The (Non)-Existence of Stable Mechanisms in Incomplete Information Environments

Nick Arnosti[1], Nicole Immorlica[2], and Brendan Lucier[2(✉)]

[1] Department of Management Science and Engineering, Stanford University, Stanford, USA
narnosti@stanford.edu
[2] Microsoft Research, Cambridge, USA
nicimm@gmail.com, brlucier@microsoft.com

Abstract. We consider two-sided matching markets, and study the incentives of agents to circumvent a centralized clearing house by signing binding contracts with one another. It is well-known that if the clearing house implements a stable match and preferences are known, then no group of agents can profitably deviate in this manner.

We ask whether this property holds even when agents have *incomplete information* about their own preferences or the preferences of others. We find that it does not. In particular, when agents are uncertain about the preferences of others, *every* mechanism is susceptible to deviations by groups of agents. When, in addition, agents are uncertain about their *own* preferences, every mechanism is susceptible to deviations in which a single pair of agents agrees in advance to match to each other.

1 Introduction

In entry-level labor markets, a large number of workers, having just completed their training, simultaneously seek jobs at firms. These markets are especially prone to certain failures, including unraveling, in which workers receive job offers well before they finish their training, and exploding offers, in which job offers have incredibly short expiration dates. In the medical intern market, for instance, prior to the introduction of the centralized clearing house (the *National Residency Matching Program*, or NRMP), medical students received offers for residency programs at US hospitals two years in advance of their employment date. In the market for law clerks, law students have reported receiving exploding offers in which they were asked to accept or reject the position on the spot (for further discussion, see Roth and Xing [17]).

In many cases, including the medical intern market in the United States and United Kingdom and the hiring of law students in Canada, governing agencies try to circumvent these market failures by introducing a centralized clearing house which solicits the preferences of all participants and uses these to recommend a matching. One main challenge of this approach is that of incentivizing

N. Arnosti — Work conducted at Microsoft Research.

E. Markakis and G. Schäfer (Eds.): WINE 2015, LNCS 9470, pp. 46–59, 2015.
DOI: 10.1007/978-3-662-48995-6_4

participation. Should a worker and firm suspect they each prefer the other to their assignment by the clearing house, then they would likely match with each other and not participate in the centralized mechanism. Roth [15] suggests that this may explain why clearing houses that fail to select a stable match have often had difficulty attracting participants.

Empirically, however, even clearing houses which produce stable matches may fail to prevent early contracting. Examples include the market for Canadian law students (discussed by Roth and Xing [17]) and the American gastroenterology match (studied by Niederle and Roth [12] and McKinney et al. [11]). This is perhaps puzzling, as selecting a stable match ensures that no group of participants can profitably circumvent the clearing house ex-post.

Our work offers one possible explanation for this phenomenon. While stable clearing houses ensure that for *fixed, known* preferences, no coalition can profitably deviate, in most natural settings, participants contemplating deviation do so without complete knowledge of others' preferences (and sometimes even their own preferences). Our main finding is that in the presence of such uncertainty, *no mechanism* can prevent agents from signing mutually beneficial side contracts.

We model uncertainty in preferences by assuming that agents have a common prior over the set of possible preference profiles, and may in addition know their own preferences. We consider two cases. In one, agents have no private information when contracting, and their decision of whether to sign a side contract depends only on the prior (and the mechanism used by the clearing house). In the second case, agents know their own preferences, but not those of others. When deciding whether to sign a side contract, agents consider their own preferences, along with the information revealed by the willingness (or unwillingness) of fellow agents to sign the proposed contract.

Note that with incomplete preference information, agents perceive the partner that they are assigned by a given mechanism to be a random variable. In order to study incentives for agents to deviate from the centralized clearing house, we must specify a way for agents to compare lotteries over match partners. One seemingly natural model is that each agent gets, from each potential partner, a utility from being matched to that partner. When deciding between two uncertain outcomes, agents simply compare their corresponding expected utilities. Much of the previous literature has taken this approach, and indeed, it is straightforward to discover circumstances under which agents would rationally contract early (see the full version of the paper for an example). Such cases are perhaps unsurprising; after all, the central clearing houses that we study solicit only ordinal preference lists, while the competing mechanisms may be designed with agents' cardinal utilities in mind.

For this reason, we consider a purely ordinal notion of what it means for an agent to prefer one allocation to another. In our model, an agent debating between two uncertain outcomes chooses to sign a side contract only if the rank that they assign their partner under the proposed contract strictly first-order stochastically dominates the rank that they anticipate if all agents participate in the clearing house. This is a strong requirement, by which we mean that it is easy for a mechanism to be stable under this definition, relative to a definition relying

on expected utility. For instance, this definition rules out examples of beneficial deviations, where agents match to an acceptable, if sub-optimal, partner in order to avoid the possibility of a "bad" outcome.

