Essays on Market Design

Thesis by
Marcelo Ariel Fernandez

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2018
Defended May 8, 2018

Marcelo Ariel Fernandez
ORCID: 0000-0002-5475-0304

ABSTRACT

This thesis investigates the impact of incomplete information and behavioral biases in the context of market design.

In chapter 2, I analyze centralized matching markets and rationalize why the arguably most heavily used mechanism in applications, the deferred acceptance mechanism, has been so successful in practice, despite the fact that it provides participants with opportunities to "game the system." Accounting for the lack of information that participants typically have in these markets in practice, I introduce a new notion of behavior under uncertainty that captures participants' aversion to experience regret. I show that participants optimally choose not to manipulate the deferred acceptance mechanism in order to avoid regret. Moreover, the deferred acceptance mechanism is the unique mechanism within an interesting class (quantile stable) to induce honesty from participants in this way.

In chapter 3, co-authored with Leeat Yariv, we study the impacts of incomplete information on centralized one-to-one matching markets. We focus on the commonly used deferred acceptance mechanism (Gale and Shapley, 1962). We characterize settings in which many of the results known when information is complete are overturned. In particular, small (complete-information) cores may still be associated with multiple outcomes and incentives to misreport, selection of equilibria can affect the set of individuals who are unmatched—i.e., there is no analogue for the Rural Hospital Theorem, and agents might prefer to be on the receiving side of the of the algorithm underlying the mechanism. Nonetheless, when either side of the market has assortative preferences, incomplete information does not hinder stability, and results from the complete-information setting carry through.

In chapter 4, co-authored with Tatiana Mayskaya, we present a dynamic model that illustrates three forces that shape the effect of overconfidence (overprecision of consumed information) on the amount of collected information. The first force comes from overestimating the precision of the next consumed piece of information. The second force is related to overestimating the precision of already collected information. The third force reflects the discrepancy between how much information the agent expects to collect and how much information he actually collects in expectation. The first force pushes an overconfident agent to collect more information, while the second and the third forces work in the other direction. We show that under

some symmetry conditions, the second and third force unequivocally dominate the first, leading to underinvestment in information.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Chapter 1

INTRODUCTION

In this thesis I investigate the consequences of incomplete information and behavioral biases in the context of market design. Chapters 2 and 3 deal with the implications of incomplete information in the context of matching markets, like the ones that are used to assign doctors to residency programs in the U.S., or students to public schools in cities like Boston and New York. Chapter 4, studies the impact of overconfidence on the quality of decisions made by a decision maker, such as a judge's verdict during a bench trial.

In chapter 2, I provide a rationalization as of why the most frequently used matching mechanism, the deferred acceptance mechanism, has been so successful in practice, despite the fact that it is know to provide participants with incentives to "game the system." To do so I rely on two novel features: First, I take into account the limited amount of information that participants typically have in these markets. Second, I define a new notion of regret, and ask what should participants do if they wish to avoid regret. With these two elements I show that a participant will uniquely and optimally choose not to game the system in order to avoid regret, when the clearinghouse uses the deferred acceptance. And, moreover, that the deferred acceptance is the unique rule within an interesting class, to induce agents not to game the system as a way to avoid regret.

In chapter 3, co-authored with Leeat Yariv, we study the impact of incomplete information in centralized matching markets that use the deferred acceptance mechanism from a Bayesian perspective. In this context, we show that several desirable features that these markets present under complete information, cease to hold when information is incomplete, even when the underlying cores are unique. For instance, among others: (i) unstable matching outcomes are supported as Bayes Nash equilibria of the game induced by the direct revelation deferred acceptance mechanism; (ii) participants may benefit from being on the receiving side of the deferred acceptance; (iii) the number of matched individuals need not be the same across Bayes Nash equilibria outcomes. Nonetheless, we show that when all participants on either side of the market agree on the ranking of their potential matching partners, these desirable features are recovered.

In chapter 4, co-authored with Tatiana Mayskaya, we study the impact of overconfidence on the decisions to invest in costly information of an agent. We identify three forces that shape the impact of overconfidence on these investment decisions. These forces relate to the overvaluing the value of the information to be acquired, overvaluing the information already acquired, and the misinterpretation of noise as signal. Under some symmetry conditions, we show that overconfidence unequivocally leads to the agent underinvesting in information, and thus to a lower quality of decisions.

Chapter 2

DEFERRED ACCEPTANCE AND REGRET-FREE TRUTH-TELLING

2.1 Introduction

The deferred acceptance mechanism (Gale and Shapley, 1962) occupies a central place in the practice of market design. Among its various applications, it is used to assign graduating medical students to their first position as residents in the U.S., as well as to allocate entering students to public high schools in New York City.[1] However, it is known that participants may have the opportunity to "game the system" when the deferred acceptance is employed. In this paper, I argue that participants of these markets may optimally choose not to manipulate the deferred acceptance mechanism, if they wish to avoid regret (in a precise sense that I define). Moreover, I show that the deferred acceptance is unique within a class of mechanisms in the way that it induces all participants to be honest. Thus, the paper provides a rationalization of the success of the deferred acceptance in practice, as well as its salience with respect to other mechanisms.

In the typical design of a matching market, a centralized clearinghouse elicits the preferences of participants over their potential partners in the form of ranked order lists. It then uses this information to generate a matching using a known rule, particularly the deferred acceptance algorithm. Participants know their preferences, and the private report that they provide to the clearinghouse. However, they generally do not know each others preferences or reports. Moreover, privacy concerns limit the amount of information that is revealed even after the matching is implemented; meaning the reports remain private even ex-post.

A crucial element for the success of a matching market is that it generates matchings that are *stable*.[2] Broadly speaking, stable matchings are immune to renegotia-

[1] For a history of the NRMP and a list of labor markets that adopted the deferred acceptance see Roth (2008, Table 1). For the design of the New York City high school match, see Abdulkadiroğlu, Pathak, Roth, and Sönmez (2005), Abdulkadiroğlu, Agarwal, and Pathak (2017) and references therein. The UK and Israeli matching markets are discussed in Roth (1991) and Bronfman, Hassidim, Afek, Romm, Shreberk, Hassidim, and Massler (2015), respectively.

[2] Stable mechanisms have been reported to outlive their unstable counterparts, and in several instances have replaced them altogether (see Roth, 2008).

tion given the agents' preferences.[3] The deferred acceptance mechanism produces matchings that are stable *with respect to the reported preferences*. Thus, the clearinghouse's capacity to generate a stable matching depends on its ability to elicit agents' preferences truthfully. However, when agents know each others' preferences, theoretical results establish that mechanisms that generate matchings that are stable with respect to their reported preferences, such as the deferred acceptance, are manipulable; that is they provide incentives to agents to lie to the clearinghouse regarding their preference ranking. Thus, as pointed out in Roth and Rothblum (1999), there is a gap in our understanding as of why the deferred acceptance works so well in practice.

This paper reconciles stable and truth-telling matching mechanisms by taking a novel approach to incentives, and leveraging the presence of incomplete information found in most real-world markets. To do so, I introduce a notion of regret: an agent suffers *regret* if she takes an action, and ex-post she finds it to be dominated. I show that the most prominent matching mechanism, the deferred acceptance (DA), provides agents (on both sides of the market) incentives to report their preferences *truthfully* if they wish to avoid regret when the environment presents the features of a typical matching market discussed above. Additionally, I show that, for any agent, truth is the *unique* report that is guaranteed to be free of regret in a market that uses the DA. Furthermore, DA is the *unique* quantile stable mechanism for which truth-telling is regret-free.

To illustrate the notion of regret consider the following example:[4] Dr. Bob is looking for a residency program. He must participate in the National Residency Matching Program which runs the DA.[5] Assume (for expositional simplicity), that the NRMP is using the Hospital-proposing DA. There are only two hospitals in the market, Johns Hopkins (JH) and Mount Sinai (MS), with one available position each. Bob prefers Johns Hopkins over Mount Sinai. He also knows there is a second doctor, Dr. Alice, participating in the matching process.

All doctors and hospitals are required to send to the NRMP a list of partners ranked according to their preference. The hospital-proposing DA, or H-DA, uses the reports to simulate a sequence of proposals and rejections leading to a matching. In it, the

[3] Stability requires that there is no pair of agents on opposite sides of the market that prefer each other to their assigned partner in the matching, and that there is no agent that rather be unmatched to their assigned partner in the matching.

[4] To see how regret-avoidance compares to other standard notions of behavior see section 2.5.

[5] The workings and properties of this mechanism are described in section 2.2.

hospitals make proposals to their favorite doctors according to their reports. Doctors tentatively hold the best offer among those received, and reject the rest. Hospitals that were rejected can make new offers to doctors that have not rejected them yet. The process iterates until there are no more rejections.

Suppose Bob knows that the true and reported preferences of everyone in the market are:

$$\begin{array}{ll} Bob \text{ :Johns Hopkins} \succ \text{Mount Sinai} & Johns\ Hopkins \text{ :Alice} \succ \text{Bob} \\ Alice \text{ :Mount Sinai} \succ \text{Johns Hopkins} & Mount\ Sinai \text{ :Bob} \succ \text{Alice} \end{array}$$

Bob decides to try to game the system and lists *only* Johns Hopkins. Given the preferences and the rules of DA, this strategy is in fact successful. Had Bob told the truth he would have been matched to Mount Sinai. By deviating from the truth, in this scenario, he is matched to Johns Hopkins which is his top choice. This strategy is known as a *truncation*, it is a salient strategy in the literature and under complete information, it is sufficient to consider deviations from truth of the form of truncation strategies.

	True Bob	Alice		False Bob	Alice
1st	**MS**	**JH**		MS	**JH**
2nd					JH, **MS**
3rd				**JH**	
Matching	**MS**	**JH**		**JH**	**MS**

Table 2.1: Matchings resulting from hospital-proposing DA if Bob tells the truth (left) versus if he truncates (right). The rows correspond to the rounds in DA. **Bold** denotes an accepted offer, while *italics* denoted a rejected offer.

However, in real-world applications the participants have *incomplete information*; they do not know the preferences of others nor their reports. Would Bob truncate if he did not know the true preferences of the remaining players? Under this assumption Bob does not know in advance the outcome of the mechanism. One possibility is that Johns Hopkins actually finds Bob unacceptable, and if so, by performing a truncation he would remain unmatched. Bob's ex-post information consists of his true preferences, what he reported, the rules of the mechanism, and the resulting matching.

Can Bob come up with an alternative report that, given the information he now has, would have yield him a better outcome than remaining unemployed? The answer

is *Yes*. From the stability of DA, it follows that had he told the truth, he could not be worse off than remaining unemployed. Moreover, he knows that there exists a possible scenario (e.g. the one depicted above) where Mount Sinai wanted to hire him, but where his truncation prevented that matching from occurring. In such scenario, had he told the truth, he would have been matched to Mount Sinai which is strictly better for him than being unemployed. That means that truncating is ex-post dominated for Bob when he finds himself unemployed. In this case we say that Bob *regrets* truncating.

Regret is not merely a theoretical construct. There is evidence both in the psychology and experimental literature that (i) people have regret, (ii) fear of regret affects behavior, and (iii) the effect of anticipated regret is related to the information the subject knows will be revealed to him; see Gilovich and Medvec (1995); Bell (1982); Loomes and Sugden (1982, 1987).[6]

Formally, I focus on two-sided one-to-one matching problems where an individual's ordinal preference over his partners is private information.[7] Agents report their preferences to a centralized mechanism that chooses a matching for the reported preferences. The rules of the mechanism are common knowledge. The resulting matching is the only information observable ex-post; preferences of others remain unobserved.

In this paper I show that when the mechanism is DA, for every agent (on both sides of the market) and every possible deviation from the truth they could try, there is a scenario where the agent regrets deviating. The argument is based on the intuition presented in the above example. Moreover I show that when an agent regrets a deviation she only needs to consider truth as the report that dominates the deviation. More importantly, I show that an agent cannot regret reporting her preferences truthfully in DA. This stems from two observations. First, there are deviations from the truth that do not affect the resulting matching, e.g. permuting the order of alternatives that are preferred to the assigned partner. Second, for those deviations that may be profitable in some situation (e.g. truncations), there is an alternative preference profile, consistent with the observed matching, such that the deviation would yield a detrimental outcome. This ensures that the agent will never

[6] There are two other observations: (iv) there are different levels of regret; and (v) regret from an negative action taken is greater than regret of benefits by omitted action. The notion of regret in this paper captures the first three observations, but abstracts from the last two.

