfectly a worker's type in a matching. The role of the assumption of knowledge is evident. It implies that the uninformed players know their ex post payoffs in a matching with asymmetric information. Thus, if there is a unique ex post individually rational matching outcome, it has to be stable; when this is not the case, Liu, Mailath, Postlewaite, and Samuelson (2014) show that monotonicity and modularity assumptions on payoffs facilitate belief-free operations over the set of types. Chen and Hu (2017) relax this knowledge assumption, but the resulting stable set becomes more permissive that includes that of Liu, Mailath, Postlewaite, and Samuelson (2014) as a subset.

1.2.3   Assumption on Observability within a Match

The present paper does not make the strong assumption of the knowledge of the firm about a worker's type in a match. That being said, our framework is flexible enough to accommodate the situation in which the uninformed players know the types of their partners and hence their ex post payoffs. More generally, each firm can observe a private signal that is correlated with the matching outcome and the underlying types of the workers. Technically, this merely amounts to additional restrictions on the on-path beliefs. See Section 6.2 for details.

For applications, the right assumption on what an uninformed firm can observe depends ultimately on the market situation we try to study. 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 to assume full information revelation within a match because this extra information is not used to stabilize the market in the first place; hence, 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 the labor markets of new college graduates, and for assignment problems and auctions where sellers do not directly observe the buyers' types when the market clears, although indirect inference can be made.

It is plausible that firms find out that their workers' true types in the continuation of the employment relationship (and this could lead to further market movements), although this information is not known before the finalization of the initial job matches (and hence should not be used to define stability of the initial job allocations). This additional information

9

should be used if we want to define stability for the market of this later stage. We indeed offer such a definition.

From a different perspective, in an equilibrium or steady-state situation of a dynamic market interaction, it is common that private information is not fully revealed (e.g., signaling and screening problems)—a cooperative solution concept is a reduced-form way to capturing it and, hence, exogenously imposing full revelation of information can be problematic. Instead, information revelation and belief updating is an equilibrium phenomenon.

2   The Model

The model is based on the matching games with transfers studied by Crawford and Knoer (1981). Following their terminology and notations, the matching is between a set of firms and a set of workers; but "firms" and "workers" are just semantics, and the model applies more generally.

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 summarized by j. Similarly, each worker i can also have publicly observable, payoff-relevant attributes that are summarized 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.[7] To ease notation, for a profile of workers' types $t = (t_i, t_{-i}) \in T$, we write $a_{ij}(t) := a_{ij}(t_i)$ and $b_{ij}(t) := b_{ij}(t_i)$. 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 each $i \in I$ and $j \in J$: (i) $\mu(i) \in J \cup \{i\}$, (ii) $\mu(j) \in I \cup \{j\}$, and

---

[7]The value is allowed to depend on 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 is worker i's observable characteristic and f_j is firm j's observable type.

10

(iii) $\mu(i) = j$ if and only if $\mu(j) = i$. Here $\mu(i) = i \in I$ means that worker i is unmatched; likewise 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 $p_{\mu(j)j} \in \mathbb{R}$ 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.[8]

2.3 Matching Function

For every $t = (t_1, ..., t_n) \in T$, some matching outcome $(\mu, \mathbf{p})$ is obtained. 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; i.e., for each type profile, $M(t) = (\mu, \mathbf{p})$ specifies a matching outcome. Three remarks are immediately needed.

Remark 1. The function $M : t \mapsto (\mu, \mathbf{p})$ describes a stable relationship between uncertainties and observables, and players agree on this relationship. This is similar to the classic rational expectations equilibrium approach to markets with asymmetric information, e.g., Radner (1979), where an equilibrium relationship is described by a mapping from unobservable uncertainties to publicly observable price vectors, and in equilibrium players correctly agree on this relationship. In our matching environment, it is natural to assume that the assignment μ is observable in addition to price vectors $\mathbf{p}$. Economists 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). ∎

Remark 2. The mapping $M : t \mapsto (\mu, \mathbf{p})$ 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 observability of match μ and transfers $\mathbf{p}$ simplifies the formulation of expected payoffs. If we are interested in the partial observability of matching outcomes, we can introduce private signals $\omega = (\omega_1, ..., \omega_{n+m})$ about $(\mu, \mathbf{p})$, and consider a mapping $M : t \mapsto \omega$. It will

---

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

not take sophisticated thinking to formalize this extension once we see the case of observable matching outcomes. This extension yields no new conceptual insights but necessitates more notations. ∎

Remark 3. We may impose the following measurability restriction on M: if any two types t_i and t'_i of the same worker i are ex ante indistinguishable (i.e., $a_{ij}(t_i) = a_{ij}(t'_i)$ 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

As explained in the Introduction, cooperative matching games make no assumption on the details of strategic interactions, and hence, unlike non-cooperative games with a well-defined extensive form, they make it difficult to formulate Bayesian consistency between exogenous prior beliefs and endogenous stable beliefs. The matching function approach helps to resolve this issue.

3.1 On-Path Beliefs and Individual Rationality

Let
$$M^{-1}(\mu, \mathbf{p}) = \{t \in T : M(t) = (\mu, \mathbf{p})\}$$

be the set of types for which $(\mu, \mathbf{p})$ is the matching outcome. The uninformed firms will update their belief to
$$\beta^1(\cdot) = \beta^0(\cdot|M^{-1}(\mu, \mathbf{p})). \tag{3.1}$$

This is an **on-path belief** of the matching M. Because firms share a common prior and identical observations, they have the same posterior on-path beliefs $\beta^1(\cdot)$. Section 6.2 considers the case of heterogeneous private on-path beliefs that are consistent with a common prior.

As we explained in the Introduction, the on-path belief is the endogenous belief when there is no longer an opportunity to deviate (pairwise or individually) from a matching. Does Equation (3.1) accomplish this? Not yet, but it will. So far, (3.1) describes the Bayesian consistency of posterior beliefs with the common prior β^0 when the putative matching M is in place. The to-be-defined stability of M (individual rationality and no pairwise blocking)

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^0 will be determined jointly rather than separately.

Definition 1. A matching M is individually rational for worker $i \in I$ if $a_{i\mu(i)}(t) + p_{i\mu(i)} \geq 0$ for all $t \in T$ and $(\mu, \mathbf{p}) = M(t)$. It is individually rational for firm $j \in J$ if $\mathbb{E}[b_{\mu(j)j}|M^{-1}(\mu, \mathbf{p})] - p_{\mu(j)j} \geq 0$ for all $(\mu, \mathbf{p}) \in M(T)$. A matching M is **individually rational** if it is individually rational for all $i \in I$ and $j \in J$.

Note that individual rationality is not necessarily ex post because β^1 may not be fully separating (see Section 1.2.2 for a discussion).

3.2 Off-Path Beliefs and Pairwise Blocking

Pairwise stability requires that there be no pairwise blocking; i.e., pairwise blocking is an off-path possibility if M is stable. A **blocking coalition** (i, j, p) consists of a worker $i \in I$, a firm $j \in J$, and a transfer $p \in \mathbb{R}$. We would like to formalize the following intuitive idea. A matching outcome $(\mu, \mathbf{p})$ is blocked by (i, j, p) if both worker i and firm j prefer to rematch with each other at a price p, and a matching M is blocked if there exists some $t \in T$ such that the matching outcome $M(t)$ is blocked by some coalition (i, j, p).