Despite the strong requirements we impose on beneficial deviations, we show that every mechanism is vulnerable to side contracts when agents are initially uncertain about their preferences or the preferences of others. On the other hand, when agents are certain about their own preferences but not about the preferences of others, then there do exist mechanisms that resist the formation of side contracts, when those contracts are limited to involving only a pair of agents (i.e., one from each side of the market).

2 Related Work

Roth [14] and Roth and Rothblum [16] are among the first papers to model incomplete information in matching markets. These papers focus on the strategic implications of preference uncertainty, meaning that they study the question of whether agents should truthfully report to the clearinghouse. Our work, while it uses a similar preference model, assumes that the clearing house can observe agent preferences. While this assumption may be realistic in some settings, we adopt it primarily in order to separate the strategic manipulation of matching mechanisms (as studied in the above papers) from the topic of early contracting that is the focus of this work.

Since the seminal work of Roth and Xing [17], the relationship between stability and unraveling has been studied using observational studies, laboratory experiments, and theoretical models. Although work by Roth [15] and Kagel and Roth [5] concluded that stability plays an important role in encouraging participation, other papers note that uncertainty may cause unraveling even if a stable matching mechanism is used.

A common theme in these papers is that unraveling is driven by the motive of "insurance." For example, the closely related models of Li and Rosen [6], Suen [19], and Li and Suen [7,8] study two-sided assignment models with transfers in which binding contracts may be signed in one of two periods (before or after revelation of pertinent information). In each of these papers, unraveling occurs (despite the stability of the second-round matching) because of agents' risk-aversion: when agents are risk-neutral, no early matches form. Yenmez [20] also considers notions of interim and ex-ante stability in a matching market with transferable utility. He establishes conditions under which stable, incentive compatible, and budget-balanced mechanisms exist.

Even in models in which transfers are not possible (and so the notion of risk aversion has no obvious definition), the motive of insurance often drives early matching. The models presented by Roth and Xing [17], Halaburda [4], and Du and Livne [2] assume that agents have underlying cardinal utilities for each match, and compare lotteries over matchings by computing expected utilities. They demonstrate that unraveling may occur if, for example, workers are willing to accept an offer from their second-ranked firm (foregoing a chance to

be matched to their top choice) in order to ensure that they do not match to a less-preferred option.[1]

While insurance may play a role in the early contracting observed by Roth and Xing [17], one contribution of our work is to show that it is not necessary to obtain such behavior. In this work, we show that even if agents are unwilling to forego top choices in order to avoid lower-ranked ones, they might rationally contract early with one another. Put another way, we demonstrate that some opportunities for early contracting may be identified on the basis of ordinal information alone (without making assumptions about agents' unobservable cardinal utilities).

Manjunath [10] and Gudmundsson [3] consider the stochastic dominance notion used in this paper; however they treat only the case (referred to in this paper as "ex-post") where the preferences of agents are fixed, and the only randomness comes from the assignment mechanism. One contribution of our work is to define a stochastic dominance notion of stability under asymmetric information. This can be somewhat challenging, as agents' actions signal information about their type, which in turn might influence the actions of others.[2]

Perhaps the paper that is closest in spirit to ours is that of Peivandi and Vohra [13], which considers the operation of a centralized exchange in a two-sided setting with transferable utility. One of their main findings is that every trading mechanism can be blocked by an alternative; our results have a similar flavor, although they are established in a setting with non-transferrable utility.

3 Model and Notation

In this section, we introduce our notation, and define what it means for a matching to be *ex-post*, *interim*, or *ex-ante* stable. There is a (finite, non-empty) set M of men and a (finite, non-empty) set W of women.

Definition 1. *Given M and W, a* **matching** *is a function $\mu : M \cup W \to M \cup W$ satisfying:*

1. *For each $m \in M$, $\mu(m) \in W \cup \{m\}$*

[1] In many-to-one settings, Sönmez [18] demonstrates that even in full-information environments, it may be possible for agents to profitably pre-arrange matches (a follow-up by Afacan [1] studies the welfare effects of such pre-arrangements). In order for all parties involved to strictly benefit, it must be the case that the firm hires (at least) one inferior worker in order to boost competition for their remaining spots (and thereby receive a worker who they would be otherwise unable to hire). Thus, the profitability of such an arrangement again relies on assumptions about the firm's underlying cardinal utility function.

[2] Liu et al. [9] have recently grappled with this inference procedure, and defined a notion of stable matching under uncertainty. Their model differs substantially from the one considered here: it takes a matching μ as given, and assumes that agents know the quality of their current match, but must make inferences about potential partners to whom they are not currently matched.

2. *For each* $w \in W$, $\mu(w) \in M \cup \{w\}$
3. *For each* $m \in M$ *and* $w \in W$, $\mu(m) = w$ *if and only if* $\mu(w) = m$.

We let $\mathcal{M}(M, W)$ be the set of matchings on M, W.