[7] Even though in practice most mechanism are many-to-one I restrict attention to one-to-one as a stepping stone in the analysis of DA and regret-free truth-telling.

suffer regret. Consequently, truth is a regret-free report. A regret-free report allows an agent to always be able to justify her action ex-post. It needs to be highlighted that in DA, truth is a regret-free report for *both* sides of the market.

Is regret-free truth-telling a property of stable mechanisms in general or is there something unique about DA? I study this question in the context of *quantile stable mechanisms*, a family that includes DA. Quantile stable mechanisms are the generalization of the *median stable mechanism*. The median stable mechanism ranks all stable matchings according to the preferences of one side of the market, and assigns each agent the partner they have in the median matching over that ordered set. This turns out to be a stable matching itself. The median stable matching has been shown to be a focal point in decentralized matching markets in the laboratory, see Echenique and Yariv (2011), and appears as a compromise solution contrasting the extremeness of DA. *Quantile stable mechanisms* are the generalization where, instead of the median, another quantile over the ordered set is selected, see Klaus and Klijn (2006); Chen, Egesdal, Pycia, and Yenmez (2014).

I show that among the set of quantile stable mechanisms described above, DA is the *unique* one that satisfies regret-free truth-telling. For any quantile stable mechanism that is not DA, there is a market and an agent who regrets reporting truthfully. To this end, I construct the report that dominates truth ex-post. To do so I fix a possible resulting matching from the mechanism and define the dominating report as respecting binary orders between alternatives, but declaring all alternatives that are worse than the resulting matching as unacceptable. I refer to this type of deviation as a "soft-truncation." I then show that all elements of the stable set under the "soft-truncation" contain weakly preferred matchings to the resulting matching. Therefore the quantile mechanism over a weakly preferred set results in a weakly better matching. For any quantile I give a construction that ensures the agent can be made strictly better through this "soft-truncation," thus obtaining that the agent regrets truth-telling when the quantile stable mechanism is not DA.

The paper contributes to several strands of literature within matching, reviewed in detail in section 2.7. The characterization of the DA through regret-free truth-telling complements the analysis of Chen, Egesdal, Pycia, and Yenmez (2016) and Pathak and Sönmez (2013) on selecting a stable mechanism based on its manipulability properties. An insightful strand of literature rationalizes the success of stable mechanisms by showing that the gains from manipulations can be small (in an appropriate sense) when the markets are large (Roth and Peranson, 1999; Immorlica

and Mahdian, 2005; Kojima and Pathak, 2009; Lee, 2017, among others). The characterization of the DA through regret-free truth-telling complements the large markets approach, since the latter is generally not able to distinguish among the incentives provided by different stable mechanisms. The paper also contributes to the analysis of stable matching under incomplete information under different behavioral notions (Barberà and Dutta, 1995; Roth and Rothblum, 1999; Ehlers, 2008; Ehlers and Massó, 2007, 2015), and to the understanding of what makes the deferred acceptance special (Kojima and Manea, 2010).

The structure of the paper is as follows: Section 2.2 introduces the framework and basic definitions in matching, including quantile stable mechanisms. Section 2.3 introduces the notion of regret. Section 2.4 presents the main characterization result. Section 2.5 discusses the relation between regret-free truth-telling and other notions of incentive compatibility. Section 2.6 concludes. Section 2.7 discusses the related literature. All proofs are relegated to Appendix A.

2.2 Framework

This section introduces basic definitions and recalls known results from the literature that are used later on, Roth and Sotomayor (1990) is the standard reference.

A one-to-one matching market is a triple $(M, W, \succ)$ where M is a finite set of men, W a finite set of women, and $\succ$ a preference profile composed of the preference relations of all men and women in the economy, that is the $(|M| + |W|)$-tuple $\succ = ((\succ_m)_{m\in M}, (\succ_w)_{w\in W})$. Each man m is endowed with a strict preference relation over $W \cup \{m\}$ denoted $(\succ_m)$, where the alternative $\{m\}$ represents the possibility of remaining single. Similarly $(\succ_w)$ is woman w's strict preference on $M \cup \{w\}$. The reflexive closure of $\succ_i$ is denoted by $\succeq_i$. $\mathcal{P}_i$ and $\mathcal{P}_{-i}$ denote the set of all possible preference relations of agent i and agents other than i respectively. The set of all possible preference profiles is denoted by $\mathcal{P}$.

A *matching* is a function $\mu : M \cup W \to M \cup W$ that assigns to each man (woman) either a woman (man) or himself (herself). Furthermore, the matching is restricted to be consistent, that is if a man is assigned a woman, that woman is getting assigned to him, reciprocally. Formally,

(i) $\mu(m) \in W \cup \{m\},\ \forall m \in M$;

(ii) $\mu(w) \in M \cup \{w\},\ \forall w \in W$;

(iii) $\mu(m) = w$ iff $\mu(w) = m,\ \forall m \in M,\ \forall w \in W$.

Denote $\mathcal{M}$ the set of all matchings for a fixed marriage market. $\mu(m)$ is m's partner under μ. Woman w is *acceptable* to m whenever $w \succ_m m$, otherwise w is *unacceptable* to m. The acceptable set for m is $A_m(\succ_m) = \{w \in W : w \succ_m m\}$, and $U_m(\succ_m) = W \setminus A_m(\succ_m)$ is the unacceptable set for m.

A matching μ is *individually rational* if every agent prefers their assigned partner to remaining single; that is, $\mu(i) \succeq_i i, \ \forall i \in M \cup W$. A matching μ is *blocked* by a pair (m, w) at $\succ$ if they prefer each other over their assigned partners; that is, $m \succ_w \mu(w)$ and $w \succ_m \mu(m)$. A matching is *stable* if it is individually rational at $\succ$ and it is not blocked by any pair (m, w) at $\succ$. $S(\succ)$ is the set of all stable matchings under preference profile $\succ$.

A *centralized matching mechanism* is an institution that receives reports of preferences from all agents in the economy and produces a matching; formally, it is a mapping $\phi : \mathcal{P} \to \mathcal{M}$. The notation $\phi(\succ)(i) = j$ means that j is i's partner under mechanism ϕ when the reported preferences are $\succ$. The mechanism ϕ is commonly known.

A matching mechanism is *stable* if $\forall \succ \in \mathcal{P}$, $\phi(\succ) \in S(\succ)$. Gale and Shapley (1962) showed that the set of stable matchings $S(\succ)$ is non-empty for any one-to-one matching market. In doing so, they introduced the deferred acceptance algorithm (DA) described below.

- *Step 0.* Given a marriage market $(M, W, \succ)$, denote the set of active men at time $t = 1$ by $\mathcal{A}_1$ and set it to be all men $\mathcal{A}_1 = M$.
- *Step 1.* Each man in the active set $\mathcal{A}_1$ proposes to his highest ranked acceptable woman according to $\succ_m$. Each woman selects the $(\succ_w)$-best acceptable partner out of those who proposed to her (if any) and they are tentatively matched; all other proposals are declared rejected. Set $\mathcal{A}_2$ to be the set of all men who made a proposal and were rejected. If $\mathcal{A}_2 = \emptyset$, stop; otherwise continue to step 2.
- *Step* $t \geq 2$. Each man in the active set $\mathcal{A}_t$ proposes to his highest ranked acceptable woman out of those who have not rejected him yet. Each woman then selects the best acceptable partner out of those who proposed to her at t and her tentative partner from $t-1$ (if any), and they are tentatively matched; all other proposals are declared rejected. Set $\mathcal{A}_{t+1}$ to be the set of all men who made a proposal at t and were rejected together with the set of men who

were tentatively matched at $t-1$ but rejected by their tentative match at t. If $\mathcal{A}_{t+1} = \emptyset$, stop; otherwise continue to step $t+1$.

This algorithm stops in finitely many steps and the resulting outcome is a stable matching. The above description corresponds to the Men-proposing or men-optimal stable matching (M-DA), where every man (weakly) prefers their assigned partner under this algorithm to the partner they would get in any other stable matching. In an analogous manner one can define the women-optimal (W-DA) stable matching.

Abusing notation, it is useful to have the binary relation $\succ_m$ hold over matchings $\mu, \mu' \in \mathcal{M}$: $\mu \succ_m \mu' \iff \mu(m) \succ_m \mu'(m)$. That is m prefers matching μ over μ' if and only if he prefers his partner under μ to his partner under matching μ'. Notice that a man is indifferent between two matches if he is matched to the same woman under both. The (side-)unanimous partial order $\succeq_M$ is then defined as: $\mu \succeq_M \mu' \iff (\forall m \in M)\,[\mu \succeq_m \mu']$ Similarly, the order $\succ_M$ is defined as $\mu \succ_M \mu' \iff [\mu \succeq_M \mu' \text{ and } \exists m \in M : \mu \succ_m \mu']$. The orders $\succeq_M, \succ_M$ represent the aligned preferences of men. A matching is $\succ_M$ preferable to another only if every man is weakly better and at least one is made strictly better off. Analogous definitions and constructions hold for women.

A known yet key fact is that the stable set forms a (distributive) lattice under order $\succeq_M$, with the men-optimal and women-optimal stable matchings as the extremal elements of this set under $\succeq_M$. Any two elements of the stable set are ranked inversely by men and women: $\mu, \mu' \in S(\succ)$, $\mu \succeq_M \mu' \iff \mu' \succeq_W \mu$, which is known as the opposition of interest property. An immediate corollary is that the men-optimal is the women-pessimal stable matching and vice versa. Another important property is that the set of agents who are matched is the same across all stable matchings. This property is referred to as the Lone Wolf or Rural Hospital Theorem (McVitie and Wilson, 1970).

An important property is that the men-optimal stable matching mechanism is strategy-proof *for men*. A matching mechanism ϕ is strategy-proof if $\forall \succ \in \mathcal{P}$ and $\forall i \in M \cup W$ it holds that $\phi(\succ_i, \tilde{\succ}_{-i}) \succeq_i \phi(\succ'_i, \tilde{\succ}_{-i})$, $\forall \succ'_i, \forall \tilde{\succ}_{-i}$. Strategy-proof *for men* requires the condition to hold only for men. That is, no matter what other men and women are reporting, a man cannot achieve a better partner by misrepresenting his preferences than he gets by reporting them truthfully.

The domain in which I focus on is the family of quantile stable mechanism, see Teo and Sethuraman (1998); Klaus and Klijn (2006); Chen, Egesdal, Pycia, and Yenmez

(2014).

Definition 2.1 (Chen, Egesdal, Pycia, and Yenmez (2014))**.** *Let* $q \in [0,1]$. *The* **q-*quantile stable matching mechanism*** *is the mapping* $\{\phi^q : \mathcal{P} \to \mathcal{M} | \mu : \forall m \in M,\ \mu(m)$ *is man* m*'s partner in his* $\lceil kq \rceil$*-th best stable matching according to order* $\succ_m$*, where* $k = |S(\succ)|\}$.[8]

An easy way to interpret quantile stable mechanisms is to think about the stable set as the size of a pie to be distributed between the set of men and set of women, and $q \in [0,1]$ as the share women are going to get. By choosing q, the designer anchors the ex-post distribution of payoffs across sides of the market, making it constant, regardless of other details such as the number of participants in the market, or their reports.

The distribution of payoffs being constant across markets also implies that quantile stable mechanisms are "easy to write," since they can be completely described with one parameter q. This is in the spirit of Wilson's critique (Wilson, 1987), posing as a desideratum for a mechanism not to depend on the fine details of the economy.[9] A particular case is that of the median stable matching mechanism, ($q = 1/2$) which assigns each individual the partner they have in the median-preferred stable matching. The median stable matching mechanism appears as a compromise solution between the two side-optimal stable mechanisms. Median stable matchings have been found to be salient in decentralized two-sided matching problem Echenique and Yariv (2011). The family of quantile stable mechanism is the family of all such compromises.

2.3 Regret and Regret-free Truth-telling

In this section I introduce the notion of regret, and define what it means for a mechanism to be regret-free truth-telling. In section 2.5 I discuss the relation of regret-free truth-telling to existing incentive compatibility notions.