We consider strict incentives for both firms and workers in defining blocking, although this is inconsequential in matching with transfers. Worker i benefits from the blocking at $t \in T$ with $M(t) = (\mu, \mathbf{p})$ if and only if

$$a_{ij}(t) + p > a_{i\mu(i)}(t) + p_{i\mu(i)}. \tag{3.2}$$

Firm j does not directly observe the types of workers. From its perspective, worker i benefits from the blocking if and only if the type profile is in the following set:

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

We suppress the dependence of D_{ijp} on $(\mu, \mathbf{p})$ for notational simplicity. If (3.2) does not hold at any $t \in T$, i.e., if D_{ijp} is empty, then worker i will reject the blocking coalition (i, j, p). Now suppose that D_{ijp} is non-empty. Conditional on worker i's willingness to join the off-path blocking (i, j, p), firm j should obtain an **off-path belief** β^2. What should this off-path belief be? In the main analysis, we adopt the following specification:

$$\beta^2(\cdot) := \beta^1(\cdot|D_{ijp}) = \beta^0(\cdot|M^{-1}(\mu, \mathbf{p}) \cap D_{ijp}). \tag{3.3}$$

That is, firm j revises the prior belief β^0 by conditioning on not only the observables $M^{-1}(\mu, \mathbf{p})$ but also D_{ijp}, the set of types that make the worker willing to join the blocking coalition, *without discriminating any types in this set*. Given the definition of the on-path belief (3.1), the off-path belief is equivalently derived from the on-path belief $\beta^1(\cdot) = \beta^0(\cdot | M^{-1}(\mu, \mathbf{p}))$ conditional on the set of deviating types D_{ijp}. The formulation thus links the three beliefs (the prior β^0, the on-path belief β^1, and the off-path belief β^2) consistently by Bayes' rule.

Before proceeding, we should provide a remark for interested readers in order to clarify how our configuration of beliefs relates to the existing literature.

Remark 4. Although we shall offer a range of alternative configurations of off-path beliefs in Section 3.4, the specification by (3.3) is perhaps the most natural and obvious. Not surprisingly, the seminal papers on two-sided markets of asymmetric information utilize the idea without elaboration. For instance, Rothschild and Stiglitz (1976) and Wilson (1977) derive the off-equilibrium belief associated with an off-equilibrium contract from the prior belief by conditioning on the set of types that find the contract attractive. The idea is further advanced by Grossman and Perry (1986) in their definition of "perfect sequential equilibrium" and Mailath, Okuno-Fujiwara, and Postlewaite (1993) in their definition of "undefeated equilibrium." Dutta and Vohra (2005) use the same idea to define the notion of the credible core.

However, what is different in our formulation is that the off-path belief β^2 is derived from the on-path belief β^1, instead of the prior belief β^0, by conditioning on the set of deviating types D_{ijp}. The on-path stable belief β^1 is in turn derived from the prior belief conditional on observable matching outcomes $M(t)$. This feature does not appear in the aforementioned papers, and we stress that endogeneity of on-path stable beliefs is a crucial feature of our notion of stability, as clearly seen in Section 1.1. The Bayesian consistency of prior β^0, the on-path belief β^1, and the off-path belief β^2 makes it possible to discuss belief-based refinements. ∎

With the specification of the off-path belief β^2, firm j's expected payoff from the blocking coalition (i, j, p), or the off-path payoff, is

$$\mathbb{E}_{\beta^2}[b_{ij}] - p = \mathbb{E}[b_{ij} | M^{-1}(\mu, \mathbf{p}) \cap D_{ijp}] - p. \tag{3.4}$$

Conditional on the deviation, firm j will need to update its expected on-path payoff under $(\mu, \mathbf{p})$ with the off-path belief β^2:

$$\mathbb{E}_{\beta^2}[b_{\mu(j)j}] - p_{\mu(j)j} = \mathbb{E}[b_{\mu(j)j} | M^{-1}(\mu, \mathbf{p}) \cap D_{ijp}] - p_{\mu(j)j}. \tag{3.5}$$

14

Firm j joins the blocking coalition (i, j, p) if its expected payoff from the blocking is higher than its expected payoff from the putative matching:

$$\mathbb{E}_{\beta^2}[b_{ij}] - p > \max\left\{0, \mathbb{E}_{\beta^2}[b_{\mu(j)j}] - p_{\mu(j)j}\right\}. \tag{3.6}$$

The "max" operator on the right-hand side is required because, under the off-path belief β^2, the firm's expected payoff in the putative matching, (3.5), may be negative, in which case firm j would no longer find it individually rational to match with worker $\mu(j)$. Therefore, (3.4) being strictly larger than (3.5) does not ensure that firm j will join the coalition (i, j, p).

Definition 2. A matching M is **blocked** by a coalition $(i, j, p) \in I \times J \times \mathbb{R}$ if there exists $t \in T$, with $M(t) = (\mu, \mathbf{p})$, such that (3.2) and (3.6) hold.

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

Remark 5. Definition 2 describes a mutually profitable coalition. It is silent about how two players find each other and how they negotiate a transfer between them, a detail that is abstracted away from the cooperative model. This is the same as in Shapley and Shubik (1971). The definition simply tests a putative matching against an arbitrary counterfactual blocking coalition.

Still, readers might get the impression from the presentation that we are studying a particular game of coalition formation in which worker i first proposes to firm j, and only after observing (i, p), firm j updates its belief and responds. This interpretation is unnecessary and perhaps misleading. If one really wants to go down the path of finding a non-cooperative interpretation, it is perfectly reasonable to consider an alternative story of the firm approaching the worker (and players making offers alternately).

The gist is that when the uninformed firm computes its payoff from a coalitional blocking, it must condition on the counterfactual that a coalition is successfully formed and hence beneficial to the worker, regardless of how the coalition is formed. The firm does not need to be physically approached by the worker (or vice versa) in order to do this computation. This kind of conditional reasoning is familiar in games with interdependent values: we see it in auctions—a bidder bids conditional on the counterfactual that he is the winner (the reasoning of the winner's curse), or in voting—a voter votes conditional on the event that he is pivotal.

A cooperative model abstracts away from non-cooperative assumptions about who proposes to whom and how the coalition is formed. The evaluation of the off-path beliefs does not hinge on finding one specific non-cooperative game. Instead, our intuition of the matching market should be captured by the restrictions on the on-path and off-path beliefs, which

is permitted in the framework.[9] ∎

Remark 6. It is tempting to argue that the matching outcome $(\mu, \mathbf{p})$ is invalidated as soon as the updated on-path payoff (3.5) is strictly negative because firm j would reject its assignment $\mu(j)$ under $(\mu, \mathbf{p})$ regardless of whether or not firm j and worker i rematch with each other. This argument suggests that the "max" operator in (3.6) is unnecessarily restrictive for a notion of blocking, and that, instead, the given matching should be viewed as defeated as long as

$$\mathbb{E}_{\beta^2}[b_{ij}] - p > \mathbb{E}_{\beta^2}[b_{\mu(j)j}] - p_{\mu(j)j}. \tag{3.7}$$

This argument is flawed for the following reason. Note that (3.7) and (3.6) differ only when firm j's off-path payoff, (3.4), 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 given match $(\mu, \mathbf{p})$, but worker i also understands that he will not be hired by firm j even after the firm fires its worker. Therefore, worker i has no incentive to join the blocking in the first place.

One might still be tempted to 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. ∎

Remark 7. We consider a hypothetical blocking (i, j, p) to the putative matching M. But we do not consider further rounds of hypothetical deviations from the hypothetical blocking (i, j, p). This issue of "farsighted blocking," which intuitively makes a blocking coalition more difficult to form, is beyond the scope of this paper; see, e.g., Mauleon, Vannetelbosch, and Vergote (2011) and Ray and Vohra (2015) for related discussions in complete-information problems. Another issue is blocking that jointly involving multiple pairs of workers and firms, which in principle opens up more blocking opportunities. This will lead to the concept of the core, which we study in Section 6.1. ∎

3.3 Stability

We now summarize the definition of stability.