Given a set S, define $\mathcal{R}(S)$ to be the set of one-to-one functions mapping S onto $\{1, 2, \ldots, |S|\}$. Given $m \in M$, let $P_m \in \mathcal{R}(W \cup \{m\})$ be m's ordinal preference relation over women (and the option of remaining unmatched). Similarly, for $w \in W$, let $P_w \in \mathcal{R}(M \cup \{w\})$ be w's ordinal preference relation over the men. We think of $P_m(w)$ as giving the *rank* that m assigns to w; that is, $P_m(w) = 1$ implies that matching to w is m's most-preferred outcome.

Given sets M and W, we let $\mathcal{P}(M, W) = \prod_{m \in M} \mathcal{R}(W \cup \{m\}) \times \prod_{w \in W} \mathcal{R}(M \cup \{w\})$ be the set of possible preference profiles. We use P to denote an arbitrary element of $\mathcal{P}(M, W)$, and use ψ to denote a probability distribution over $\mathcal{P}(M, W)$. We use P_A to refer to the preferences of agents in the set A under profile P, and use P_a (rather than the more cumbersome $P_{\{a\}}$) to refer to the preferences of agent a.

Definition 2. *Given* M *and* W, *and* $P \in \mathcal{P}(M, W)$, *we say that matching* μ *is* **stable at preference profile** P *if and only if the following conditions hold.*

1. *For each* $a \in M \cup W$, $P_a(\mu(a)) \leq P_a(a)$.
2. *For each* $m \in M$ *and* $w \in W$ *such that* $P_m(\mu(m)) > P_m(w)$, *we have* $P_w(\mu(w)) < P_w(m)$.

This is the standard notion of stability; the first condition states that agents may only be matched to partners whom they prefer to going unmatched, and the second states that whenever m prefers w to his partner under μ, it must be that w prefers her partner under μ to m.

In what follows, we fix M and W, and omit the dependence of $\mathcal{M}$ and $\mathcal{P}$ on the sets M and W. We define a *mechanism* to be a (possibly random) mapping $\phi : \mathcal{P} \to \mathcal{M}$. We use A' to denote a subset of $M \cup W$.

We now define what it means for a coalition of agents to *block* the mechanism ϕ, and what it means for a *mechanism* (rather than a matching) to be stable. Because we wish to consider randomized mechanisms, we must have a way for agents to compare lotteries over outcomes. As mentioned in the introduction, our notion of blocking relates to stochastic dominance. Given random variables $X, Y \in \mathbb{N}$, say that X *first-order stochastically dominates* Y (denoted $X \succ Y$) if for all $n \in \mathbb{N}$, $\Pr(X \leq n) \geq \Pr(Y \leq n)$, with strict inequality for at least one value of n.

An astute reader will note that this definition reverses the usual inequalities; that is, $X \succ Y$ implies that X is "smaller" than Y. We adopt this convention because below, X and Y will represent the ranks assigned by each agent to their partner (where the most preferred option has a rank of one), and thus by our convention, $X \succ Y$ means that X is preferred to Y.

Definition 3 (Ex-Post Stability). *Given* M, W *and a profile* $P \in \mathcal{P}(M, W)$, *coalition* A' **blocks mechanism** ϕ **ex-post** *at* P *if there exists a mechanism* ϕ' *such that for each* $a \in A'$,

1. $\Pr(\phi'(P)(a) \in A') = 1$, *and*
2. $P_a(\phi'(P)(a)) \succ P_a(\phi(P)(a))$.

Mechanism ϕ *is* **ex-post stable at profile** P *if no coalition of agents blocks* ϕ *ex-post at* P.
Mechanism ϕ *is* **ex-post stable** *if it is ex-post stable at* P *for all* $P \in \mathcal{P}(M, W)$.
Mechanism ϕ *is* **ex-post pairwise stable** *if for all* P, *no coalition consisting of at most one man and at most one woman blocks* ϕ *ex post at* P.

Note that in the above setting, because P is fixed, the mechanism ϕ' is really just a random matching. The first condition in the definition requires that the deviating agents can implement this alternative (random) matching without the cooperation of the other agents; the second condition requires that for each agent, the random variable denoting the rank of his partner under the alternative ϕ' stochastically dominates the rank of his partner under the original mechanism.

Note that if the mechanism ϕ is deterministic, then it is ex-post pairwise stable if and only if the matching it produces is stable in the sense of Definition 2.

The above notions of blocking and stability are concerned only with cases where the preference profile P is fixed. In this paper, we assume that at the time of choosing between mechanisms ϕ and ϕ', agents have incomplete information about the profile P that will eventually be realized (and used to implement a matching). We model this incomplete information by assuming that it is common knowledge that P is drawn from a prior ψ over $\mathcal{P}$. Given a mechanism ϕ, each agent may use ψ to determine the ex-ante distribution of the rank of the partner that they will be assigned by ϕ. This allows us to define what it means for a coalition to block ϕ ex-ante, and for a mechanism ϕ to be ex-ante stable.