[8] For simplicity of exposition we take $\lceil 0 \rceil = 1$ such that $\phi^0(\cdot) = \phi^M(\cdot)$; that is M-DA. Nothing depends on this assumption, alternatively one could define quantile over $\succeq_W$ and let $q = 1$.

[9] To illustrate this point consider a mechanism (see details in appendix A.3) that partitions the set of possible preference profiles into sets B and C, such that $\phi(\succ) = \phi^M(\succ)$ for $\succ \in B$ and $\phi(\succ) = \phi^W(\succ)$ for $\succ \in C$. This mechanism is stable since it always coincides with one of the allocations that a DA would assign. However one might find this mechanism undesirable since inessential changes (e.g. rearranging the order among the alternatives in an agent's unacceptable set) in the reported preference profile can lead to large jumps in the distribution of payoffs across men and women.

Regret captures the idea that individuals extremely dislike being *proven* wrong. That is, being able to infer given the outcome and observables that they could have done better (for sure) by taking a different action. There is strong evidence both in the psychology and experimental literature that (i) people have regret, (ii) fear of regret affects behavior, and (iii) the effect of anticipated regret is related to the information the subject knows will be revealed to him; see Gilovich and Medvec (1995); Bell (1982); Loomes and Sugden (1982, 1987).[10]

In what follows I assume that each agent knows their own preference, but not that of others. After the mechanism has generated a matching, the whole matching is observable to all agents, but the reports given to the mechanism remain private. Fix a matching market $(M, W, \succ)$ and a mechanism ϕ, and suppose agent i reports $\succ_i'$ to mechanism $\phi(\cdot)$. $\mathcal{M}|_{\succ_i'} = \{\mu \in \mathcal{M} : (\exists \succ_{-i} \in \mathcal{P}_{-i}) \ [\phi(\succ_i', \succ_{-i}) = \mu]\}$ is the set of matchings that are consistent with i's report and the known rules of the mechanism. This means that when reporting $\succ_i'$ agent i *knows* that the resulting matching μ must belong to $\mathcal{M}|_{\succ_i'}$. Suppose (ex-post) one such $\mu \in \mathcal{M}|_{\succ_i'}$ is observed by i. Then the *inference set* $\mathcal{I}(\mu; \succ_i', \phi) = \{\succ_{-i} \in \mathcal{P}_{-i} : \phi(\succ_i', \succ_{-i}) = \mu\}$ identifies the preference reports that are consistent with the observed matching, given his report, and the known rules of the mechanism. Player i *knows* that the reported preference profile is in this set.

Definition 2.2. *i **regrets** (reporting) $\succ_i'$ **at** $\mu \in \mathcal{M}|_{\succ_i'}$ **through** $\succ_i''$ **in** $\phi(\cdot)$ if $\exists \succ_i'' \in \mathcal{P}_i$ such that*

(i) for each $\succ_{-i} \in \mathcal{I}(\mu; \succ_i', \phi)$ it holds that $[\phi(\succ_i'', \succ_{-i}) \succeq_i \mu]$; and

(ii) for some $\tilde{\succ}_{-i} \in \mathcal{I}(\mu; \succ_i', \phi)$ it holds that $[\phi(\succ_i'', \tilde{\succ}_{-i}) \succ_i \mu]$.

In this context regretting a report simply means that the agent knows ex-post that his report is weakly dominated. Agent i knows that the reported preference profile lies within the inference set and, furthermore, there existed an alternative report that would have resulted in either matching him/her to the same or a strictly preferred partner.

Definition 2.3. *A report $\succ_i'$ is **regret-free** in $\phi(\cdot)$ if $\nexists (\mu, \succ_i'') \in (\mathcal{M}|_{\succ_i'}, \mathcal{P}_i)$ such that i regrets $\succ_i'$ at μ through $\succ_i''$ in $\phi(\cdot)$.*

[10] There are two other observations: (iv) there are different levels of regret; and (v) regret from an negative action taken is greater than regret of benefits by omitted action. The notion of regret in this paper captures the first three observations, but abstracts from the last two since the environment has ordinal preferences.

A regret-free report guarantees agents that they will never face regret. The criterion is stringent in that if a report is susceptible of regret at some matching, then no matter how unlikely that matching is, it is not regret-free. Standard and familiar notions use the same type of criteria. For example, dominant strategy incentive compatibility requires that reporting truthfully is weakly better than any other report for *every* possible report others could make (regardless of the probability that they will effectively use such report).

Definition 2.4. *A mechanism $\phi(\cdot)$ is **regret-free truth-telling** if for every market $(M, W, \succ)$ and every agent in it, truth-telling is regret-free.*

A regret-free truth-telling mechanism is one that assures agents they will not be *proven* wrong by stating their true preferences, using arguments that depend only in the same information the agent has.

2.4 Main Result

In this section I show that the DA (both men- and women-proposing) provides incentives to report truthfully to agents on *both* sides of the market if they want to avoid regret. Moreover, these incentives characterize the DA among quantile stable mechanisms (Theorem 2.1). The incentives are strict in the sense that truth is found to be the *unique* regret-free report in DA (Proposition 2.1).

Theorem 2.1. *A mechanism $\phi(\cdot)$ is a quantile stable regret-free truth-telling mechanism if and only if it is (either the men- or women-proposing) deferred acceptance mechanism.*

Proof. See Appendix A.1. □

I discuss the structure of the argument that carries the proof aided by an example. First, I argue that in the case of DA, no agent ever regrets reporting their preferences truthfully. Then, I show that for any quantile stable mechanism that is not DA, truth is susceptible of regret.

(DA is regret-free truth-telling). Suppose $\phi(\cdot)$ is the men-proposing DA, then $\phi(\cdot)$ is strategy-proof *for men* which means truth is a dominant-strategy in the induced direct revelation game, and therefore a fortiori truth is regret-free for men. Consequently the argument needs to address the incentives to report truthfully that the DA provides to the receiving side (women). I show that no deviation can dominate truth ex-post

in DA, for each type of possible deviation a woman can have; that is, there is no alternative report through which the agent may regret reporting truthfully.

For concreteness, consider a market $(M, W, \succ)$ with $|M| = |W| = 5$[11] where woman W_1 has preferences

$$W_1 : m_1 \succ m_2 \succ m_3 \succ m_4 \succ m_5.$$

The set of possible outcomes she faces by reporting truthfully can be divided into those in which she is matched to a man and those where she remains single. By the Rural Hospital Theorem, if she is not matched in the M-DA then she is not matched in any stable matching. This implies she would have remained single even if the mechanism had been W-DA. Therefore, there is no deviation that could have generated a better outcome. Basically, if she was unmatched in M-DA, it means she did not receive any proposals in the course of the algorithm. Consequently how she ranked her partners was irrelevant. Thus, she will not regret reporting her preferences truthfully if the matching she observes has her unmatched.

On the other hand, for any man that is acceptable to her there exist reports of others $(-W_1)$ such that each of man is her assigned stable partner in M-DA. The argument that follows will apply to any such matching, but it serves to go through a specific case: Suppose W_1 reports her preferences honestly and observes the matching:

$$\phi(\succ_{W_1}, \cdot) = \begin{pmatrix} W_1 & W_2 & W_3 & W_4 & W_5 \\ m_3 & m_2 & m_1 & m_4 & m_5 \end{pmatrix}.$$

Any report that could lead W_1 to regret must be belong to one (or a combination) of the following:

(i) permute the order among alternatives that are preferred by her to her assigned match,

(ii) declare someone who is preferred to the observed match as less preferred to it,

(iii) declare as preferable to the observed match someone who is not,

(iv) permute the order among alternatives that are less preferred to the observed match.

[11] I use a market of size five so that there is enough richness to portray all types of deviations a woman may have.

Examples of these are,

$$W_1 : m_1 \succ m_2 \succ m_3 \succ m_4 \succ m_5, \quad \text{(true)}$$
$$W_1' : \mathbf{m_2} \succ' \mathbf{m_1} \succ' m_3 \succ' m_4 \succ' m_5, \quad \text{(i)}$$
$$W_1' : m_1 \succ' \mathbf{m_3} \succ' \mathbf{m_2} \succ' m_4 \succ' m_5, \quad \text{(ii)}$$
$$W_1' : m_1 \succ' m_2 \succ' \mathbf{W_1} \succ' \mathbf{m_3} \succ' \ldots, \quad \text{(iii)}$$
$$W_1' : m_1 \succ' m_2 \succ' \mathbf{m_4} \succ' \mathbf{m_3} \succ' m_5, \quad \text{(iii')}$$
$$W_1' : m_1 \succ' m_2 \succ' m_3 \succ' \mathbf{m_5} \succ' \mathbf{m_4}. \quad \text{(iv)}$$

In order to affect the outcome of M-DA by changing her report, the woman needs to change the set of offers that are made to her during the process of the algorithm. The only way to do so is to change some decision she made over the offers she received when she was honest. Otherwise the same matching would ensue.

Permuting the order of the alternatives that are preferred to the observed match as in (i) is innocuous, that is it results in the same matching outcome. The reason is that, as the DA algorithm progresses, women are made weakly better off in each round. If W_1 had gotten an offer from either m_1 or m_2, given that she reported her preferences truthfully, she would have accepted it. That partner would be a lower bound to her matching outcome, and thus contradict the observed matching. Therefore no such offers could have been made, and how W_1 ranked them was irrelevant. It follows that changing the order among men in the upper contour set of the observed matching does not affect the result of the algorithm.

By the same reason as (i), declaring someone who is preferred to the assigned match as less preferred, as in (ii), is also innocuous. The set of decisions over *received offers* made by the algorithm on behalf of W_1 remains unchanged, and the same matching would result.

Reports of the form (iii) are known as truncations. Truncation happens when W_1 declares several acceptable men as unacceptable without changing their relative order. Truncations are known to be profitable deviations from the truth for agents in the receiving side in the context of complete information (Roth, 1982). This means there exist preference profiles for which truncating would yield W_1 a strictly better outcome. The question then becomes whether W_1 can be *sure* of a truncation's profitability given an observed matching. If so, truth would not be regret-free. I argue that, given that W_1 observes the resulting matching, but not the exact report that gave rise to the observe matching, she cannot conclude that a truncation dominates

truth-telling. A truncation involves risk, even ex-post when the agent has learned the resulting match.

To see that W_1 cannot regret reporting truthfully through a truncation, consider the preference profile where each individual finds their assigned partner as the only acceptable alternative, $\succ_i$: $\phi(\succ)(i) \succ i, \forall i \neq W_1$. It is straightforward to verify that this profile belongs to W_1's inference set.[12] Particularly, she might only be acceptable to m_3.[13] If so, by reporting m_3 to be unacceptable as in (iii) she would remain single which is a strictly worse outcome according to her true preferences. This shows that a truncation never dominates truth-telling ex-post (given the assumed information structure), and therefore W_1 cannot regret reporting truthfully through a truncation. A similar argument holds for (iii').

Truncation reports are salient in the literature for several reasons. First, truncations play a key role in proving that no stable strategy-proof mechanism exists (Roth, 1982). Second, considering truncations is sufficient to analyze profitable manipulations in the context of complete information, since the outcome of any profitable deviation can be achieved through a truncation (Roth and Vande Vate, 1991). Third, in a Bayesian setup and under a symmetry condition on agents beliefs, truncations are shown to first order stochastically dominate other untruthful reports (Roth and Rothblum, 1999). However, no unequivocal order between truncation and truthful reporting arises in their analysis. Ruling truncations susceptible of regret and truth regret-free is therefore relevant.

Lastly consider whether W_1 may regret truth-telling through an alternative report as in (iv). The preferences of the rest of the economy depicted in Figure 2.1a are in W_1's inference set, as Figure 2.1b shows. If these are the reported preferences by other participants, it means that by reporting honestly W_1 is able to obtain m_3 because at some point during the course of the DA she was faced with the decision of choosing between m_4 and m_5. By accepting m_4 over m_5 she generated a chain reaction of proposals that lead m_3 to propose to her (Figure 2.1b). Had she decided to accept m_5 over m_4 as (iv) prescribes, no such chain would have occurred. She would have been matched to m_5 who is less preferred by her to m_3, see Figure 2.1c.

Thus there does not exist a report through which W_1 can regret reporting her true

[12] If this was the case, then the DA process would be: every man made only one offer, to a distinct woman, in particular to their observed partner. Each woman accepted the offer received. The observed matching resulted.