Definition 3. A matching M is **stable** if it is individually rational and is not blocked by any $(i, j, p) \in I \times J \times \mathbb{R}$. If M is a stable matching and $(\mu, \mathbf{p}) \in M(T)$, we say that $(\mu, \mathbf{p})$ is a **stable matching outcome**; the posterior beliefs $\beta^1(\cdot) = \beta^0(\cdot | M^{-1}(\mu, \mathbf{p}))$ and $\beta^2(\cdot) = \beta^0(\cdot | M^{-1}(\mu, \mathbf{p}) \cap D_{ijp})$ for each (i, j, p) are, respectively, an **on-path stable belief** and an **off-path stable belief** that support the matching outcome $(\mu, \mathbf{p})$.

Our approach is completely cooperative: we test a putative matching against all potential blocking coalitions. This is fully consistent with the approach of stability under complete information. It is easy to verify that when T is a singleton, the stability notion in Definition 3 coincides with the existing complete-information notion of pairwise stability.

Proposition 1. *A stable matching, in particular a measurable[10] stable matching, exists for each matching game (a, b, β^0).*

As a first step of the proof, we merge all types of worker $i \in I$ that are ex ante indistinguishable, and redefine each firm's matching values by taking the weighted average of the firm's original matching values over these indistinguishable types with respect to the prior belief. The remaining 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. This resulting M satisfies Definition 3 even though M is not invertible.

3.4 Refinements through Off-Path Beliefs

The literature on belief-based equilibrium refinements for non-cooperative games offers useful insights into alternative definitions of stability through the configuration of off-path beliefs β^2. Here we shall offer several alternative configurations. In all of these cases, the previous formulations of individual rationality (Definition 1), blocking (Definition 2), and stability (Definition 3) can be reproduced verbatim.

Arbitrary Belief. The most obvious alternative configuration is to require $\beta^2 \in \Delta(M^{-1}(\mu, \mathbf{p}) \cap D_{ijp})$ but make no further restriction, in the spirit of Fudenberg and Tirole's (1991) perfect Bayesian equilibrium. The arbitrariness of the off-path belief generally supports more stable outcomes and, consequently, the corresponding notion of stability is a coarsening of that in Definition 3.

Pessimistic Belief. We can define β^2 to be the belief under which the putative blocking appears to be the least favorable to the firm. That is, for a blocking coalition (i, j, p), the

---

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

corresponding off-path belief is such that

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

It bears emphasis that the same belief is used to compute the expected payoff from the off-path deviation and the expected payoff from the putative matching. This off-path belief makes blocking more difficult and, consequently, it leads to a coarser notion of stability.

Optimistic Belief. We can define β^2 to be the optimistic belief of the firm. That is, for a blocking coalition (i, j, p),

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

This off-path belief makes blocking easier and, consequently, leads to a strict refinement.

Dominance. We may follow Banks and Sobel (1987) and Cho and Kreps (1987) to examine how likely a worker will join a blocking (i, j, p). The idea of dominance (e.g., D1) compares the sets of responses of the uninformed players upon a deviation. In the cooperative matching game, a firm's "response" is simply the decision of whether to join the coalition, and hence we cannot directly apply the idea of dominance from the refinement literature. One possible approach in the spirit of D1 is to consider the maximal set of prices p' that induce blocking by certain worker types. This leads us to consider the set of worker types that benefit the *most* from joining a putative blocking (i, j, p), because, due to transferability, these types have the maximal set of prices p' under which (i, j, p') may block. Therefore, in the spirit of D1, we require that the off-path belief assign positive probability only to these types. Formally, consider a coalition (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

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

Thus B_{ijp} 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_{ijp}, that is,

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

Whether these restrictions make blocking easier or more difficult depends on whether the firm's preference is aligned with the worker's (recall that most dominance-based refinements are developed in signaling games with aligned preferences). If the matching value is such that a deviation is more attractive to the firm whenever it is more attractive to the worker, this off-path belief will make blocking much easier, and we obtain a refinement of our solution

concept.

Other Ideas. There are still many ways to define refinements. For instance, one plausible story is that to decide whether to join a coalition (i, j, p), firm j makes the assumption that any other coalition (i, j', p') with $j' \neq j$ is less attractive to worker i (not surprisingly, the mutual dominance of (i, j, p) and (i, j', p') could lead to non-existence of stable matching except for special cases). We could also utilize **tremble-based refinements** to model beliefs indirectly through the likelihood of trembles across different types. The experience with the refinement literature taught 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 applications, which is better left for future research. What is essential is that the separation of on-path and off-path beliefs opens up many possibilities and enables a coherent discussion of the (im)plausibility of stable matching outcomes in a purely cooperative framework without mixing cooperative and non-cooperative elements together.

3.5   Restrictions of On-Path Beliefs

In our model, all firms in a matching share a common on-path belief, because they share a common prior and make identical observations. In some applications, firms observe some attributes of their *own* workers *after* an initial match relationship is formed (see, e.g., Liu, Mailath, Postlewaite, and Samuelson 2014), and based on this new information, further market movements ensue until the matching market stabilizes. Stability at this stage can also be studied. To model this, 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 the unobservables. 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 $t_{\mu(j)}^2$ of its assigned worker $\mu(j)$, firm j's private on-path belief over T is

$$\beta_j^1(\cdot) := \beta^0(\cdot | M^{-1}(\mu, \mathbf{p}) \cap (T_{\mu(j)}^1 \times \{t_{\mu(j)}^2\} \times T_{-\mu(j)})),$$

and firm j's private off-path belief at a deviating coalition (i, j, p) is $\beta_j^2(\cdot) = \beta_j^1(\cdot | D_{ijp})$, where $D_{ijp} = \left\{ t' \in T : a_{ij}(t') + p > a_{i\mu(i)}(t') + p_{i\mu(i)} \right\}$.

Notions of individual rationality, blocking, and stability can be defined with respect to $\{\beta_j^1\}_{j \in J}$ and $\{\beta_j^2\}_{j \in J}$ in the same way as in Definitions 1, 2, and 3, respectively. The resulting stability notion captures a situation in which there is not pairwise blocking even after firms learn additional information about their own workers' types.

In Section 6.2, we shall consider a more general model of firms' heterogeneous private observations and their private beliefs.

4 Structural Properties of Stability

4.1 Notions of Match Efficiency

We are interested in (in)efficiency of matchings for the following reasons.

First, Shapley and Shubik (1971) observe that, under complete information (i.e., T is a singleton), *all* stable matching outcomes maximize the sum of individual players' surpluses, although multiplicity and indeterminacy of transfers and payoff distributions are generally inevitable. It is thus worthwhile to explore how this robust *allocative* efficiency property identified by Shapley and Shubik (1971) extends to asymmetric information.

Secondly, our definition of stability with asymmetric information is a consistency requirement for beliefs and matching outcomes. Hence efficiency, besides being a welfare property, is a joint restriction of matches and information. Indeed, we shall see that stability and competitive equilibrium, two outcome-equivalent concepts under complete information, imposes very different restrictions.

Thirdly, although efficiency is an enduring topic in economics, there is a tendency to shrug off efficiency under asymmetric information as implausible. However, efficiency has been established in classic models of two-sided markets under asymmetric information and with perfectly *transferrable* utilities and *independent* values, such as the Vickrey–Clarke–Groves (VCG) mechanisms and durable goods monopolies, demonstrating the limit of informational friction. Identifying stable matching environments with certain efficiency properties is therefore a reasonable goal. We are interested in both *independent* values (to compare stability with outcomes derived from the well-understood mechanisms) and *interdependent* values.