Definition 4 (Ex-Ante Stability). *Given* M, W *and a prior* ψ *over* $\mathcal{P}(M, W)$, *coalition* A' **blocks mechanism** ϕ **ex-ante** *at* ψ *if there exists a mechanism* ϕ' *such that if* P *is drawn from the prior* ψ, *then for each* $a \in A'$,

1. $\Pr(\phi'(P)(a) \in A') = 1$, *and*
2. $P_a(\phi'(P)(a)) \succ P_a(\phi(P)(a))$.

Mechanism ϕ *is* **ex-ante stable at prior** ψ *if no coalition of agents blocks* ϕ *ex-ante at* ψ.
Mechanism ϕ *is* **ex-ante stable** *if it is ex-ante stable at* ψ *for all priors* ψ.
Mechanism ϕ *is* **ex-ante pairwise stable** *if, for all priors* ψ, *no coalition consisting of at most one man and at most one woman blocks* ϕ *ex-ante at* ψ.

Note that the only difference between ex-ante and ex-post stability is that the randomness in Definition 4 is over both the realized profile P and the matching produced by ϕ, whereas in Definition 3, the profile P is deterministic. Put another way, the mechanism ϕ is ex-post stable if and only if it is ex-ante stable at all deterministic distributions ψ.

The notions of ex-ante and ex-post stability defined above are fairly straightforward because the information available to each agent is identical. In order to study the case where each agent knows his or her own preferences but not

the preferences of others, we must define an appropriate notion of a blocking coalition. In particular, if man m decides to enter into a contract with woman w, m knows not only his own preferences, but also learns about those of w from the fact that she is willing to sign the contract. Our definition of what it means for a coalition to block ϕ in the interim takes this into account.

In words, given the common prior ψ, we say that a coalition A' *blocks* ϕ *in the interim* if there exists a preference profile P that occurs with positive probability under ψ such that when preferences are P, all members of A' agree that the outcome of ϕ' stochastically dominates that of ϕ, given their own preferences and the fact that other members of A' also prefer ϕ'. We formally define this concept below, where we use the notation $\psi(\cdot)$ to represent the probability measure assigned by the distribution ψ to the argument.

Definition 5 (Interim Stability). *Given M, W, and a prior ψ over $\mathcal{P}(M, W)$, coalition A'* **blocks mechanism** *ϕ* **in the interim** *if there exists a mechanism ϕ', and for each $a \in A'$, a subset of preferences $\mathcal{R}_a$ satisfying the following:*

1. *For each $P \in \mathcal{P}$, $\Pr(\phi'(P)(a) \in A') = 1$.*
2. *For each agent $a \in A'$ and each preference profile $\tilde{P}_a$, $\tilde{P}_a \in \mathcal{R}_a$ if and only if*
   (a) *$\psi(Y_a(\tilde{P}_a)) > 0$, where $Y_a(\tilde{P}_a) = \{P\colon P_a = \tilde{P}_a\} \cap \{P\colon P_{a'} \in \mathcal{R}_{a'}\ \forall a' \in A' \backslash \{a\}\}$*
   (b) *When P is drawn from the conditional distribution of ψ given $Y_a(\tilde{P}_a)$, we have $P_a(\phi'(P)(a)) \succ P_a(\phi(P)(a))$.*

Mechanism ϕ is **interim stable at** *ψ if no coalition of agents blocks ϕ in the interim at ψ.*
Mechanism ϕ is **interim stable** *if it is interim stable at ψ for all distributions ψ.*
Mechanism ϕ is **interim pairwise stable** *if, for all priors ψ, no coalition consisting of at most one man and at most one woman blocks ϕ in the interim at ψ.*

To motivate the above definition of an interim blocking coalition, consider a game in which a moderator approaches a subset A' of agents, and asks each whether they would prefer to be matched according to the mechanism ϕ (proposed by the central clearing house) or the alternative ϕ' (which matches agents in A' to each other). Only if all agents agree that they would prefer ϕ' is this mechanism used. Condition 1 simply states that the mechanism ϕ' generates matchings among the (potentially) deviating coalition A'.

We think of $\mathcal{R}_a$ as being a set of preferences for which agent a agrees to use mechanism ϕ'. The set $Y_a(\tilde{P}_a)$ is the set of profiles which agent a considers possible, conditioned on the events $P_a = \tilde{P}_a$ and the fact that all other agents in A' agree to use mechanism ϕ'. Condition 2 is a consistency condition on the preference subsets $\mathcal{R}_a$: (2a) states that agents in A' should agree to ϕ' only if they believe that there is a chance that the other agents in A' will also agree to ϕ' (that is, if ψ assigns positive mass to Y_a); moreover, (2b) states that in the cases when $P_a \in \mathcal{R}_a$ *and the other agents select* ϕ', it should be the case that a "prefers" the mechanism ϕ' to ϕ (here and in the remainder of the paper,

when we write that agent a prefers ϕ' to ϕ, we mean that *given the information available to* a, the rank of a's partner under ϕ' stochastically dominates the rank of a's partner under ϕ).