[13] There is no need to assume she is unacceptable to everyone except m_3. The same argument holds if every man other than m_3 finds her acceptable, but less preferred than the observed partner.

$$
\begin{aligned}
W_2 &: m_2 \succ W_2 \\
W_3 &: m_1 \succ W_3 \\
W_4 &: m_4 \succ W_4 \\
W_5 &: m_5 \succ m_3 \\
m_1 &: W_3 \succ m_1 \\
m_2 &: W_2 \succ m_2 \\
m_3 &: W_5 \succ W_1 \\
m_4 &: W_1 \succ W_4 \\
m_5 &: W_1 \succ W_5
\end{aligned}
$$

(a) Preferences of $-W_1$

$$
\left(\begin{array}{ccccc}
W_1 & W_2 & W_3 & W_4 & W_5 \\
\hline
\mathbf{m_4}, m_5 & m_2 & m_1 & & m_3 \\
 & & & & m_3, \mathbf{m_5} \\
\mathbf{m_3}, m_4 & & & & \\
 & & & \mathbf{m_4} &
\end{array}\right)
$$

(b) DA with $\succ_{W_1}$

$$
\left(\begin{array}{ccccc}
W_1 & W_2 & W_3 & W_4 & W_5 \\
\hline
m_4, \mathbf{m_5} & \mathbf{m_2} & \mathbf{m_1} & & \mathbf{m_3} \\
 & & & \mathbf{m_4} &
\end{array}\right)
$$

(c) DA with $\succ'_{W_1}$

Figure 2.1

preference profile.

Although the argument was shown in the context of a specific example, Appendix A.1 shows it works for any agent and any market. Therefore DA is *regret-free truth-telling*. □

Having argued that the DA is regret-free truth-telling, I now turn to other quantile stable mechanisms and show that no quantile mechanism except DA is regret-free truth-telling. To illustrate the argument I provide an example where an agent regrets truth-telling in the context of the median mechanism. The proof for general quantile mechanisms can be found in the Appendix A.1.

(Regretting truth in the median mechanism). Consider a matching market with $|M| = |W| = 5$, and let W_1's preferences be

$$W_1 : m_1 \succ m_2 \succ m_3 \succ m_4 \succ m_5.$$

Suppose that W_1 reports her preferences honestly to a clearinghouse that uses the median stable mechanism ($q = 1/2$) to generate matchings, and that she observes the matching:

$$\phi(\succ_{W_1}, \cdot) = \begin{pmatrix} W_1 & W_2 & W_3 & W_4 & W_5 \\ m_3 & m_2 & m_1 & m_4 & m_5 \end{pmatrix} = \mu.$$

I claim that W_1 regrets truth-telling when she observes μ, and she does so through the alternative report:

$$W_1' : m_1 \succ' m_2 \succ' m_3 \succ' W_1 \succ' m_4 \succ' m_5.$$

To show the claim I procede in two steps: Step 1. Under the alternative $\succ'_{W_1}$, W_1 would obtain a matching that is *weakly* better than the observed one for *any* report of others in W_1's inference set; thus satisfying condition (i) of the regret definition. Step 2. Under the alternative $\succ'_{W_1}$, W_1 would obtain a matching that is *strictly* better than the observed one for *some* report of others in W_1's inference; thus satisfying condition (ii) of the regret definition.

Step 1. In order to show that W_1 regrets truth-telling through $\succ'_{W_1}$ one needs to understand how the matching outcome when W_1 reports $\succ'_{W_1}$ compares to the one obtained when W_1 reports $\succ_{W_1}$, given the observed matching μ. Since a q-quantile stable matching mechanism depends on the whole set of stable matchings it is important to see how these relate.

The following lemma says that any stable matching under $(\succ'_{W_1}, \succ_{-W_1})$ is also a stable matching under $(\succ_{W_1}, \succ_{-W_1})$. This provides crucial information about what the stable set would look like under each type of report, and consequently about what matching would be implemented under each report. In fact, the stable matchings under $(\succ'_{W_1}, \succ_{-W_1})$ are exactly those that are stable under $(\succ_{W_1}, \succ_{-W_1})$ and that are weakly preferred by W_1 to μ. The lemma has the key implication that ex-post W_1 knows she would have incurred in no risk by using $\succ'_{W_1}$ instead of $\succ_{W_1}$.

Lemma 2.1. $S(\succ'_{W_1}, \succ_{-W_1}) = \{\mu' \in S(\succ_{W_1}, \succ_{-W_1}) : \mu' \succeq_{W_1} \mu = \phi^q(\succ)\}$, $\forall \succ_{-W_1} \in \mathcal{I}(\mu; \succ_{W_1}, \phi^q)$.

Proof. In Appendix A.1. □

The lemma implies that instead of having to compute the stable set under $\succ'_{W_1}$, it is sufficient to look at the set of stable matchings under $\succ_{W_1}$ that are weakly preferred to μ; a fact I will leverage on later. Since W_1 takes the reports of others in the inference set as given, one only has to deal with the change in the stability constraints that involve W_1; that is, whether by changing her report she is creating a new block, individual or pairwise. For W_1, any individual rational matching that is weakly better than μ is also an individually rational matching under $\succ_{W_1}$, and vice versa. This follows immediately from $A_{W_1}(\succ'_{W_1}) = \{m_1, m_2, m_3\} \subseteq$

$\{m_1, m_2, m_3, m_4, m_5\} = A_{W_1}(\succ_{W_1})$ and $\phi^q(\cdot)$ being a stable mechanism. The fact that $\succ'_{W_1}$ respects m_1's preference relations amongst men ensures she does not create new blocking pairs. Lastly, since W_1 is matched under μ ($\mu(W_1) = m_3$) it follows from the Rural Hospital Theorem that she must be matched in every stable matching, in particular in the Women-optimal stable matching, which by its optimality must match her to a weakly preferred partner: $\phi^W(\succ_{W_1}, \succ_{-W_1})(W_1) \in \{m_1, m_2, m_3\}$. Consequently $\phi^W(\succ'_{W_1}, \succ_{-W_1})(W_1) \in \{m_1, m_2, m_3\}$, since the W-DA algorithm progresses using the same sequence of proposals under both $\succ_{W_1}$ and $\succ'_{W_1}$. Again, by the Rural Hospital Theorem, it implies that W_1 would be matched in the Women-pessimal/Men-optimal stable matching under $\succ'_{W_1}$; that is, $\phi^M(\succ'_{W_1}, \succ_{-W_1})(W_1) \in \{m_1, m_2, m_3\}$

An immediate but important corollary of lemma 2.1 is that the matching that M-DA would produce under $\succ'_{W_1}$ must be the same that the q-quantile mechanism produces under $\succ_{W_1}$. In terms of the example, it means W_1 must be getting m_3 as her assigned partner in M-DA when she reports $\succ'_{W_1}$, which coincides with her assigned partner in the median matching when reporting $\succ_{W_1}$.

Corollary 2.1. $\phi^q(\succ'_{W_1}, \succ_{-W_1}) \succeq_{W_1} \phi^M(\succ'_{W_1}, \succ_{-W_1}) = \phi^q(\succ_{W_1}, \succ_{-W_1})$.

Consequently, given the observed matching, W_1 knows she would have done at least as well by reporting $\succ'_{W_1}$ than by reporting $\succ_{W_1}$. Thus, $\succ'_{W_1}$ satisfies requirement (i) from the definition of regret at μ.

Step 2. It remains to be shown that for some report of others in the inference set, reporting $\succ'_{W_1}$ would generate a matching that is *strictly* preferred by W_1 to μ; i.e. $\succ'_{W_1}$ satisfies condition (ii) from the regret definition at μ. Given corollary 2.1, it is enough to show that $\exists \succ_{-W_1} \in \mathcal{I}(\mu; \succ_{W_1}, \phi^{q=\frac{1}{2}})$ such that $\phi^q(\succ'_{W_1}, \succ_{-W_1}) \succ_{W_1} \phi^M(\succ'_{W_1}, \succ_{-W_1})$.

The structure of the argument is as follows: I show that there is a preference profile in the inference set such that W_1 is matched to each of her acceptable partners in the stable set. Moreover she is matched to a different man in each stable matching. By performing a soft-truncation she forces the mechanism to calculate the quantile with respect to a set that contains only weakly preferred matchings to the one obtained through truth-telling. I show that in fact the quantile over this set selects a partner for W_1 that she strictly prefers.

Consider the preferences:

$$
\begin{array}{ll}
m_1 : W_2 \succ W_3 \succ W_4 \succ W_5 \succ W_1 & W_1 : m_1 \succ m_2 \succ m_3 \succ m_4 \succ m_5 \\
m_2 : W_3 \succ W_4 \succ W_5 \succ W_1 \succ W_2 & W_2 : m_2 \succ m_3 \succ m_4 \succ m_5 \succ m_1 \\
m_3 : W_4 \succ W_5 \succ W_1 \succ W_2 \succ W_3 & W_3 : m_3 \succ m_4 \succ m_5 \succ m_1 \succ m_2 \\
m_4 : W_5 \succ W_1 \succ W_2 \succ W_3 \succ W_4 & W_4 : m_4 \succ m_5 \succ m_1 \succ m_2 \succ m_3 \\
m_5 : W_1 \succ W_2 \succ W_3 \succ W_4 \succ W_5 & W_5 : m_5 \succ m_1 \succ m_2 \succ m_3 \succ m_4
\end{array} \qquad (*)
$$

In this case there exist exactly five stable matchings

$$
\mu_1 = \begin{pmatrix} W_1 & W_2 & W_3 & W_4 & W_5 \\ m_1 & m_2 & m_3 & m_4 & m_5 \end{pmatrix}, \qquad \mu_2 = \begin{pmatrix} W_1 & W_2 & W_3 & W_4 & W_5 \\ m_2 & m_3 & m_4 & m_5 & m_1 \end{pmatrix},
$$

$$
\mu_3 = \begin{pmatrix} W_1 & W_2 & W_3 & W_4 & W_5 \\ m_3 & m_4 & m_5 & m_1 & m_2 \end{pmatrix} = \mu,
$$

$$
\mu_4 = \begin{pmatrix} W_1 & W_2 & W_3 & W_4 & W_5 \\ m_4 & m_5 & m_1 & m_2 & m_3 \end{pmatrix}, \qquad \mu_5 = \begin{pmatrix} W_1 & W_2 & W_3 & W_4 & W_5 \\ m_5 & m_1 & m_2 & m_3 & m_1 \end{pmatrix}.
$$

$|S(\succ_{W_1}, \succ_{-W_1})| = 5$ implies that $\lceil kq \rceil = \lceil 5\frac{1}{2} \rceil = 3$, so the median stable mechanism selects the partner associated with each man/woman's 3-rd best stable matching for this economy. Since all women share the same ranking/preference over the set of stable matchings ($\mu_1 \succ_W \mu_2 \succ_W \mu_3 \succ_W \mu_4 \succ_W \mu_5$) then

$$
\phi^{q=\frac{1}{2}}(\succ_{W_1}, \succ_{-W_1}) = \mu_3 = \mu.
$$

Thus, preference report $(*)$ is consistent with W_1's report, the rules of the median stable mechanism ($q = 1/2$) and the observed matching μ; that is, $\succ_{-W_1} \in \mathcal{I}(\mu; \succ_{W_1}, \phi^{q=\frac{1}{2}})$.

What would have happened if she had reported $\succ'_{W_1}$?,

$$
W'_1 : m_1 \succ' m_2 \succ' m_3 \succ' W_1 \succ' m_4 \succ' m_5.
$$

By lemma 2.1 it follows that $S(\succ'_{W_1}, \succ_{-W_1}) = \{\mu_1, \mu_2, \mu_3\}$, $|S(\succ'_{m_1}, \succ_{m_1})| = 3$ implies $\lceil kq \rceil = \lceil 3\frac{1}{2} \rceil = 2$. The median stable mechanism selects the partner associated with each man/woman's 2-nd best stable matching, therefore

$$
\phi^q(\succ'_{W_1}, \succ_{-W_1}) = \mu_2 \succ_{W_1} \mu_3 = \phi^q(\succ_{W_1}, \succ_{-W_1}).
$$

Then, $\succ'_{W_1}$ satisfies condition (ii) of the definition of regret at μ.