Lastly, we must stress that our match efficiency for stable matching is different from the traditional efficiency notion. Take the classic Akerlof (1970) market for lemons as an example. Suppose there is an uninformed buyer and a privately informed seller. If the condition is such that there is a complete breakdown of trade (a problem Akerlof initially considered), no trade will lead to no belief revision, and the inefficient no-trade outcome is unavoidably a stable outcome according to our definition. If instead Akerlof's equilibrium prescribes partial trade, there will be non-trivial belief updating conditional on no trade, and this equilibrium outcome cannot be stable. In the traditional analysis of the lemons problem, the updated belief after no trade is irrelevant because the market is exogenously shut down after a one-shot interaction. The updated belief becomes relevant if the market interactions are not exogenously shut down. This is a situation the solution concept of stability captures: there are always further opportunities for voluntary interactions. A good comparison is with the infinite-horizon dynamic market for lemons studied by Deneckere and Liang (2006). They show that gains from trade are eventually realized with probability 1,

albeit slowly. Our match efficiency is precisely about this kind of allocative efficiency, but it does not take into account the time that it takes to achieve this.

We study two notions of efficiency. The following is an obvious candidate.

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

The following is a less demanding notion of efficiency.

Definition 5. 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 overdoing it, we make the following comments to address questions raised by audiences.

Remark 8. Bayesian efficiency differs from full-information efficiency in only one aspect: it concerns surplus maximization conditional on each $M(t)$ instead of each $t \in T$. This is clear from the following equivalent statement: M is Bayesian efficient if for all $t \in T$ and $(\mu, p) = M(t)$, the match μ maximizes

$$\mathbb{E}\left[\sum_{i=1}^{n} (a_{i\mu'(i)} + b_{i\mu'(i)}) | \{t' \in T : M(t') = M(t)\}\right]$$

over all matchings $\mu' : I \cup J \to I \cup J$. It then follows immediately that a full-information efficient matching M is Bayesian efficient. ∎

Remark 9. In Definition 5, $\mathbb{E}\left[\sum_{i=1}^{n} a_{i\mu'(i)} | M^{-1}(\mu, \mathbf{p})\right]$ 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 the welfare from the viewpoint of an *outside observer* or an *analyst* 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' stable beliefs about the workers are *correct* (just as the equilibrium beliefs are correct in non-cooperative theory). ∎

Remark 10. To some readers, the fact that the match μ maximizes the expected surplus conditional on *endogenous* information $M(t) = (\mu, \mathbf{p})$ may seem unusual. In the matching model we consider, matching outcomes are publicly observable and reveal public information about the workers' types. This information should not be ignored when we evaluate matching outcomes, and this new notion of efficiency makes economic sense. In fact, readers who know the literature on incomplete markets with incomplete information (e.g., Radner 1979 and Grossman 1981) will immediately realize that conceptually this is not very different from evaluating efficiency conditional on publicly observable price information (albeit that literature focuses mainly on fully revealing equilibria because partial revealing equilibria are difficult to handle. We will study efficiency properties of all stable matchings of a given matching game, regardless of whether they are fully revealing or not). ∎

Remark 11. A different notion of efficiency has been suggested: each $(\mu, \mathbf{p}) \in M(T)$ maximizes

$$\mathbb{E}\left[\sum_{i=1}^{n}(a_{i\mu'(i)} + b_{i\mu'(i)})|M^{-1}(\mu', \mathbf{p}')\right] \tag{4.1}$$

over all $(\mu', \mathbf{p}')$. Notice that this definition is meaningful only if $M^{-1}(\mu', \mathbf{p}') \neq \emptyset$. The suggestion is mathematically equivalent to the requirement that (4.1) be constant over all outcomes in $M(T)$ in spite of the differences in conditional probabilities. Thus this alternative requirement has no reasonable economic interpretations. ∎

Bayesian inefficiency can persist and no pairwise recontracting arrangement can correct it (this is in sharp contrast with the competitive equilibrium notion studied in Section 5 where inefficiency can be unilaterally corrected). In the following example, there are multiple stable matchings, and only one of them is Bayesian efficient.

Example 3. Consider two workers and two firms. Suppose that $\beta^0 = \beta_1^0 \times \beta_2^0$, where $\beta_1^0(t_1) = \beta_1^0(t_1') = \beta_2^0(t_2) = \beta_2^0(t_2') = \frac{1}{2}$. The value matrix is as follows:

	firm 1	firm 2
t_1	$0, 0$	$0, 0$
t_1'	$-2, 5$	$-2, 0$
t_2	$-3, 0$	$-3, 4$
t_2'	$-4, 9$	$-4, 4$

Consider two stable matchings. In M^1, firm 1 is matched with worker 1, irrespective of his types, at a price of 2, and firm 2 is matched with worker 2, irrespective of his types, at a price of 4. This matching is Bayesian efficient. In M^2, firm 1 is matched with worker 2,

irrespective of his types, at a price of 4; worker 1 and firm 2 are unmatched. This matching is not Bayesian efficient. ∎

4.2 Stability and Efficiency

We now turn to economically interpretable assumptions on the matching game (a, b, β^0). Our payoff assumptions bring the model closer to the assignment problems studied by Koopmans and Beckmann (1957) and Shapley and Shubik (1971), but the model is more general and includes both multiple-object auctions and adverse-selection problems as special cases.

4.2.1 Full-Information Efficiency

Full-information efficiency is too demanding to be generally satisfied by a stable matching.

Assumption 1. $b_{ij}(t) = b_{ij}(t')$ for any $t, t' \in T$, $i \in I$ 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 summarized by i and j. A special case is $b_{ij}(\cdot) \equiv 0$, where the uninformed players care only about the transfers, and a privately informed player values the types of both players in a match $(a_{ij}(\cdot) + p)$. One application of this setting is multiple-object auctions in which privately informed bidders (workers) acquire heterogeneous objects (jobs). We do not make any restrictions on a_{ij}, although in auction applications, the bidder's valuation $a_{ij}(t)$. Under Assumption 1, b_{ij} can vary with j, which can be interpreted as the object owners' heterogeneous reservation values.

Proposition 2. *Suppose that Assumption 1 holds. Then a matching $M : t \mapsto (\mu, \mathbf{p})$, where $\mathbf{p} = (p_{i\mu(i)})_{i \in I}$, is stable if and only if for all $t \in T$, $M(t)$ is a complete-information stable matching when t is common knowledge. Consequently, M is full-information efficient.*

Proposition 2 is a basic test of our notion of stability. We know from the auction literature that under Assumption 1 a Vickrey–Clarke–Groves mechanism implements ex post efficient allocations. Assumption 1 is almost necessary (because of discrete types) for this result. Our stability notion conforms to this classic result.

4.2.2 Bayesian Efficiency

We are interested in conditions under which all stable matchings of a given matching game (a, b, β^0) are Bayesian efficient.

Assumption 2. $a_{ij}(t) = a_{ij}(t')$ for any $t, t' \in T$, $i \in I$, 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). There is no restriction on b_{ij}.

A special case of Assumption 2 that is of applied interest is that $a_{ij} \equiv 0$. This case captures a labor market in which workers care only about the salaries they receive ($a_{ij} + p = p$), while firms value the workers' private types (productivity) ($b_{ij} - p$).

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

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

A special case of Assumption 3 is the following familiar assumption adopted in many classic adverse-selection models such as signaling and screening: $a_{ij}(t) = g(i, t)$. 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 $a_{ij}(t)$ to vary with the worker's private type t and the worker's identity i, which summarize 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 argument of Shapley and Shubik (1971). Unlike their case of complete information, the proof for Proposition 3 is not quite immediate because surplus maximization and its dual are defined by on-path stable beliefs, but stability and blocking utilize off-path stable beliefs.

Proposition 3. *A stable matching 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.*

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 the welfare comparison is restricted only to matched agents. 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 $a_{ij}(\cdot) + p > 0$ and $b_{ij}(\cdot) - p > 0$). 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.

Assumption 2 and Assumption 3 cannot be dispensed with. The following two examples show that the full-match restriction in condition (ii) is tight.