4 Results

We begin with the following observation, which states that the three notions of stability discussed above are comparable, in that ex-ante stability is a stronger requirement than interim stability, which is in turn a stronger requirement than ex-post stability.

Lemma 1. *If ϕ is ex-ante (pairwise) stable, then it is interim (pairwise) stable. If ϕ is interim (pairwise) stable, then it is ex-post (pairwise) stable.*

Proof. We argue the contrapositive in both cases. Suppose that ϕ is not ex-post stable. This implies that there exists a preference profile P, a coalition A', and a mechanism ϕ' that only matches agents in A' to each other, such that all agents in A' prefer ϕ' to ϕ, given P. If we take ψ to place all of its mass on profile P, then (trivially) A' also blocks ϕ in the interim, proving that ϕ is not interim stable.

Suppose now that ϕ is not interim stable. This implies that there exists a distribution ψ over $\mathcal{P}$, a coalition A', a mechanism ϕ' that only matches agents in A' to each other, and preference orderings $\mathcal{R}_a$ satisfying the following conditions: the set of profiles $Y = \{P : \forall a \in A', P_a \in \mathcal{R}_a\}$ has positive mass $\psi(Y) > 0$; and conditioned on the profile being in Y, agents in A' want to switch to ϕ' , i.e., for all $a \in A'$ and *for all* $P_a \in \mathcal{R}_a$ agent a prefers ϕ' to ϕ conditioned on the profile being in Y. Thus, agent a must prefer ϕ' even ex ante (conditioned only on $P \in Y$).

If we take ψ' to be the conditional distribution of ψ given $P \in Y$, it follows that under ψ', all agents $a \in A'$ prefer mechanism ϕ' to mechanism ϕ ex-ante, so ϕ is not ex-ante stable.

4.1 Ex-Post Stability

We now consider each of our three notions of stability in turn, beginning with ex-post stability. By Lemma 1, ex-post stability is the easiest of the three conditions to satisfy. Indeed, we show there not only exist ex-post stable mechanisms, but that any mechanism that commits to always returning a stable matching is ex-post stable.

Theorem 1. *Any mechanism that produces a stable matching with certainty is ex-post stable.*

Note that if the mechanism ϕ is deterministic, then (trivially) it is ex-post stable if and only if it always produces a stable matching. Thus, for deterministic mechanisms, our notion of ex-post stability coincides with the "standard" definition of a stable mechanism. Theorem 1 states further that any mechanism that

randomizes among stable matchings is also ex-post stable. This fact appears as Proposition 3 in [10].[3]

We next show in Example 1 that the converse of Theorem 1 does not hold. That is, there exist randomized mechanisms ϕ which sometimes select unstable matches but are nevertheless ex-post stable. In this and other examples, we use the notation $P_m : w_1, w_2, w_3$ as shorthand indicating that m ranks w_1 first, w_2 second, w_3 third, and considers going unmatched to be the least desirable outcome.

Example 1.

$$\begin{array}{ll} P_{m_1} : w_1, w_2, w_3 & P_{w_1} : m_3, m_2, m_1 \\ P_{m_2} : w_1, w_3, w_2 & P_{w_2} : m_2, m_1, m_3 \\ P_{m_3} : w_2, w_1, w_3 & P_{w_3} : m_3, m_2, m_1 \end{array}$$

There is a unique stable match, given by $\{m_1w_2, m_2w_3, m_3w_1\}$.

Lemma 2. *For the market described in Example 1, no coalition blocks the mechanism that outputs a uniform random matching.*

Proof. Because the random matching gives each agent their first choice with positive probability, if agent a is in a blocking coalition, then it must be that the agent that a most prefers is also in this coalition. Furthermore, any blocking mechanism must always match all participants, and thus any blocking coalition must have an equal number of men and women. Thus, the only possible blocking coalitions are $\{m_2, m_3, w_1, w_2\}$ or all six agents. The first coalition cannot block; if the probability that m_2 and w_2 are matched exceeds $1/3$, m_2 will not participate. If the probability that m_3 and w_2 are matched exceeds $1/3$, then w_2 will not participate. But at least one of these quantities must be at least $1/2$.