Step 1. and 2. together imply that, given her inference set after reporting truthfully and observing μ, W_1 knows she would have done at least as well by reporting $\succ'_{W_1}$ as by reporting $\succ_{W_1}$; additionally, she knows there exists a preference profile consistent with the observed matching where the report $\succ'_{W_1}$ would have yielded a strictly preferred matching to the one obtain through truth. That is, W_1 regrets $\succ_{W_1}$ at μ through $\succ'_{W_1}$ in the median stable mechanism. The same argument can be extended to quantile stable mechanisms in general, as shown in Appendix A.1. □

The following proposition shows that in the DA truth-telling is the *unique* regret-free report. Thus obtaining truthful reports from all agents is not a consequence of making the behavioral criterion (regret) arbitrarily coarse. If so, not only truth-telling, but also other reports to be regret-free.

Proposition 2.1 (Uniqueness)**.** *Truth is the essentially unique regret-free report in the DA mechanism. Moreover, an agent regrets any other report through truth.*

The qualifier “essentially" refers to the existence of regret-free reports that are essentially equivalent to truth-telling; i.e. those that differ from it only in how they rank the alternatives in the unacceptable set.

Proof. See Appendix A.1. □

To illustrate how an agent may regret misrepresenting their preferences in the DA, I reproduce the example from the introduction in the notation of the marriage market.[14] The example shows that if an agent performs a truncation, and observes an outcome where she is unmatched, then the agent regrets the truncation. Particularly, she regrets truncating by considering the outcomes that would have been generated if she had told the truth.

(Regretting a truncation). Consider a market with two men and two women that runs the M-DA. Suppose W_1's preferences are $W_1 : m_2 \succ m_1 \succ W_1$, but she chooses to report man m_1 as unacceptable; i.e. W_1 performs a truncation. Namely

[14] Dr. Bob would be W_1, and Johns Hopkins would be m_2.

she reports $W_1' : m_2 \succ' W_1 \succ' m_1$.

$$\phi_M(\succ'_{W_1}, \succ'_{-W_1}) = \begin{pmatrix} m_1 & m_2 & \cdot \\ W_2 & \cdot & W_1 \end{pmatrix} \qquad \begin{aligned} m_1 &: W_1 \succ' W_2 \succ' m_1 \\ m_2 &: W_2 \succ' m_2 \succ' W_1 \\ W_2 &: m_1 \succ' m_2 \succ' W_2 \end{aligned}$$

where $\succ'_{-W_1}$ is one of the possible preference profile for agents other than W_1 that are consistent with the observed outcome, the rules of M-DA and what W_1 reported; that is, $\succ'_{-W_1} \in \mathcal{I}$.

Does W_1 regret reporting $\succ'_{W_1}$? Yes. When W_1 finds herself single after performing a truncation she *knows* that, had she told the truth, (i) she could not have done worse since the mechanism is individually rational; $\forall \succ_{-W_1} \in \mathcal{I}$,

$$\phi(\succ)(W_1) \succeq_{W_1} W_1 = \phi_M(\succ'_{W_1}, \succ_{-W_1})(W_1),$$

and (ii) there exists the possibility (namely $\succ'_{-W_1}$) that she could have been matched to m_1 by just reporting truthfully,

$$\phi_M(\succ_{W_1}, \succ'_{-W_1}) = \begin{pmatrix} m_1 & m_2 \\ W_1 & W_2 \end{pmatrix},$$

$$\phi_M(\succ_{W_1}, \succ'_{-W_1})(W_1) = m_1 \succ_{W_1} W_1 = \phi_M(\succ'_{W_1}, \succ'_{-W_1})(W_1).$$

Woman W_1 has conclusive evidence at the end of the game that she would have done better just by reporting truthfully. Consequently she regrets performing the truncation through truth-telling. □

The example shows an agent regretting a truncation. The arguments that address other type of misrepresentations are tackled in Appendix A.1. The basic intuition is that a reversion in the ordering of partners and the option of remaining single, makes the agent to susceptible to being matched to a less preferred alternative. In those situations, there is always an outcome that could have been improved upon by truth-telling. Thus, in order to regret a misrepresentation, the agent only needs to bear in mind truth-telling, and not necessarily some other complex reports.

2.5 Further Remarks

This section briefly discusses how regret-free truth-telling relates to standard notions of incentive compatibility.

The first observation is that if a mechanism is strategy-proof, then it is regret-free truth-telling since if truth is a dominant strategy (in the associated direct revelation game) then it is necessarily undominated ex-post. Yet, the converse is not true. The fact that DA is regret-free truth-telling yet not strategy-proof for agents in the receiving side establishes this.

Regret-free truth-telling does not imply ex-post incentive compatibility. Ex-post incentive compatibility requires truth-telling to be optimal even if the agent were to learn the true state of the world. Regret-free is a weaker concept in this regard. If a mechanism is ex-post incentive compatible, then truth is weakly dominant in every state of the world and therefore regret-free.

The notion of regret and the implied regret-minimization presented in this paper does not coincide with worst-case minimization *à la* von Neumann-Morgenstern. Under worst-case minimization, the agent expects to remain single irrespective of her report in a stable matching. Therefore, she is indifferent among all reports. In contrast, truth-telling is the unique regret-free report in the case of DA.

The definition of regret presented in this paper is by no means the only possible one. The literature on regret is too vast to survey. Instead I address the most prominent notion, that of minimizing maximum regret in the sense of Stoye (2011). Under this notion an individual computes her regret as the difference in outcomes between what would have been the optimal action at the true state of the world and the action she takes (see Bell, 1982) This requires a knowledge of the true preference profile at the end of the game which is absent in the setup analyzed in this paper, since in most applications privacy concerns prevent this sort of information revelation. Moreover, it requires a cardinal computation, both absent in this paper's setup, as well as in the data that is elicited from participants. The report that would minimize an agent's maximum regret in the sense of Stoye (2011) does not generally coincide with a regret-free report for that agent.

The question of existence and uniqueness of a regret-free report are beyond the scope of the present paper, and are left for future research. However, a few remarks are in order.

Remark 2.1. There exists a mechanism ϕ such that no agent has a regret-free report, namely the Boston Mechanism. In contrast, every report is regret-free in a constant mechanism.

Remark 2.2. $\phi(\cdot)$ being stable does not imply, nor is it implied by, truth being

regret-free in $\phi(\cdot)$.

Remark 2.3. Regret is not transitive, i regretting $\succ_i'$ through $\succ_i''$ and regretting $\succ_i''$ through $\succ_i'''$ does not imply i regrets $\succ_i'$ through $\succ_i'''$.

Proofs. See Appendix A.2. □

Remark 2.1 shows that neither existence nor uniqueness are guaranteed in general. Remark 2.2 says that a regret-free truth-telling mechanism may not be stable. Lastly, remark 2.3 highlights that this notion of regret is not transitive since regret may be happening at different matches. Intransitivity is not surprising since Bikhchandani and Segal (2011) showed that intransitivity is built into any definition of regret-based behavior that is not expected utility.

2.6 Conclusion

In this paper I present the notion of regret-free truth-telling and through it show that the most salient matching mechanism, the deferred acceptance, provides incentives to agents on *both* sides of the market to report their preferences truthfully. Moreover, the incentives provided by the DA are strict: if an agent wishes to avoid regret, she has a unique choice, to report her preferences honestly. The main result is that DA is the *unique* quantile stable matching mechanisms to provide such incentives. The result helps to understand the success *and* salience of the DA in practice, despite being manipulable.

2.7 Related Literature

This paper contributes to understanding the incentives DA provides for truth-telling. The main result in the literature in this regard is negative, Roth (1982) and Dubins and Freedman (1981) show that there exists no stable strategy-proof mechanism. Additionally, in recent contributions Chen, Egesdal, Pycia, and Yenmez (2016) and Pathak and Sönmez (2013) show that ranking stable mechanisms by their manipulability is equivalent to ranking them by preferences. The important implication of their analysis is that stable mechanisms cannot be ranked by their manipulability *for all agents*. Based on these observations it would seem we are left with multiple stable mechanisms and no clear way of choosing among them in terms of their incentives for truthful reporting. This identifies incentives that DA provides agents to report truthfully that are not provided by any other quantile stable mechanism. These incentives are a relaxation of the concepts of strategy-proofness and ex-post incentive compatibility.

Trying to understand the incentives in the context of incomplete information, Roth and Rothblum (1999) and Ehlers (2008) look at the problem of incentives in DA from the Bayesian point of view, where expected utility maximizing participants have prior beliefs over each other's preferences. They find that truncation strategies first order stochastically dominate other false reports under some symmetry conditions on the priors. However, there is no clear order between truncation strategies and truth-telling since it depends on the agent's attitude towards risk.

The most recent and fruitful approach has focused on the incentives agents face in large markets; see Roth and Peranson (1999), Immorlica and Mahdian (2005), Kojima and Pathak (2009) and Lee (2017). The main message is that as the market grows large, the core of the market shrinks (in an appropriate sense) and so do the gains of deviating from truth-telling. Lee (2017) distinguishes itself from the rest by not having to rely on limited acceptability assumptions. His analysis hinges on both large markets and on agents having (random) cardinal utilities. In general this approach does not offer unequivocal advice for participants of small markets and requires participants to be able to compare cardinally the possible matches which is a departure from the traditional ordinal approach in matching literature. In contrast the treatment in this paper maintains the ordinal preference approach tradition in matching and looks for incentives in the *spirit* of the "Wilson doctrine"[15] (Wilson, 1987) and robust mechanism (Ledyard, 1977; Bergemann and Morris, 2008) to search for mechanisms whose properties do not critically depend on the common knowledge of agents about distributions of beliefs, types, etc.

The closest papers are Kojima and Manea (2010) and Barberà and Dutta (1995).[16] Kojima and Manea (2010) also provide a characterization in the case of two-sided matching as the unique stable weakly Maskin monotonic mechanism. The basic interpretation of their axiom is that when an agent makes fewer claims over partners on the other side of the market, his side of the market must be made weakly better off. Surprisingly, weak Maskin monotonicity is defined using only the primitives of one side of the market. They do not approach the issue of incentives for truthful reporting.[17] In contrast I characterize DA through stability and regret-free truth-

[15] "Game theory has a great advantage in explicitly analyzing the consequences of trading rules that presumably are really common knowledge; it is deficient to the extent it assumes other features to be common knowledge, such as one agent's probability assessment about another's preferences or information. I foresee the progress of game theory as depending on successive reductions in the base of common knowledge required to conduct useful analyses of practical problems. Only by repeated weakening of common knowledge assumptions will the theory approximate reality." —Wilson (1987).

[16] I am indebted to Laura Doval for bringing Barberà and Dutta (1995) to my attention.

[17] Other known characterizations are Gale and Shapley (1962), in their seminal paper, characterize

telling in the class of quantile stable mechanism.

Barberà and Dutta (1995) introduce the notion of protective strategies and show that truth-telling is the unique protective strategy equilibrium in the revelation game associated with the deferred acceptance. The concept of protective strategies is a refinement of worst-case minimization. Two strategies that have the same worst outcome are distinguished by their 2nd-worst outcome; if the latter coincides, then they are that distinguished by the 3rd-worst outcome, etc. If one assumes that the agents observe only their own matching partner, and nothing else, then the regret-free criterion would coincide with protective strategies. However, in general these two criteria are distinct. For example, in an environment under complete information, truth-telling is a protective strategy for any participant of the DA, but it is not regret-free.

The family of quantile stable mechanisms is discussed in Teo and Sethuraman (1998), Klaus and Klijn (2006) and Chen, Egesdal, Pycia, and Yenmez (2014) among others. Teo and Sethuraman (1998) take a polyhedral approach to stable marriage problem exploiting linear programming techniques. They show that every fractional stable matching as a convex combination of solutions to the stable marriage problem. Using this technique they prove existence of the quantile stable mechanisms. Klaus and Klijn (2006) provides a simple proof of the existence through purely lattice-theoretical arguments. Chen, Egesdal, Pycia, and Yenmez (2014) take on quantile stable mechanism in a context of matching with contracts. They show that if preference satisfy strong substitutability condition and the law of aggregate demand then quantile stable matchings exist and are well defined. This paper builds upon their work analyzing regret based incentives and use these notions to characterize DA.

DA as constrained efficiency subject to stability. Balinski and Sönmez (1999) characterize DA as the unique stable mechanism that respects improvements. Alcalde and Barberà (1994) characterize DA by stability and strategy-proofness under the domain restriction of preferences satisfying top-dominance.