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 for the following reasoning. Any blocking 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 stable 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 stable matching for any prices: 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. ∎

A natural question is what happens when firms are on the short side of the market, so workers cannot be fully matched—in this case condition (ii) of Proposition 3 does not apply. Proposition 3 makes assumptions only on the payoffs (a, b) and this means that all stable matchings of these matching games must be Bayesian efficient 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 4. *A stable matching M is Bayesian efficient if Assumption 3 holds, $a_{ij}(\cdot)$ and $b_{ij}(\cdot)$ are co-monotonic[11] for all $i \in I$ and $j \in J$, and the on-path beliefs are independent.*

The order over T with respect to which a_{ij} and b_{ij} are co-monotonic may vary with i and j. Thus this assumption 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, the lemons problem (Akerlof 1970) does not satisfy co-monotonicity and the inefficient no-trade outcome is in fact stable.

The independence assumption in Proposition 4 cannot be relaxed, as shown in the following example.

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 $\beta^0(t_1, t_2) = \beta^0(t_1', t_2') = \frac{1}{2}$. 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. Therefore, the 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. The firm's expected payoff in this matching is $\frac{1}{2} \times 4 + \frac{1}{2} \times 12 - 2 = 6$. Consider a potential blocking coalition that involves the firm and worker 1 with a price p. If the price p is such that only the type t_1 of worker 1 joins the blocking coalition, then the firm's payoff cannot be higher than 5.5, the total surplus produced by the pair. In this case the firm rejects the blocking coalition. If the price p is such that both types of worker 1 joins the blocking 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 blocking coalition. If the price p is such that only the type t_1' of worker 1 joins the blocking coalition, then the firm's payoff cannot be higher than 7, the total surplus produced by the pair. But because the two workers' types

---

[11] Two functions $a_{ij}(\cdot)$ and $b_{ij}(\cdot)$ are co-monotonic if $(a_{ij}(t) - a_{ij}(t'))(b_{ij}(t) - b_{ij}(t')) \geq 0$ for any $t, t' \in T$.

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 $12 - 2 = 10$. So the firm rejects the blocking with worker 1 in this case as well. ∎

5   Competitive Equilibrium

We now turn to the notion of competitive equilibrium, which is easier to formulate as it involves merely unilateral deviations and hence no off-path beliefs. The contrast of competitive equilibrium and stability elucidates the role of asymmetric information.

5.1   Motivation and Definition

For complete information matching and assignment problems, Koopmans and Beckmann (1957) and Shapley and Shubik (1971) construe the following notion of competitive equilibrium.[12] 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 each $i \in I$ 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 switching 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 process summarized by a competitive equilibrium has two critical differences from stability. 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. However, the postulates of **unilateral deviation** and **price-taking behavior** have different implications under asymmetric information, and manifest the differences of competitive equilibrium and stability.

A natural notion of a competitive equilibrium in an economy with uncertainty and without state-contingent contracts is rational expectations equilibrium (see Radner 1979 and Grossman 1981).[13] We now construe such a notion of rational expectations competitive

---

[12]Becker (1973) studies competitive equilibria in two-sided matching with a continuum of agents.

[13]Complete state-contingent contracts brings the problem back to complete information. Consequently, the Arrow–Debreu formulation of competitive equilibrium under uncertainty is not a suitable solution concept for our purpose.

equilibrium for two-sided matching markets.

A competitive matching with asymmetric information is a function $M : t \mapsto (\mu, \mathbf{p})$, where $\mathbf{p} = (p_{ij})_{i \in I, j \in J}$. We can 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 6. A matching $M : t \mapsto (\mu, \mathbf{p})$ is a **competitive equilibrium** if the following holds 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, because a unilateral deviation does not require mutual agreement. When T is a singleton, this definition reduces to the familiar notion of competitive equilibrium under complete information.

Proposition 5. *A (measurable) competitive equilibrium matching exists for each matching game (a, b, β^0).*

There are no general existence results for a standard rational expectations equilibrium in the prior literature (Kreps 1977 provides an early counterexample), except for generic existence results for a fully revealing equilibrium (the proof strategy is straightforward: to study an artificial economy with full information revelation and then verify that the equilibrium mapping is invertible for generic payoffs). But general existence is straightforward in the matching problem: the equilibrium in the artificial full information problem is a competitive equilibrium regardless of the invertibility of M. This is because the match, μ, is publicly observable.

5.2   Stability and Competitive Equilibrium

Stability and competitive equilibrium are two different ways of looking at a matching problem. With complete information, the two solution concepts lead to the same outcomes. This property is no longer true under informational asymmetry. 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 full 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 has to stem from the *incentives* and *information* embedded in their definitions.

Definition 7. 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 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; that is, a competitive equilibrium can support outcomes that are unstable.

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 the worker 2 is $p_{21} = -3$. By deviating to worker 2, the firm's expected payoff is $\frac{1}{2} \times (-4) + \frac{1}{2} \times 4 - (-3) = 3$. So 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 $2 - 3 = -1$. So neither type of worker 2 deviates. Hence, this matching is a competitive equilibrium.

The matching outcome of this competitive equilibrium cannot be stable. 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. ∎

This example demonstrates that flexible off-path prices allow for more information revelation, so it would suggest that stability refines competitive equilibrium with fixed prices. This intuition is again **incorrect**. The point is that having more information does not necessarily facilitate blocking when the rematch is ex post undesirable, but unilateral deviation may still be possible in a competitive environment with less information revelation. To confirm this point, the example below presents a stable matching that 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: 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 $\frac{1}{2} \times 1 + \frac{1}{2} \times 6 - p < 4.5$. So the firm does not deviate.

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

Although stable matchings and competitive matchings are generally not the same, they must have overlap. Formally, we have the following result.

Proposition 6. (i) *For any matching game (a, b, β^0), there exists a stable matching M^s that can be extended to a competitive matching. There exists some matching game (a, b, β^0) with a stable matching M^s 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. There exists a matching game (a, b, β^0) with a competitive matching M^c that is not an extension of a stable matching.*

5.3   Match Efficiency and 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 5, respectively, by taking into account $\mathbf{p} = (p_{ij})_{i \in I, j \in J}$. Remarks 8–11 apply here as well.

Recall that a stable matching is not guaranteed to be Bayesian efficient. In 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 an 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 somewhat surprising:

the new information generated from a blocking pair can prevent the inefficiency from being corrected.

We would like to reiterate that our notion of matching efficiency concerns different things from existing notions of efficiency. Our concepts of competitive equilibrium and stability capture situations in which further opportunities for voluntary interactions are always possible (as contrast to one-shot interactions). The match efficiency is about allocative efficiency, but it does not take into the process that it takes to achieve it.

We summarize our result here.

Proposition 7. *A competitive equilibrium matching $M : t \mapsto (\mu, \mathbf{p})$ is Bayesian efficient. If Assumption 1 holds, then 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

We shall offer two extensions of the stability concept. The first one concerns blockings by large coalitions. Without requiring new models, this concept of the core is indicative of the usefulness of the idea to markets with more general coalitional structures. The second extension concerns heterogeneous private observations of the uninformed players. It has the feature of two-sided private information and the formulation can be reinterpreted as introducing stochasticity into the matching function M.

6.1   The Core

Pairwise deviations are natural in two-sided markets. Conceptually, it is interesting to consider deviations by a coalition of multiple pairs of firms and workers. This leads us to consider the concept of the core. In complete information matching games, the core and stability coincide, but they differ under asymmetric information.

Endogenous beliefs should continue be a component of the core. Given a matching $M : t \mapsto (\mu, \mathbf{p})$, individual rationality is defined as in Definition 1. Consider the following blocking possibility: a subset of workers $I' \subset I$ and a subset of firms $J' \subset J$ walk away from the putative matching M 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'.[14] They block the matching if each of them is strictly better off from this rematch, conditional on the information revealed by their agreement to participate in the blocking coalition; again, the order in which players join the coalition is irrelevant, because each player, in evaluating his expected payoff from joining the coalition, must condition on the counterfactual event that the coalition is formed, i.e., everyone's agreeing to join the coalition.