Considering a mechanism that all agents participate in, for any set of weights on the six possible matchings, we can explicitly write inequalities saying that each agent must get their first choice with probability at least $1/3$, and their last with probability at most $1/3$. Solving these inequalities indicates that any random matching μ that (weakly) dominates a uniform random matching must satisfy

$$\begin{aligned} \Pr(\mu = \{m_1w_1, m_2w_2, m_3w_3\}) &= \Pr(\mu = \{m_1w_2, m_2w_3, m_3w_1\}) \\ &= \Pr(\mu = \{m_1w_3, m_2w_1, m_3w_2\}), \end{aligned}$$

$$\begin{aligned} \Pr(\mu = \{m_1w_1, m_2w_3, m_3w_2\}) &= \Pr(\mu = \{m_1w_2, m_2w_1, m_3w_3\}) \\ &= \Pr(\mu = \{m_1w_3, m_2w_2, m_3w_1\}). \end{aligned}$$

But any such mechanism gives each agent their first, second and third choices with equal probability, and thus does not strictly dominate the uniform random matching.

Finally, the following lemma establishes a simple necessary condition for ex-post incentive compatibility. This condition will be useful for establishing non-existence of stable outcomes under other notions of stability.

[3] We thank an anonymous reviewer for the reference.

Lemma 3. *If mechanism ϕ is ex-post pairwise stable, then if man m and woman w rank each other first under P, it follows that* $\Pr(\phi(P)(m) = w) = 1$.

Proof. This follows immediately: if $\phi(P)$ matches m and w with probability less than one, then m and w can deviate and match to each other, and both strictly benefit from doing so.

4.2 Interim Stability

The fact that a mechanism which (on fixed input) outputs a uniform random matching is ex-post stable suggests that our notion of a blocking coalition, which relies on ordinal stochastic dominance, is very strict, and that many mechanisms may in fact be stable under this definition even with incomplete information. We show in this section that this intuition is incorrect: despite the strictness of our definition of a blocking coalition, it turns out that *no* mechanism is interim stable.

Theorem 2. *No mechanism is interim stable.*

Proof. In the proof, we refer to *permutations* of a given preference profile P, which informally are preference profiles that are equivalent to P after a relabeling of agents. Formally, given a permutation σ on the set $M \cup W$ which satisfies $\sigma(M) = M$ and $\sigma(W) = W$, we say that P' is the **permutation of P obtained by** σ if for all $a \in M \cup W$ and a' in the domain of P_a, it holds that $P_a(a') = P'_{\sigma(a)}(\sigma(a'))$.

The proof of Theorem 2 uses the following example.

Example 2. Suppose that each agent's preferences are iid uniform over the other side, and consider the following preference profile, which we denote P:

$$\begin{array}{ll} P_{m_1} : w_1, w_2, w_3 & P_{w_1} : m_1, m_2, m_3 \\ P_{m_2} : w_1, w_3, w_2 & P_{w_2} : m_1, m_3, m_2 \\ P_{m_3} : w_3, w_1, w_2 & P_{w_3} : m_3, m_1, m_2 \end{array}$$

Note that under profile P, m_1 and w_1 rank each other first, as do m_3 and w_3. By Lemma 1, if ϕ is interim stable, it must be ex-post stable. By Lemma 3, given this P, any ex-post stable mechanism must produce the match $\{m_1w_1, m_2w_2, m_3w_3\}$ with certainty. Furthermore, if preference profile P' is a permutation of P, then the matching $\phi(P')$ must simply permute $\{m_1w_1, m_2w_2, m_3w_3\}$ accordingly. Thus, on any permutation of P, ϕ gives four agents their first choices, and two agents their third choices.

Define the mechanism ϕ' as follows:

- If P' is the permutation of P obtained by σ, then

$$\phi'(P') = \{\sigma(m_1)\sigma(w_2), \sigma(m_2)\sigma(w_1), \sigma(m_3)\sigma(w_3)\}.$$

- On any profile that is not a permutation of P, ϕ' mimics ϕ.

Note that on profile P, ϕ' gives four agents their first choices, and two agents their second choices. If each agent's preferences are iid uniform over the other side, then each agent considers his or herself equally likely to play each role in the profile P (by symmetry, this is true even after agents observe their own preferences, as they know nothing about the preferences of others). Thus, conditioned on the preference profile being a permutation of P, all agents' interim expected allocation under ϕ offers a 2/3 chance of getting their first choice and a 1/3 chance of getting their third choice, while their interim allocation under ϕ' offers a 2/3 chance of getting their first choice and a 1/3 chance of getting their second choice. Because ϕ' and ϕ are identical on profiles which are not permutations of P, it follows that all agents strictly prefer ϕ' to ϕ ex-ante.

The intuition behind the above example is as follows. Stable matchings may be "inefficient", meaning that it might be possible to separate a stable partnership (m_1, w_1) at little cost to m_1 and w_1, while providing large gains to their new partners (say m_2 and w_2). When agents lack the information necessary to determine whether they are likely to play the role of m_1 or m_2, they will gladly go along with the more efficient (though ex-post unstable) mechanism.