Chapter 3

CENTRALIZED MATCHING WITH INCOMPLETE INFORMATION

3.1 Introduction

Overview

The large literature on two-sided matching has been successful in its application to various markets, ranging from the matching of newly-minted doctors and residency positions, to the matching of kids to schools, to the matching of medical patients and organ donors. Most of this literature assumes complete information of preferences in the market: market participants are assumed to be perfectly informed of all other participants' preferences, in addition to their own. This assumption allowed for important theoretical advances. However, in applications, it is by no means innocuous. Many matching markets are large and therefore imply limited communication between participants. For instance, the National Resident Matching Program (NRMP) involves around $70,000$ participants annually, the number of kids and schools in big cities such as NYC and Boston is in the hundreds of thousands, etc. The goal of the current paper is to characterize some of the potential consequences of incomplete information in centralized matching markets.

One of the most commonly used clearinghouses for two-sided matching markets is the celebrated deferred acceptance algorithm, DA for short, first introduced by Gale and Shapley (1962). In practice, market participants report their preferences simultaneously to the clearinghouse, which emulates the Gale and Shapley (1962) algorithm. This mechanism has several desirable features (see Roth and Sotomayor, 1990). The resulting matching is stable with respect to reported preferences and therefore Pareto optimal. Furthermore, the generated stable matching is the most preferred, with respect to the reported preferences, for the "proposing" side (doctors in the NRMP; kids in centralized school choice system). It is known, however, that the DA matching mechanism is generally not incentive compatible: whenever the market has more than one stable matching, there are incentives to misreport preferences. Nonetheless, with complete information, and under mild assumptions on behavior, the set of equilibrium outcomes coincides with the set of stable matchings. Therefore, strategic behavior is still expected to generate outcomes retaining

the desirable features of stability. Moreover, the Rural Hospital Theorem (McVitie and Wilson, 1970) guarantees that the set of unmatched individuals is the same across stable matchings. Therefore, with complete information, the selection of equilibrium does not impact the set of matched participants.

In this paper, we ask whether these insights carry over to the case in which information is incomplete. Under what conditions would we still expect equilibrium outcomes of DA to be stable with respect to the underlying preferences? Would the proposing side be expected to maintain an advantage? Would selection of equilibria remain an ineffective instrument for inducing different sets of matched participants?

In our model, firms and workers—that could be a metaphor for doctors and hospitals, kids and schools, etc.—match using the firm-proposing DA mechanism. At the outset, preferences over match partners are realized according to a commonly-known distribution. We consider a setting in which deviations from the complete information environment are "small." In particular, we assume that firms, who act as proposers, know the realized preferences in the market. Each worker, however, knows only his own realized match preferences over others.

In the complete information world, markets with a unique stable matching do not exhibit incentive compatibility issues. Furthermore, a significant body of work tackling incentive compatibility issues has focused on environments in which small cores emerge as markets grow large, ones where incomplete information is indeed more likely to be important (see Immorlica and Mahdian, 2005; Kojima and Pathak, 2009; Ashlagi, Kanoria, and Leshno, 2017). We therefore assume that for any realized preference profile, the market entails a unique stable matching. Certainly, with incomplete information on both sides, or markets that entail multiple stable matchings, some of the negative results we illustrate are easier to achieve.[1] In that respect, our assumptions on preferences are designed to stack the cards in favor of stable outcomes.

In this setting, our first result characterizes a rich class of environments in which equilibrium outcomes do not correspond to stable matchings with respect to the underlying preferences. Such environments can emerge even when the (complete-information) stable matching is the same across possible preference realizations; in other words, when there is no uncertainty over what is the stable matching. Our

[1] Roth (1989) suggested one example with two-sided incomplete information and multiple stable matchings for realized markets that yields an equilibrium that is unstable in each realized market. See our discussion in the literature review below.

characterization result is a “one and a half cycle" theorem.

In rough terms, the result identifies two conditions on a preference profile such that there is an incomplete-information economy with that profile in its support and a Bayesian Nash equilibrium that does not coincide with the stable matching for each realized preference profile. The first condition is the existence of a preference cycle. Namely, there is a sub-market in which the induced preferences exhibit multiple stable matchings. Due to our assumption of a unique stable matching in the market as a whole, there must be a worker who is a “spoiler" and blocks one of the matchings in the sub-market with a firm participating in the cycle. The second condition is that this spoiler takes part in a “half cycle." That is, workers' preferences are such that it is possible to construct (alternative) firms' preferences to create a cycle involving the spoiler. The idea of the proof is then to consider an additional state in which that cycle exists. As it turns out, such a construction allows workers to “trade" match partners across possible preference realizations. Intuitively, the spoiler in one market drops his claim for the firm he prefers under one preference realization to gain a firm he prefers in another.

Equilibria of these sort may be desirable for the receiving side, i.e. the workers. In fact, we show settings in which all workers prefer to be on the receiving side of the DA mechanism and select an “unstable" equilibrium. This stands in contrast with the common wisdom held for the complete information case, suggesting that the proposing side is advantaged. Indeed, this common wisdom was an important argument for the switch in the late 1990s from the hospital-proposing to the doctor-proposing version of DA in the NRMP (see Roth and Peranson, 1999). It is also the central reason for having children, or their parents, propose in most applications of DA to school choice. Our results suggest that market designers should carefully inspect the information and preferences market participants hold before deeming which segment of the market should be on the proposing side of the DA mechanism.

Our second result characterizes settings in which, with incomplete information, equilibria may be associated with different distributions over the set of unmatched individuals. Namely, we show that whenever we start with an economy in which there is an equilibrium outcome that is not (complete-information) stable for some preference realizations, one can add one worker and one firm to generate an economy with different sets of matched individuals across equilibria. In such settings, selection could potentially constitute an instrument for generating more matches of particular sets of participants, affecting e.g. the number of filled residency positions

in rural hospitals, or the number of kids of certain demographics who get matched.

It is important to contrast our results with some of the existing results tackling issues of incentive compatibility in the DA mechanism. As alluded to above, the literature has taken three approaches, all focusing on large markets. The first approach argues empirically (Roth and Peranson, 1999) and theoretically (Immorlica and Mahdian, 2005; Kojima and Pathak, 2009) that in large markets cores are small. In other words, as markets grow large, with high likelihood there is a unique stable matching and, therefore, no incentive compatibility issues. In the same spirit, the second approach argues that when markets are imbalanced, cores shrink rapidly (Ashlagi, Kanoria, and Leshno, 2017). Again, with imbalanced markets, under some conditions, we expect small cores and incentive compatibility issues disappear. The third approach argues that in large markets, the incentives to misreport vanish, when everyone else is reporting truthfully (Lee, 2017). In comparison, our results illustrate that even when cores are small, with a unique stable matching for each realization of preferences, incomplete information may imply strategic misreporting. In other words, to the extent that incomplete information may be important in some of the matching applications, small cores are no theoretical panacea. Empirically, even if we expect cores to be small for any realization of preferences, we cannot assume that participants are truthfully reporting their preferences.

Our last results are more positive in nature. We identify classes of preferences in which incomplete information has no impact. Namely, whenever preferences are assortative as in Becker (1973), on either side—either firms share ordinal rankings over workers in each preference realization, or workers share ordinal rankings over firms (that are the same across preference realizations)—there is a unique equilibrium of the induced game with incomplete information that realizes the unique stable matching for any realized preferences. This result holds regardless of the underlying assortative preferences that occur with positive probability in the economy, the match utilities that represent them, or the precise probability distribution that governs which are more likely to arise.

Related Literature

Roth and Sotomayor (1990) offers a review of the seminal work on matching markets with complete information. In the years since, several papers considered the potential impacts of incomplete information in both decentralized and centralized settings.

In the decentralized, non-cooperative, setting, there has been a long-standing quest

for a natural notion of stability with incomplete information. Liu, Mailath, Postlewaite, and Samuelson (2014) offer such a notion for matching markets with one-sided incomplete information, similar to ours, and transfers. Bikhchandani (2017) also considers markets with one-sided incomplete information and offers a stability notion that does not entail transfers.[2] Several papers allow for incomplete information when modeling decentralized interactions as a non-cooperative dynamic game (see Niederle and Yariv, 2009, and references therein). Similar to our setting, these papers often look for a characterization of economies in which equilibrium results in the complete-information stable matching in all realized markets.

With respect to centralized settings, Roth (1989) is possibly the closest to the current paper. He offers an example of a market with incomplete information on both sides in which there exists no stable mechanism implementing the (complete-information) stable matching for each realization of preferences. Importantly, in the example, preference profiles that can conceivably be realized are associated with multiple stable matchings. The message that incomplete information may upset some of the basic conclusions derived with complete information is common to both this paper and ours. Our contribution relative to Roth (1989) is fourfold. First, we provide a general characterization of when stable mechanisms fail in guaranteeing stable outcomes in equilibrium when information is incomplete. Second, we illustrate that the impacts of incomplete information can be severe even when incomplete information is only one-sided, there is no uncertainty on the stable matching itself, and there is a unique stable matching for each realized preference profile. This last item is important as multiplicity of stable matchings is associated with incentive compatibility issues even in the complete information setting. Third, we illustrate other results that can be overturned when incomplete information is present: the optimality of truncation strategies, the usefulness of equilibrium selection for affecting the set of unmatched participants, etc. Last, we also characterize a class of economies in which incomplete information does not impact centralized outcomes.[3]

[2] Chakraborty, Citanna, and Ostrovsky (2010) study a many-to-one school choice setting with one-sided incomplete information. They suggest that stability should be defined in conjunction with the mechanism the generates the matching in place as different mechanisms allow for different learning patterns from outcomes. The characterize settings in which there exist mechanisms allowing for an ex-post notion of stability.

[3] Roth and Rothblum (1999) consider economies where information is "symmetric." Roughly, for any two firms f and f', the likelihood of every profile of preferences for workers is the same as the likelihood of that profile where f and f' are swapped in all rankings. Under this symmetry assumption, there are always optimal truncation strategies in the firm-proposing DA. We show that

Ehlers and Massó (2007) identify a link between small cores and the existence of ordinal Bayesian Nash equilibria implementing the (complete-information) stable outcome for any realized market when information is incomplete. In our setting, we assume cores are small and, indeed, (complete-information) stable outcomes can always be implemented through a Bayesian Nash equilibrium. In contrast, our focus is on cases that give rise to other equilibrium outcomes.[4]

A growing empirical literature estimates preferences in matching markets using constraints implied by stability (see review in Chiappori and Salanié, 2016; Agarwal, 2015; Hsieh, 2011). Our results provide caution regarding the validity stability constraints when analyzing outcomes of centralized settings. Even when markets are such that one expects the set of stable matchings to be small, equilibrium outcomes may not be stable. The assumption that outcomes are stable is then an equilibrium-selection assumption. It is also important to note that empirical studies of matching markets often consider rather few, if not one, instantiation of a market. There is then a form of omitted variable as markets that could have been realized but did not are not observed, yet could have an important impact on equilibrium play.

3.2 Motivating Example

Small Cores and Multiple Outcomes

Consider a market with three firms: f_1, f_2, and f_3 and three workers: w_1, w_2, w_3. There are two states of the world that determine preferences: θ_1 and θ_2, which are equally likely. For $i = 1, 2$, the preferences in state θ_i are given by $U(\theta_i)$ as follows:

$$U(\theta_1) = \begin{pmatrix} \mathbf{3,2} & \underline{1,4} & 2,2 \\ \underline{2,4} & \mathbf{3,2} & 1,4 \\ 1,1 & 2,1 & \underline{\mathbf{3,1}} \end{pmatrix} \qquad U(\theta_2) = \begin{pmatrix} \mathbf{3,2} & \underline{1,4} & 2,2 \\ 1,4 & \mathbf{3,2} & \underline{2,4} \\ \underline{1,1} & 2,1 & \mathbf{3,1} \end{pmatrix}$$

Figure 3.1: Payoff matrix by state.

where $U_{jk}(\theta_i) = (U^f_{jk}(\theta_i), U^w_{jk}(\theta_i))$ describing the payoff $U^f_{jk}(\theta_i)$ to firm f_j from matching with worker w_k in state θ_i, and the payoff $U^w_{jk}(\theta_i)$ to worker w_k from matching with firm f_j in state θ_i. That is, in the matrix notation above, each row corresponds to one of the three firms and each column corresponds to one of the

this symmetry assumption is, in fact, quite stringent and illustrate cases in which truncation strategies are sub-optimal.