Definition 8. A matching M is **blocked** by a coalition $(I', J', \mu', \mathbf{p}')$, where $I' \subset I$, $J' \subset J$, $\mu' : I' \cup J' \to I' \cup J'$ is a match, and $\mathbf{p}' = (p'_{i\mu'(i)})_{i \in I'}$ is a transfer scheme associated with the match μ', if there exists $t \in T$ with $M(t) = (\mu, \mathbf{p})$ 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) $\mathbb{E}\left[b_{\mu'(j)j} | M^{-1}(\mu, \mathbf{p}) \cap D \right] - p'_{\mu'(j)j} > \max\left\{ 0, \mathbb{E}\left[b_{\mu(j)j} | M^{-1}(\mu, \mathbf{p}) \cap D \right] - p_{\mu(j)j} \right\}$ for all $j \in J'$ where

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

Condition (ii) needs two remarks. First, 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). Secondly, the formulation implicitly excludes the possibility that $\mu'(j) = j$ for some $j \in J'$, i.e., j joins the blocking coalition but staying alone in the rematch, because otherwise the left-hand side of condition (ii) becomes $\mathbb{E}\left[b_{\mu'(j)j} | M^{-1}(\mu, \mathbf{p}) \cap D \right] - p'_{\mu'(j)j} = 0$, thus violating the condition. But this exclusion is without loss of generality because an unmatched firm j does not contribute any information or value to the blocking by other players.

We summarize the definition of the core below.

Definition 9. A matching M is in the **core** if it is individually rational and is not blocked by any coalition $(I', J', \mu', \mathbf{p}')$, where $I' \subset I$, $J' \subset J$, $\mu' : I' \cup J' \to I' \cup J'$ is a match, and $\mathbf{p}' = (p'_{i\mu'(i)})_{i \in I'}$ is a transfer scheme associated with the match μ'.

The core defined above is a strict refinement of stability.

Proposition 8. *If M is in the core, then it is stable; however, a stable matching M is not necessarily in the core.*

One direction is straightforward. A pairwise coalition $(i, j, p) \in I \times J \times \mathbb{R}$ is a special coalition $(I', J', \mu', \mathbf{p}')$ with $I' = \{i\}$, $J' = \{j\}$, $\mu'(i) = j$, and $p'_{ij} = p$. So if M is in the

---

[14]We have assumed that a player receives transfers only from his matched partner. An extension is straightforward.

core, then it is not blocked by any coalition including a pairwise deviation, and hence M is stable. 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 a deviating coalition 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. The existence of such a cycle is necessary for the core to strictly refines stability.

Example 8. Consider two workers and two firms. Suppose that $\beta^0 = \beta_1^0 \times \beta_2^0$, where $\beta_1^0(t_1) = \beta_1^0(t_1') = \beta_2^0(t_2) = \beta_2^0(t_2') = \frac{1}{2}$. 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 for the following reason. Let us consider pairwise blocking of worker i and firm $j = 3 - i$. For the firm to join the blocking, its expected payoff from this blocking 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. The blocking coalition 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 by deviating 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 following is an immediate corollary of Propositions 2, 3, and 8.

Corollary 1. *Suppose that a matching M is in the core. Then M is full-information efficient*

if Assumption 1 holds. Furthermore, 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, $a_{ij}(\cdot)$ and $b_{ij}(\cdot)$ are co-monotonic in t_i for all $i \in I$ and $j \in J$, and on-path beliefs are independent.

6.2 Private Beliefs and Stochastic Matching Functions

In our model, all firms in a matching share a common on-path belief, because they share a common prior and make identical observations. A natural question is how to model the firms' heterogeneous private observations and their private beliefs. We have also considered deterministic matching functions of the form $M : t \to (\mu, \mathbf{p})$. Naturally, we are interested in stochastic matching functions. The two tasks can be accomplished together.

6.2.1 Modeling Private Beliefs with Private Signals

For each $j \in J$, let S_j be the finite set of 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 $\sigma^0 \in \Delta(T \times S)$—modeling heterogeneous priors is straightforward.

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 $s_j \in S_j$, updates its belief to the conditional probability measure $\sigma^0(\cdot | T \times \{s_j\} \times S_{-j}) \in \Delta(T \times S)$, which we shall denote simply by $\sigma^0_{s_j}(\cdot)$.

To ease notation, we adopt the following harmless convention: for a non-empty subset $E \subset T \times S$, we write

$$\sigma^0_{s_j}(t | E) := \sigma^0_{s_j}(\{t\} \times S | E)$$

and for a function $f : T \to \mathbb{R}$, we write

$$\mathbb{E}_{\sigma^0_{s_j}}[f | E] := \sum_{t \in T} f(t) \sigma^0_{s_j}(t | E).$$

Definition 10. 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 when s is a public signal.

6.2.2 Stability with Private Beliefs

Each firm $j \in J$, after observing its private signal s_j and the matching outcome $(\mu, \mathbf{p})$, obtains a **private on-path belief** $\sigma^0_{s_j}(\cdot | M^{-1}(\mu, \mathbf{p}))$.

Definition 11. A matching M is **individually rational** if

$$a_{i\mu(i)}(t) + p_{i\mu(i)} \geq 0$$

and

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

for all $i \in I$, $j \in J$, $t \in T$, $s \in S$, and $(\mu, \mathbf{p}) = M(t, s)$.

Consider a blocking coalition $(i, j, p) \in I \times J \times \mathbb{R}$. The worker benefits from this coalition at $t \in T$ if

$$a_{ij}(t) + p > a_{i\mu(i)}(t) + p_{i\mu(i)}. \tag{6.1}$$

Let $D_{ijp} = \left\{ t' : a_{ij}(t') + p > a_{i\mu(i)}(t') + p_{i\mu(i)} \right\}$ be the set of types such that worker i prefers to join (i, j, p) rather than stay in the putative matching with firm $\mu(i)$. If D_{ijp} is empty, i.e., (6.1) is invalid, then (i, j, p) fails to be a successful blocking coalition. Thus suppose that D_{ijp} is non-empty. Conditional on D_{ijp}, firm j with a private on-path belief $\sigma^0_{s_j}(\cdot | M^{-1}(\mu, \mathbf{p}))$ will update its belief to $\sigma^0_{s_j}(\cdot | M^{-1}(\mu, \mathbf{p}) \cap (D_{ijp} \times S))$ by Bayes' rule. This is firm j's **private off-path belief**. Thus, firm j benefits from joining the blocking coalition (i, j, p) if

$$\mathbb{E}_{\sigma^0_{s_j}} \left[b_{ij} | M^{-1}(\mu, \mathbf{p}) \cap (D_{ijp} \times S) \right] - p > \max \left\{ 0, \mathbb{E}_{\sigma^0_{s_j}} \left[b_{\mu(j)j} | M^{-1}(\mu, \mathbf{p}) \cap (D_{ijp} \times S) \right] - p_{\mu(j)j} \right\}. \tag{6.2}$$

Definition 12. A matching M is **blocked** by $(i, j, p) \in I \times J \times \mathbb{R}$ if there exists $(t, s) \in T \times S$ with $(\mu, \mathbf{p}) \in M(t, s)$ such that both (6.1) and (6.2) are satisfied.

We are ready to define the notion of stability with private beliefs.

Definition 13. A matching $M : (t, s) \mapsto (\mu, \mathbf{p})$ is **stable** if it is individually rational and is not blocked by any $(i, j, p) \in I \times J \times \mathbb{R}$.