In addition to proving that no mechanism is interim stable *for all priors*, Example 2 demonstrates that when the priori ψ is (canonically) taken to be uniform on $\mathcal{P}$, there exists no mechanism which is interim stable *at the prior* ψ. Indeed, if ϕ sometimes fails to match pairs who rank each other first, then such pairs have a strict incentive to deviate; if ϕ always matches mutual first choices, then all agents prefer to deviate to the mechanism ϕ' described above.

Theorem 2 establishes that it is impossible to design a mechanism ϕ that eliminates profitable deviations, but the deviating coalition in Example 2 involves six agents, and the contract ϕ' is fairly complex. In many settings, such coordinated action may seem implausible. One might ask whether there exist mechanisms that are at least immune to deviations by *pairs* of agents. The following theorem shows that the complexity of Example 2 is necessary: any mechanism that always produces a stable match is indeed interim pairwise stable.[4]

Lemma 4. *Any mechanism that produces a stable match with certainty is interim pairwise stable.*

Proof. Seeking a contradiction, suppose that ϕ always produces a stable match. Fix a man m, and a woman w with whom he might block ϕ in the interim. Note that m must prefer w to going unmatched; otherwise, no deviation with w can strictly benefit him. Thus, the best outcome (for m) from a contract with w is that they are matched with certainty. According to the definition of an interim blocking pair, m must believe that receiving w with certainty stochastically dominates the outcome of ϕ; that is to say, m must be certain that ϕ will give

[4] This result relies crucially on the fact that we're using the notion of stochastic dominance to determine blocking pairs. If agents instead evaluate lotteries over matches by computing expected utilities, it is easy to construct examples where two agents rank each other second, and both prefer matching with certainty to the risk of getting a lower-ranked alternative from ϕ (see the full version of the paper for an example).

him nobody better than w. Because ϕ produces a stable match, it follows that in cases where m chooses to contract with w, ϕ always assigns to w a partner that she (weakly) prefers to m, and thus she will not participate.

4.3 Ex-Ante Stability

In some settings, it is natural to model agents as being uncertain not only about the rankings of others, but also about their own preferences. One might hope that the result of Theorem 4 extends to this setting; that is, that if ϕ produces a stable match with certainty, it remains immune to pairwise deviations ex-ante. Theorem 3 states that this is not the case: ex-ante, no mechanism is even pairwise stable.

Theorem 3. *No mechanism is ex-ante pairwise stable.*

Proof. The proof of Theorem 3 uses the following example.

Example 3. Suppose that there are three men and three women, and fix $p \in (0, 1/4)$. The prior ψ is that preferences are drawn independently as follows:

$$P_{m_1} = \begin{cases} w_1, w_3, w_2 \text{ w.p. } 1-2p \\ w_2, w_1, w_3 \text{ w.p. } p \\ w_3, w_2, w_1 \text{ w.p. } p \end{cases} \qquad P_{w_1} = \begin{cases} m_1, m_3, m_2 \text{ w.p. } 1-2p \\ m_2, m_1, m_3 \text{ w.p. } p \\ m_3, m_2, m_1 \text{ w.p. } p \end{cases}$$

$$P_{m_2} = w_1, w_2 \qquad P_{w_2} = m_1, m_2$$

$$P_{m_3} = w_3 \qquad P_{w_3} = m_3$$

Because m_3 and w_3 always rank each other first, we know by Lemmas 1 and 3 that if mechanism ϕ is ex-ante pairwise stable, it matches m_3 and w_3 with certainty. Applying Lemma 3 to the submarket $(\{m_1, m_2\}, \{w_1, w_2\})$, we conclude that

1. Whenever m_1 prefers w_2 to w_1, ϕ must match m_1 with w_2 (and m_2 with w_1) with certainty.
2. Whenever w_1 prefers m_2 to m_1, ϕ must match w_1 with m_2 (and m_1 with w_2) with certainty.
3. Whenever m_1 prefers w_1 to w_2 and w_1 prefers m_1 to m_2, ϕ must match m_1 with w_1.

After doing the relevant algebra, we see that w_1 and m_1 each get their first choice with probability $1-3p+4p^2$, their second choice with probability p, and their third choice with probability $2p-4p^2$. If w_1 and m_1 were to match to each other, they would get their first choice with probability $1-2p$, their second with probability p, and their third with probability p; an outcome that they both prefer. It follows that ϕ is not ex-ante pairwise stable, completing the proof.

The basic intuition for Example 3 is similar to that of Example 2. When m_1 ranks w_1 first and w_1 does not return the favor, it is unstable for them to

match and m_1 will receive his third choice. In this case, it would (informally) be more "efficient"(considering only the welfare of m_1 and w_1) to match m_1 with w_1; doing so improves the ranking that m_1 assigns his partner by two positions, while only lowering the ranking that w_1 assigns her partner by one. Because men and women play symmetric roles in the above example, ex-ante, both m_1 and w_1 prefer the more efficient solution in which they always match to each other.