[4] Ehlers and Massó (2015) focus on many-to-one matching markets with incomplete information and provide restrictions on strategies to constitute ordinal Bayesian Nash equilibria.

three workers. An entry then captures the payoffs of the corresponding firm and worker pair. We assume that remaining unmatched generates a payoff of 0 for any market participant.

The "centralized matching game" we consider proceeds as follows. At the outset, the state is determined. Firms learn the state, while workers do not. Workers only know their own vector of match payoffs, which does not vary across states, and the (uniform) prior over states. Then, both firms and workers simultaneously submit preference rankings to a firm-proposing DA mechanism that generates a matching.

Notice that in both states, there is a unique complete-information stable matching highlighted in bold in the payoff matrices above. That is, if all participants knew the state to be θ_i, the resulting unique stable matching and, unique equilibrium outcome of the DA mechanism (see Roth and Sotomayor, 1990), would be the matching μ, $\mu(f_j) = w_j$, $j = 1, 2, 3$. In particular, in our incomplete information setting, there is no uncertainty on the (complete information) stable matching itself. In fact, the only difference between the two states appears in the preferences of f_2—she ranks w_1 and w_3 differently in the two states.

There are several results that carry over from the complete information setting, the details of which will be illustrated in greater generality below. First, truthful reporting is still weakly dominant for the firms, for much the same reasons it is in the complete-information setting. In what follows, we will therefore assume firms report truthfully and focus on the induced game between workers. Second, the complete-information stable matching in each state is an equilibrium outcome of this game. Indeed, workers reporting their preferences truthfully constitutes an equilibrium. However, we will now show that it is not the unique equilibrium outcome. In fact, we will now show that there is another equilibrium generating an outcome that is not (complete-information) stable in both states.

Specifically, we will show that the matchings corresponding to the underlined entries in the matrices above can be implemented in equilibrium. That is, we will show that the matching $\lambda(\theta_i)$ in state $\theta_i, i = 1, 2$, given by:

$$\begin{array}{lcl} \lambda(\theta_1)(f_1) = w_2 & & \lambda(\theta_2)(f_1) = w_2 \\ \lambda(\theta_1)(f_2) = w_1 & \text{and} & \lambda(\theta_2)(f_2) = w_3 \\ \lambda(\theta_1)(f_3) = w_3 & & \lambda(\theta_2)(f_3) = w_1 \end{array}$$

is an equilibrium outcome.

Indeed, consider the following profile of strategies for workers:

$$\begin{aligned} w_1: &\quad f_2 \succ f_3 \succ w_1 \\ w_2: &\quad f_1 \succ f_2 \succ f_3 \\ w_3: &\quad f_2 \succ f_3 \succ w_3. \end{aligned}$$

That is, w_2 reports his preferences truthfully, while w_1 and w_3 *drop* f_1 from their preference list and declare her unacceptable. Notice that these strategies are not weakly dominated.

It is straightforward to check that these reports yield $\lambda(\theta_i)$ for $i = 1, 2$. Why are they a best response? Consider w_1, who receives an expected payoff of $0.5 * 4 + 0.5 * 1 = 2.5$ in this candidate equilibrium. There are two natural deviations to consider: truthful reporting and truncation. Given others' reports, if w_1 reports his preferences truthfully, the unique (complete-information) stable matching would be implemented in each state and he would expect a payoff of $2 < 2.5$. Suppose w_1 truncates his preferences and demands only his favorite firm, f_2. In this case, he would still be matched with f_2 in state θ_1, but remain unmatched in state θ_2, yielding an expected payoff of $0.5 * 4 + 0.5 * 0 = 2 < 2.5$. Other deviations are not profitable more directly and similar arguments follow for the other two workers. Thus, the profile of strategies above is indeed an equilibrium, generating $(\lambda(\theta_1), \lambda(\theta_2))$ as an outcome.

There are several notable features of this equilibrium. First, workers uniformly prefer this equilibrium to the one generating the (complete-information) stable outcome in each state. Furthermore, they might prefer being on the receiving side of DA and selecting this equilibrium over being on the proposing side of DA. Indeed, when on the proposing side, truth-telling would be weakly dominant for workers. If they use truth-telling, since firms are informed of the state, the market in each state operates as if all participants had complete information and there is a unique equilibrium generating the (complete-information) stable matching in each state.

Second, there are only two pure-strategy equilibria corresponding to our incomplete-information market "game," and neither is fragile to the precise specification of payoffs. Indeed, a small perturbation of payoffs would not affect this result, and the emergence of the additional equilibrium identified above is not a knife-edge result.

Third, the equilibrium involves *dropping* of firms from one's rankings and truncation strategies are not best responses. This is in contrast with the results from the complete information setting that illustrate that there are always truncation best responses,

regardless of the profile of strategies used by others (see Roth and Vande Vate, 1991). As shown above, in this example, truncation is not a best response for workers.

What is driving this example? Consider state θ_1. Were f_3 and w_3 out of the market, the remaining market would have two stable matching: the firm-optimal one $\mu, \mu(f_j) = w_j$ for $j = 1, 2$, and the worker-optimal one $\mu', \mu'(f_j) = w_{3-j}$ for $j = 1, 2$. When firms report truthfully in the sub-market without f_3 and w_3, w_1 could guarantee μ' by "truncating" his preferences restricted to that sub-market and declaring only f_2 as acceptable. However, in the full market, w_3 and f_1 would block any such attempt at a swap of partners by w_1 and w_2. In our terminology, w_3 is a "spoiler." However, the existence of state θ_2 helps incentivize w_3 to drop his claim for f_1. Indeed, in state θ_2, without w_1 and f_3, the remaining market has two stable matchings and w_3 would benefit by "truncating" his preferences restricted to that sub-market and, indeed, declaring f_1 as unacceptable. In fact, payoffs are such that this "truncation" is beneficial despite it meaning the foregone possible match with f_1 in θ_1. Now, in state θ_2, w_1 is a "spoiler" and, with complete information, would block the matching with f_1. However, the existence of state θ_1 induces w_1 to drop his claim for f_1 as payoffs make it useful to do so when considering both states possible ex-ante.

Theorem 3.1 below offers a generalization of this example and provides a characterization of markets that entail equilibria yielding outcomes that do not coincide with the (complete-information) stable matching in each state. As in the example, dropping strategies, rather than truncation strategies, play an important role in such markets.

Augmented Example and The Rural Hospital Theorem

With complete information, the set of stable matchings has two extrema—the firm-optimal stable matching, which is the worker-pessimal stable matching, and the worker-optimal stable matching, which is firm-pessimal stable matching. The existence of an extremal stable matching generates The Rural Hospital Theorem (McVitie and Wilson, 1970). That important theorem states that the set of unmatched individuals is constant across stable matchings. Under the assumption that firms use their weakly dominant strategy of truth-telling, the set of equilibrium outcomes of the firm-proposing DA coincides with the set of stable matchings. The Rural Hospital Theorem then suggests that selection of equilibria cannot be useful for

affecting the set of unmatched individuals.

As it turns out, this is no longer the case when information is incomplete. As an example, consider an augmentation of our basic example above, where an additional firm, f_4, and additional worker, w_4, are included. The market operates as before. At the outset, one of two states, θ_1 or θ_2 is determined with equal probabilities. The match payoffs in each state are given using the same notation as before as follows:

$$U(\theta_1) = \begin{pmatrix} \mathbf{3,2} & \underline{1,4} & 2,2 & \varnothing,\varnothing \\ \underline{2,4} & \mathbf{3,2} & 1,4 & \varnothing,\varnothing \\ 1,0 & 2,1 & \underline{\mathbf{3,1}} & \varnothing,\varnothing \\ \varnothing,1 & \varnothing,\varnothing & \varnothing,\varnothing & \underline{\mathbf{1,1}} \end{pmatrix} \quad U(\theta_2) = \begin{pmatrix} \mathbf{3,2} & \underline{1,4} & 2,2 & \varnothing,\varnothing \\ 1,4 & \mathbf{3,2} & \underline{2,4} & \varnothing,\varnothing \\ 1,0 & 2,1 & \mathbf{3,1} & \varnothing,\varnothing \\ \underline{2,1} & \varnothing,\varnothing & \varnothing,\varnothing & \mathbf{1,1} \end{pmatrix}$$

Figure 3.2: Payoff matrix by state.

where $\varnothing$ stands for the corresponding partner being unacceptable (and can be thought of as a large and negative payoff). Remaining unmatched still generates a payoff of 0 for any participant.

In this case, there is still a unique (complete-information) stable matching that is identical across the two states and corresponds to the diagonal highlighted in bold in the matrices above. A similar construction to that described above generated the underlined entries in each state as an equilibrium outcome. In this example, there is no "extremal" equilibrium. Indeed, f_3 and w_4 prefer the equilibrium in which the (complete-information) stable is implemented in each state. However, other market participants prefer the other pure-strategy equilibrium. In line with this observation, the set of matched individuals differs across equilibrium outcomes, with one outcome entailing all agents being matched and another having f_3 and w_4 unmatched in state θ_2.

Theorem 3.2 below provides a generalization of this augmented example and illustrates a class of markets in which equilibrium outcomes vary in terms of the set of unmatched individuals they generate.

3.3 General Setup

The Environment

A *matching market* is a triplet $(\mathcal{F}, \mathcal{W}, U)$ composed of a finite set of firms $\mathcal{F} = \{f_1, ..., f_n\}$, a finite set of workers $\mathcal{W} = \{w_1, ..., w_m\}$, and match utilities $U = \left(\left\{u_{ij}^f\right\}, \left\{u_{ij}^w\right\}\right)$.

For each $i = 1, ..., n$ and $j = 1, ..., m$, u_{ij}^f is firm f_i's utility from matching with worker w_j and u_{ij}^w is worker w_j's utility from matching with firm f_i. $u_{i\varnothing}^f$ stands for the utility of firm f_i from remaining unmatched and, respectively, $u_{\varnothing j}^w$ stands for the utility of worker w_j from remaining unmatched. We assume that all preferences are strict.[5] We say a worker w_j is *acceptable* to firm f_i whenever $u_{ij}^f > u_{i\varnothing}^f$. Similarly, we say a firm f_i is *acceptable* whenever $u_{ij}^w > u_{\varnothing j}^w$.

Match utilities U induce a profile of ordinal preferences of firms over workers and workers over firms, which we will denote by $\succ = \left(\{\succ^{f_i}\}, \{\succ^{w_j}\}\right)$. That is, $\succ^{f_i}$ is the ordinal ranking of workers, as well as staying unmatched, that $u_{i\cdot}^f$ represents; $\succ^{w_j}$ is defined analogously.

An *economy* is a quintuple $(\mathcal{F}, \mathcal{W}, \{U(\theta)\}_{\theta \in \Theta}, \Theta, \Psi)$, where Θ is a finite set of states with each $\theta \in \Theta$ corresponding to a different market $(\mathcal{F}, \mathcal{W}, U(\theta))$ with the same set of firms $\mathcal{F}$ and workers $\mathcal{W}$ and Ψ is a probability distribution over states. Without loss of generality, we assume Ψ has full support on Θ.

We consider a *centralized matching economy game* in which, at the outset, a state $\theta \in \Theta$ is selected according to Ψ. Firms are all informed of the realized state θ, while each worker w_j is privately informed only of her preferences, $u_{ij}^w(\theta)$ for $i = 1, ..., n, \varnothing$. Workers and firms then participate in a firm-proposing DA matching mechanism. In particular, they simultaneously submit ordinal rankings that are translated into a matching using the firm-proposing DA algorithm. The structure of the game is common knowledge among all participants. In particular, workers can update on the distribution over states, and therefore the realized preferences, after observing their private information.

Certainly, if workers' utilities are all different across states, workers' private information would allow them to distinguish between states and the analysis of the game would follow that of the complete information variant, state by state. We will assume that workers' private information is not revealing of the state. In other words, we will assume that for any $\theta', \theta'' \in \Theta$, $u_{ij} \equiv u_{ij}^w(\theta') = u_{ij}^w(\theta'')$. This implies that workers are symmetrically informed.[6]

Much as in the complete information case, truthful reporting is weakly dominant for firms.[7] We will therefore assume throughout that firms report truthfully their

[5] That is there exists no i, i' and j, j' such that $u_{ij}^f = u_{ij'}^f, u_{ij}^f = u_{i\varnothing}^f, u_{ij}^w = u_{i'j}^w$, or $u_{ij}^w = u_{\varnothing j}^w$.