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 lays the foundation for further developments. The theory has immediate implications for empirical analysis of matching; see, e.g., Chiappori (2017) and Salanié (2015) for discussions of marriage models with transferable utilities. Existing empirical research utilizes the complete-information stability as the solution concept. Although these empirical models often allow certain characteristics of players to be unobservable to the analysts, they assume that the players themselves have complete information. The assumption is obviously restrictive and unrealistic.

We do not pretend that the insight developed in this paper is immediately applicable to market design questions. However, providing a Bayesian theory of stability is a necessary step toward understanding how players respond to the "rules and infrastructures" that coordinate and facilitate both decentralized and centralized transactions in matching environments with asymmetric information; see Kominers, Teytelboym, and Crawford (2017) and Sönmez and Ünver (2017) for discussions of the intricacies of practical market design.

It should also be clear that the idea of a Bayesian formulation of prior and stable beliefs is useful beyond the two-sided matching problem with transfers. It can be extended to market with more complex structures or more general coalitional games (see Forges and Serrano 2013 for a survey). The concept of the core studied in Section 6.1 is indicative in this direction.

A   Appendix

A.1   Proof of Proposition 1

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 $t \in T$, $i \in I$ and $j \in J$, define

$$a^*_{ij}(E(t)) \;\; = \;\; a_{ij}(t); \tag{A.1}$$

$$b^*_{ij}(E(t)) \;\; = \;\; \frac{1}{\beta^0(E(t))} \sum_{t' \in E(t)} b_{ij}(t')\beta^0(t'). \tag{A.2}$$

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.

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.1) 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,

$$
\begin{aligned}
\mathbb{E}\left[b_{\mu(j)j} | M^{-1}(\mu, \mathbf{p})\right] &= \frac{1}{\beta^0(M^{-1}(\mu, \mathbf{p}))} \sum_{t' \in M^{-1}(\mu, \mathbf{p})} b_{ij}(t') \beta^0(t') \qquad &\text{(A.3)} \\
&= \frac{1}{\beta^0(\cup_{\ell=1}^k E_\ell)} \sum_{\ell=1}^k \beta^0(E_\ell) \left(\frac{1}{\beta^0(E_\ell)} \sum_{t' \in E_\ell} b_{ij}(t') \beta^0(t') \right) \qquad &\text{(A.4)} \\
&= \frac{1}{\beta^0(\cup_{\ell=1}^k E_\ell)} \sum_{\ell=1}^k \beta^0(E_\ell) b^*_{ij}(E_\ell), \qquad &\text{(A.5)}
\end{aligned}
$$

where the last equality follows from (A.2). 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 any blocking coalition (i, j, p), any $t \in T$, and $(\mu, \mathbf{p}) = M(t)$ such that $a_{ij}(t) + p > a_{i\mu(i)}(t) + p_{i\mu(i)}$. Let $D_{ijp} = \left\{ t' \in T : a_{ij}(t') + p > a_{i\mu(i)}(t') + p_{i\mu(i)} \right\}$. If $t' \in D_{ijp}$, then $E(t') \subset D_{ijp}$. Therefore, $D_{ijp} \cap M^{-1}(\mu, \mathbf{p})$ can be written as a union of equivalent classes $F_1, ..., F_h$.

Following the same arguments as in (A.3), (A.4), and (A.5), we have

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

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,

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

That is, the matching M is not blocked by (i, j, p). ∎

A.2 Proof of Proposition 2

It follows from a similar construction as in Proposition 1 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 individual rationality of M, for any $t \in T$ with $M(t) = (\mu, \mathbf{p})$, we have

$$a_{i\mu(i)}(t) + p_{i\mu(i)} \geq 0 \text{ for all } i \in I \tag{A.6}$$

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

$$b_{\mu(j)j}(t) - p_{\mu(j)j} \geq 0 \text{ for all } j \in J. \tag{A.7}$$

Hence, (A.6) and (A.7) imply that $(\mu, \mathbf{p})$ is individually rational when there is complete information about t.

Consider any $(i, j, p) \in I \times J \times \mathbb{R}$ such that $a_{ij}(t) + p > a_{i\mu(i)}(t) + p_{i\mu(i)}$ for some $t \in T$ and $M(t) = (\mu, \mathbf{p})$. Then $D_{ijp} \neq \emptyset$. Since M is not blocked by (i, j, p), we have

$$\mathbb{E}\left[b_{ij}|M^{-1}(\mu, \mathbf{p}) \cap D_{ijp}\right] - p \leq \max\left\{0, \mathbb{E}\left[b_{\mu(j)j}|M^{-1}(\mu, \mathbf{p}) \cap D_{ijp}\right] - p_{\mu(j)j}\right\}. \tag{A.8}$$

By Assumption 1, $\mathbb{E}\left[b_{ij}|M^{-1}(\mu, \mathbf{p}) \cap D_{ijp}\right] = b_{ij}(t)$ and $\mathbb{E}\left[b_{\mu(j)j}|M^{-1}(\mu, \mathbf{p}) \cap D_{ijp}\right] = b_{\mu(j)j}(t)$ for any t. Inequality $(A.8)$ 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.7). Therefore, (i, j, p) does not block $(\mu, \mathbf{p})$ when there is complete information about t. We have thus proved that $(\mu, \mathbf{p})$ is complete-information stable for any $t \in T$. A stable matching under complete information maximizes the sum of surpluses, and hence the stable matching M is full-information efficient. ∎

A.3 Proof of Propositions 3 and 4

A.3.1 Duality of Bayesian Efficiency

Consider a stable matching M and any outcome $(\mu, \mathbf{p}) \in M(T)$. Let $\beta^1 = \beta^0(\cdot | M^{-1}(\mu, \mathbf{p}))$. Then M is Bayesian efficient if and only if μ maximizes

$$\sum_{i \in I} \left(\sum_{t \in T} \beta^1(t) a_{i\mu'(i)}(t) + \sum_{t \in T} \beta^1(t) b_{i\mu'(i)}(t) \right) \tag{A.9}$$

over all matches $\mu' : I \cup J \to I \cup J$. We proceed to show that μ indeed maximizes (A.9).

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

$$V(x) := \sum_{i \in I} \sum_{j \in J} x_{ij} \left(\sum_{t \in T} \beta^1(t) a_{ij}(t) + \sum_{t \in T} \beta^1(t) b_{ij}(t) \right)$$

subject to

$$\sum_{j \in J} x_{ij} \leq 1;$$
$$\sum_{i \in I} x_{ij} \leq 1;$$
$$x_{ij} \geq 0, \ i \in I, \ j \in J.$$

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 β^1. 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 the objective

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

subject to

$$u_i + v_j \ \geq \ \sum_{t \in T} \beta^1(t) a_{ij}(t) + \sum_{t \in T} \beta^1(t) b_{ij}(t), \ i \in I, \ j \in J; \tag{A.10}$$
$$u_i \ \geq \ 0, \ i \in I;$$
$$v_j \ \geq \ 0, \ j \in J.$$

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 conditional probability measures $\beta^1(\cdot|\cdot)$ are used to define stability whereas the probability β^1 appears in the dual problem (that is, a firm needs to update its beliefs before joining a blocking coalition).

A.3.2 Proof of Propositions 3 and 4

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

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

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 β^1. 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 3 and 4 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 β^1, and any $i \in I$ and $j \in J$,

$$a_{i\mu(i)}(t) + p_{i\mu(i)} + \sum_{t \in T} \beta^1(t)b_{\mu(j)j}(t) - p_{\mu(j)j} \qquad \text{(A.11)}$$
$$\geq \quad a_{ij}(t) + \sum_{t \in T} \beta^1(t)b_{ij}(t).$$

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 β^1 and some pair $(i,j) \in I \times J$, $\mu(i) \neq j$. Then, there exists $p \in \mathbb{R}$ such that

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

and

$$\sum_{t \in T} \beta^1(t)b_{\mu(j)j}(t) - p_{\mu(j)j} < \sum_{t \in T} \beta^1(t)b_{ij}(t) - p. \qquad \text{(A.13)}$$

Inequality (A.12) captures worker i's incentive to form a blocking coalition with firm j. Consider the set

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

By assumption, D_{ijp} is a non-empty set that contains $\bar{t}$.