5 Discussion

In this paper, we extended the notion of stability to settings in which agents are uncertain about their own preferences and/or the preferences of others. We observed that when agents can sign contracts before preferences are fully known, every matching mechanism is susceptible to unraveling. While past work has reached conclusions that sound similar, we argue that our results are stronger in several ways.

First, previous results have assumed that agents are expected utility maximizers, and relied on assumptions about the utilities that agents get from each potential partner. Our work uses the stronger notion of stochastic dominance to determine blocking coalitions, and notes that there may exist opportunities for profitable circumvention of a central matching mechanism even when agents are unwilling to sacrifice the chance of a terrific match in order to avoid a poor one.

Second, not only can every mechanism be blocked under *some* prior, but also, for some priors, it is impossible to design a mechanism that is interim stable *at that prior*. This striking conclusion is similar to that of Peivandi and Vohra [13], who find (in a bilateral transferable utility setting) that for some priors over agent types, every potential mechanism of trade can be blocked.

In light of the above findings, one might naturally ask how it is that many centralized clearing houses have managed to persist. One possible explanation is that the problematic priors are "unnatural" and unlikely to arise in practice. We argue that this is not the case: Example 2 shows that blocking coalitions exist when agent preferences are independent and maximally uncertain, Example 3 shows that they may exist even when the preferences of most agents are known, and in the full version of the paper we show that they may exist even when one side has perfectly correlated (i.e. ex-post identical) preferences.

A more plausible explanation for the persistence of centralized clearing houses is that although mutually profitable early contracting opportunities may exist, agents lack the ability to identify and/or act on them. To take one example, even when profitable early contracting opportunities can be identified, agents may lack the ability to write binding contracts with one another (whereas our work assumes that they possess such commitment power). We leave a more complete discussion of the reasons that stable matching mechanisms might persist in some cases and fail in others to future work.

References

1. Afacan, M.O.: The welfare effects of pre-arrangements in matching markets. Econ. Theor. **53**(1), 139–151 (2013)
2. Du, S., Livne, Y.: Rigidity of transfers and unraveling in matching markets. Available at SSRN (2014)
3. Gudmundsson, J.: Sequences in Pairing Problems: A new approach to reconcile stability with strategy-proofness for elementary matching problems, November 2014. (Job Market Papers)
4. Halaburda, H.: Unravelling in two-sided matching markets and similarity of preferences. Games Econ. Behav. **69**(2), 365–393 (2010)
5. Kagel, J.H., Roth, A.E.: The dynamics of reorganization in matching markets: a laboratory experiment motivated by a natural experiment. Q. J. Econ. **115**(1), 201–235 (2000)
6. Li, H., Rosen, S.: Unraveling in matching markets. Am. Econ. Rev. **88**(3), 371–387 (1998)
7. Li, H., Suen, W.: Risk sharing, sorting, and early contracting. J. Polit. Econ. **108**(5), 1058–1091 (2000)
8. Li, H., Suen, W.: Self-fulfilling early-contracting rush. Int. Econ. Rev. **45**(1), 301–324 (2004)
9. Liu, Q., Mailath, G.J., Postlewaite, A., Samuelson, L.: Stable matching with incomplete information. Econometrica **82**(2), 541–587 (2014)
10. Manjunath, V.: Stability and the core of probabilistic marriage problems. Technical report, Working paper (2013)
11. McKinney, C.N., Niederle, M., Roth, A.E.: The collapse of a medical labor clearinghouse (and why such failures are rare). Am. Econ. Rev. **95**(3), 878–889 (2005)
12. Niederle, M., Roth, A.E.: The gastroenterology fellowship match: how it failed and why it could succeed once again. Gastroenterology **127**(2), 658–666 (2004)
13. Peivandi, A., Vohra, R.: On fragmented markets (2013)
14. Roth, A.E.: Two-sided matching with incomplete information about others' preferences. Games Econ. Behav. **1**(2), 191–209 (1989)
15. Roth, A.E.: A natural experiment in the organization of entry-level labor markets: regional markets for new physicians and surgeons in the united kingdom. Am. Econ. Rev. **81**(3), 415–440 (1991)
16. Roth, A.E., Rothblum, U.G.: Truncation strategies in matching markets in search of advice for participants. Econometrica **67**(1), 21–43 (1999)
17. Roth, A.E., Xing, X.: Jumping the gun: Imperfections and institutions related to the timing of market transactions. Am. Econ. Rev. **84**(4), 992–1044 (1994)
18. Sonmez, T.: Can pre-arranged matches be avoided in two-sided matching markets? J. Econ. Theor. **86**(1), 148–156 (1999)
19. Suen, W.: A competitive theory of equilibrium and disequilibrium unravelling in two- sided matching. RAND J. Econ. **31**(1), 101–120 (2000)
20. Yenmez, M.B.: Incentive-compatible matching mechanisms: consistency with various stability notions. Am. Econ. J. Microeconomics **5**(4), 120–141 (2013)