Under condition (i) Proposition 3, i.e, Assumption 2, $a_{i\mu(i)}(t)$ and $a_{ij}(t)$ are independent of t. Therefore, $D_{ijp} = \left\{ t \in T : p_{i\mu(i)} < p \right\}$ if $\mu(i) \in J$, and $D_{ijp} = \left\{ t \in T : p_{i\mu(i)} < h(i,j) + p \right\}$ if $\mu(i) = i$, where $h(i,j) = a_{ij}(t)$ and $a_{ii}(t) = 0$. In either case, since $\bar{t} \in D_{ijp}$, $D_{ijp} = T$.

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

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

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

Under both conditions (i) and (ii) in Proposition 3, $\beta^1(D_{ijp}) = 1$. If we replace $\beta^1(t)$ by $\beta^1(t|D_{ijp})$ in (A.13), the inequality is unchanged. Therefore, (A.13) implies that firm j is willing to deviate with worker i. That is, (i,j,p) blocks M, a contradiction.

Suppose that the conditions of Proposition 4 holds, and $\mu(i) = i$ (the case of $\mu(i) \neq i$ is covered by the proof under condition (ii) of Proposition 3 already). Then

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

Since $a_{ij}(\cdot)$ and $b_{ij}(\cdot)$ are co-monotonic, there exists some linear order on T_i that is specific

41

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} depends only on t_i, the linear order naturally extends to an order on T). Therefore, D_{ijp} 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

$$\sum\nolimits_{t \in T} \beta^1(t) b_{ij}(t) - p \le \sum\nolimits_{t \in T} \beta^1(t|D_{ijp}) b_{ij}(t) - p. \tag{A.14}$$

Since $\mu(i) \ne j$, it follows from the independence of β^1 that

$$\beta^1(\{t_{\mu(j)}\} \times T_{-\mu(j)}) = \beta^1(\{t_{\mu(j)}\} \times T_{-\mu(j)}|D_{ijp}).$$

Hence

$$\sum\nolimits_{t \in T} \beta^1(t) b_{\mu(j)j}(t) - p_{\mu(j)j} = \sum\nolimits_{t \in T} \beta^1(t|D_{ijp}) b_{\mu(j)j}(t) - p_{\mu(j)j}. \tag{A.15}$$

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

$$\sum\nolimits_{t \in T} \beta^1(t|D_{ijp}) b_{\mu(j)j}(t) - p_{\mu(j)j} < \sum\nolimits_{t \in T} \beta^1(t) b_{ij}(t) - p. \tag{A.16}$$

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

$$\sum\nolimits_{t \in T} \beta^1(t|D_{ijp}) b_{\mu(j)j}(t) - p_{\mu(j)j} \le \sum\nolimits_{t \in T} \beta^1(t|D_{ijp}) b_{ij}(t) - p.$$

That is, firm j is willing to deviate with worker i. Thus, (i, j, p) blocks M, a contradiction. This establishes the claim that (A.11) holds.

Multiplying both sides of (A.11) by $\beta^1(t)$ and summing it up over t, we obtain

$$u_i^* + v_j^* \ge \sum\nolimits_{t \in T} \beta^1(t) a_{ij}(t) + \sum\nolimits_{t \in T} \beta^1(t) b_{ij}(t).$$

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

A.4   Proof of Proposition 5

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.1) and (A.2). 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.1),

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

That is, condition (i) of Definition 6 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.3)–(A.5), for all $j \in J$ and $i \in I \cup \{j\}$, we have

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

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

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,

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

Thus condition (ii) in Definition 6 is obtained. ∎

A.5   Proof of Proposition 6

By the proofs of Propositions 1 and 5, $M^s : t \mapsto (\mu, \mathbf{p}^s)$ 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.6   Proof of Proposition 7

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

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

where the expectation is taken over $\beta_1(\cdot) = \beta(\cdot | M^{-1}(\mu, \mathbf{p}))$. Since M is a competitive equilibrium,

$$\mathbb{E}_{\beta^1}\left[a_{i\mu(i)} + p_{i\mu(i)}\right] \geq \mathbb{E}_{\beta^1}\left[a_{i\mu'(i)} + p_{i\mu'(i)}\right] \tag{A.18}$$

for all $i \in I$, and

$$\mathbb{E}_{\beta^1}\left[b_{\mu(j)j} - p_{\mu(j)j}\right] \geq \mathbb{E}_{\beta^1}\left[b_{\mu'(j)j} - p_{\mu'(j)j}\right] \tag{A.19}$$

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

$$\mathbb{E}_{\beta^1}\left[\sum_{i=1}^{n} a_{i\mu(i)} + \sum_{j=1}^{m} b_{\mu(j)j}\right] \geq \mathbb{E}_{\beta^1}\left[\sum_{i=1}^{n} a_{i\mu'(i)} + \sum_{j=1}^{m} b_{\mu'(j)j}\right],$$

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

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

This contradicts (A.17).

We now show the statement about full-information efficiency. It follows from the construction of Proposition 5 that if $M(t)$ is a competitive equilibrium matching when t is common knowledge, then M is a competitive equilibrium. Suppose 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 competitive equilibrium when t is common knowledge. Hence $(\mu, \mathbf{p})$ is full-information efficient. ∎

References

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

[2] Alston, Max (2017). "On the non-existence of stable matches with incomplete information." Working Paper, Australian National University.

[3] 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.

[4] Banks, Jeffrey S., and Joel Sobel (1987). "Equilibrium selection in signaling games." *Econometrica* 55(3): 647–661.

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

[6] Blum, Yosef, Alvin E. Roth, and Uriel G. Rothblum (1997). "Vacancy chains and equilibration in senior-level labor markets." *Journal of Economic Theory* 76(2): 362–411.

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

[8] Chen, Yi-Chun and Gaoji Hu (2017). "Learning by matching." Working Paper, National University of Singapore.

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

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

[11] Crawford, Vincent P. (1985). "Efficient and durable decision rules: A reformulation." *Econometrica* 53(4): 817–835.

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

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

[16] Forges, Françoise, and Roberto Serrano (2013). "Cooperative games with incomplete information: Some open problems." *International Game Theory Review* 15(2): 1–17.

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

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

[19] Glycopantis, Dionysius, and Nicholas C. Yannelis, eds. (2006). *Differential Information Economies.* Vol. 19. Springer Science & Business Media.

[20] Grossman, Sanford J. (1981). "An introduction to the theory of rational expectations under asymmetric information." *The Review of Economic Studies* 48(4): 541–559.

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

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

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

[24] Kominers, Scott Duke, Alexander Teytelboym, and Vincent P. Crawford (2017). "An invitation to market design." *Oxford Review of Economic Policy* 33(4): 541–571.

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

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

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

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

[29] Mailath, George J., Masahiro Okuno-Fujiwara, and Andrew Postlewaite (1993). "Belief-based refinements in signalling games." *Journal of Economic Theory* 60(2): 241–276.

[30] 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.

[31] Myerson, Roger B. (2007). "Virtual utility and the core for games with incomplete information." *Journal of Economic Theory* 136(1): 260–285.

[32] Nash, John (1953). "Two-person cooperative games." *Econometrica* 21(1): 128–140.

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

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

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

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

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

[38] 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.

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

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

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

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

[43] Sönmez, Tayfun and M. Utku Ünver (2017). "Market design for living-donor organ exchanges: An economic policy perspective." *Oxford Review of Economic Policy* 33(4): 676–704.

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

[45] Wilson, Charles (1977). "A model of insurance markets with incomplete information." *Journal of Economic Theory*, 16(2): 167–207.

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