[6] This assumption is similar to that made by Liu, Mailath, Postlewaite, and Samuelson (2014), who offer a stability notion for incomplete-information matching markets.

[7] The proof follows the same lines as the proof for the complete information environment, see

preferences and focus on the induced game between workers. We will consider the Bayesian Nash equilibria of this game.

It is important to note the impact of the proposing side in our economy. Suppose workers were the proposing side in the DA matching mechanism, and using their weakly dominant strategy of truthful reporting. In that case, in each state, workers would behave just as they would were they fully informed and using their weakly dominant strategy of truthful reporting. The resulting set of equilibrium outcomes would then automatically correspond to the set of equilibrium outcomes emerging when information is complete, namely the set of stable matchings in each state (Roth, 1984; Gale and Sotomayor, 1985). This is why, to inspect the impact of incomplete information on matching clearinghouses, we concentrate on the case in which the uninformed side of the market is on the receiving side.

As mentioned in the Introduction, in the complete information benchmark, there are no incentive compatibility issues when there is a unique stable matching. In order to isolate the impacts of incomplete information on strategic behavior, we therefore assume that each realized market corresponding to the support of the distribution Ψ has a unique (complete-information) stable matching. Formally, we assume that for each $\theta \in \Theta$, the matching market $(\mathcal{F}, \mathcal{W}, U(\theta))$ has a unique stable matching, denoted $\mu(\theta)$. In what follows, we will drop the allusion to complete information whenever referring to the unique (complete information) stable matching in any state.

Basic Definitions

In now define several notions that will be useful for our characterization. These definitions identify features of sub-markets in particular states.

Definition 3.1 (Cycles)**.** *In a given state θ, a set of firms $G = \{g_1, \ldots, g_K\} \subset F$ and a set of workers $V = \{v_1, \ldots, v_K\} \subset W$ form a cycle if the following conditions hold for each $k \in \{1, \ldots, K\}$:*[8]

(1) $\mu(\theta; v_k) = g_k$*;*

(2) $g_{k+1} \succ_{v_k} g_k \succ_{v_k} \varnothing$*;*

(3) $v_k \succ_{g_k} v_{k-1} \succ_{g_k} \varnothing$.

Dubins and Freedman (1981).

[8] All indices are modulo K wherever necessary.

Condition (1) requires that the firms and workers involved in the cycle are stable partners in state θ. Conditions (2) and (3) imply that workers would like to trade their stable partners among themselves, but firms would prefer to stick with the stable allocation of partners.

The existence of a cycle involving firms G and workers V implies that the sub-market involving these same firms and workers in state θ, with preferences induced by the original preferences, has multiple (complete-information) stable matchings.[9]

Graphically, consider the addition of arrows to a matching payoff matrix $(u^f_{ij}, u^w_{ij})_{i\neq\varnothing, j\neq\varnothing}$ as follows, assuming there is a unique stable matching in the corresponding market. We draw a horizontal arrow originating from entry (i, j) and pointing at entry (i, k) if and only if $u^f_{ik} > u^f_{ij}$. Similarly, we draw a vertical arrow originating from entry (i, j) and pointing at entry (k, j) if and only if $u^w_{kj} > u^w_{ij}$. The existence of a cycle defined as above corresponds to a graphical cycle in the matrix involving firms and workers who are matched under the unique stable matching.

For example, consider $U(\theta_1)$ in section 3.2. Firms $\{f_1, f_2\}$ and workers $\{w_1, w_2\}$ form a cycle. Figure 3.3(a) corresponds to the visual cycle involving $\{f_1, f_2\}$ and $\{w_1, w_2\}$.

$$\begin{pmatrix} 3,2 \leftarrow 1,4 & 2,2 \\ \downarrow \quad \uparrow & \\ 2,4 \rightarrow 3,2 & 1,4 \\ & \\ 1,1 \quad 2,1 & 3,1 \end{pmatrix} \qquad \begin{pmatrix} 3,2 & 1,4 & 2,2 \\ & \uparrow \quad \downarrow & \\ 1,4 & 3,2 & 2,4 \\ & & \\ 1,1 & 2,1 & 3,1 \end{pmatrix}$$

(a) (b)

Figure 3.3: One and a half cycle

In principle, we could have defined cycles without restriction condition (1) in our definition. That is, one could envision cycles involving firms and workers who need not be matched under the stable matching. As it turns out, this condition will be useful in our characterization. Intuitively, a cycle among firms and workers that are

[9] $\mu(\theta)$ is stable in this sub-market. To find an additional stable matching, construct the following. First, all workers in G point to their favorite firms in F. Whenever a worker points to a firm no one has pointed to, that worker and firm are matched. Whenever more than one worker points to the same firm, match that firm to its favorite worker among those. Proceed recursively with the remaining unmatched workers and firms. The resulting matching is stable, and does not coincide with $\mu(\theta)$ restricted to the sub-market.

all dominated choices for one another should not impact strategic incentives much. However, a cycle among stable match partners implies viable incentives to generate unstable matches for the workers.

Suppose there is a cycle in state θ involving firms $G = \{g_1, ..., g_K\}$ and workers $V = \{v_1, ..., v_k\}$ with the notation of Definition 3.1 above. Consider the matching $\mu'(\theta; v_k) = g_{k+1}$ for $k = 1, ..., K$ (where, again, we interpret $g_{K+1} = g_1$) and $\mu'(\theta; v) = \mu(\theta; v)$ for any $v \in \mathcal{W} \backslash V$. That is, workers in V implement the desired swap of firms that defined the cycle. We assumed there is a unique stable matching in each state θ. Therefore, $\mu'(\theta)$ must be blocked. In fact there must be a worker w that does not benefit from this trade, $\bar{w} \in \mathcal{W} \backslash V$ and a firm that suffers from the trade, $g = g_{\bar{k}} \in G$ that block $\mu'(\theta)$. Any such worker $\bar{w}$ "spoils" the potential matching $\mu'(\theta)$ for workers in V. We refer to any such worker as a *spoiler*.

In our incomplete-information setting, preferences of workers coincide across states. Therefore, for a cycle to be present in any particular state, the corresponding workers' preferences need to satisfy certain restrictions across states. Specifically, they need to form a *half cycle*, defined as follows.

Definition 3.2 (Half Cycles)**.** *$G = \{g_1, \ldots, g_K\} \subset F$ and a set of workers $V = \{v_1, \ldots, v_K\} \subset W$ form a half cycle if, for each $k \in \{1, \ldots, K\}$, $g_{k+1} \succ_{v_k} g_k \succ_{v_k} \emptyset$ (where indices are modulo K as before).*

The definition of a half cycle resembles the definition of a cycle, without placing restrictions on firms' preferences or requiring that the relevant firms and workers are matched under a stable matching. Notice that the existence of a half cycle is not tied to a particular state since it relies only on workers' preferences, which are state-invariant. Whenever there is a half cycle involving firms G and workers V in state θ, the preferences of firms G could be modified so that both conditions (2) and (3) of Definition 3.1 are satisfied. Graphically, using the same convention of drawing arrows on the payoff matrices described above, with the modified preferences of the firms in G, there would be a cycle in the matrix.[10]

Figure 3.3(b) illustrates a half cycle involving $\{f_1, f_2\}$ and $\{w_2, w_3\}$. Notice that if preferences of firms f_1 and f_2 were modified to those in Figure 3.3(b), there would be a visual cycle in the matrix. Nonetheless, it would not be a cycle according to our

[10] We note that the cycles introduced in Ergin (2002) and Kesten (2006) are special cases of our half cycles when objects or schools are interpreted as firms in our setting and agents or students are interpreted as workers in our setting.

Definition 3.1 since the stable partners of $\{f_1, f_2\}$ are not $\{w_2, w_3\}$ in when match payoffs are determined according to the matrix in Figure 3.3(a).

3.4 Economies with Unstable Equilibrium Outcomes

In the type of economies we consider, truth-telling by all participants always constitutes a Bayesian Nash equilibrium. Furthermore, truth-telling is not weakly dominated for any participant. In particular, we have the following:

Proposition 3.1 (Implementation of Stable Outcomes)**.** *For any economy* $(\mathcal{F}, \mathcal{W}, \{U(\theta)\}_{\theta\in\Theta}, \Theta, \Psi)$ *with a unique stable matching* $\mu(\theta)$ *in each market* $(\mathcal{F}, \mathcal{W}, U(\theta)), \theta \in \Theta$*, there is a Bayesian Nash equilibrium of the centralized matching economy game whose outcome is* $\mu(\theta)$ *for all* $\theta \in \Theta$*.*

We note that Proposition 3.1 relies on there being a unique stable matching in each market in the support of the economy. Indeed, even with complete information, multiplicity of stable matchings implies that truthful reporting by all participants is not an equilibrium. The setting we study is, therefore, conducive to implementing stable outcomes, as we designed it to be.

Our goal in this section is to characterize economies in which there exist additional equilibria that generate outcomes that are not stable in at least some markets in the support of the economy.

We characterize preferences in one state, call it θ_1, that can correspond to an economy with multiple equilibrium outcomes. In other words, for any set of firms $\mathcal{F}$ and workers $\mathcal{W}$ we characterize a preference profile $\succ = \left(\{\succ^{f_i}\}, \{\succ^{w_j}\}\right)$ such that we can find an economy $(\mathcal{F}, \mathcal{W}, \{U(\theta)\}_{\theta\in\Theta}, \Theta, \Psi)$ that exhibits multiple Bayesian Nash equilibrium outcomes such that $U(\theta_1)$ represents $\succ$.

One and a Half Cycle Condition *We say that a preference profile* $\succ$*, corresponding to a unique stable matching* μ*, satisfies the* One and Half Cycle Condition *if the following hold:*

1. $\succ$ *has a cycle with a unique spoiler* $\bar{w}$*. In that cycle, the spoiler is listed between two workers belonging to the cycle by only one firm* $f_{\bar{k}}$*;*
2. *There is a half cycle that includes the spoiler* $\bar{w}$ *and the firm* $f_{\bar{k}}$*, but does not include the worker* $w_{\bar{k}} = \mu(f_{\bar{k}})$*. Within this half cycle,* $f_{\bar{k}}$ *is not* $\bar{w}$*'s most preferred firm.*

When the One and a Half Cycle Condition holds, preferences exhibit a cycle. Suppose workers implement the swaps that the cycle suggests, while all other workers and firms maintain their (unique) stable match partners. The resulting matching is not stable and, from our discussion above, has a blocking pair $(\bar{w}, f_{\bar{k}})$, where $f_{\bar{k}}$ is part of the cycle and the spoiler $\bar{w}$ is not. For ease of presentation alone, restriction 1 requires that the spoiler $\bar{w}$ and the firm with which he blocks the matching resulting from the swap are unique for at least one cycle. This restriction is weakened in the Online Appendix.

Intuitively, our characterization builds on the idea of constructing other states to provide the spoiler $\bar{w}$ with incentives to drop his claim for $f_{\bar{k}}$. Restriction 2 plants the seeds for such a construction by requiring that the spoiler $\bar{w}$ and firm $f_{\bar{k}}$ be part of a half cycle and that the upper contour set of $f_{\bar{k}}$ is non-empty for $\bar{w}$. This implies that worker $\bar{w}$ prefers forgoing a match with $f_{\bar{k}}$ for a match with some other firm in the half cycle. As we will see, the requirement that the worker $w_{\bar{k}}$, whose stable match partner is $f_{\bar{k}}$, does not take part in this half cycle is important in our construction.

As an example, consider the preferences induced by the match utilities in Figure 3.3(a) and corresponding to state θ_1 in our introductory example. Notice that the One and a Half Cycle condition holds. The spoiler for the cycle involving $\{w_1, w_2\}$ and $\{f_1, f_2\}$ is $w_{\bar{k}} = w_3$. An attempt by workers $\{w_1, w_2\}$ to implement the swap prescribed by the cycle and match w_i with f_{3-i}, $i = 1, 2$, will be blocked by the spoiler w_3 and the firm $f_{\bar{k}} = f_1$. Furthermore, $\{f_1, f_2\}$ and $\{w_2, w_3\}$ form a half cycle that contains $f_2 \succ_{w_1} f_1$ and does not contain $\mu(f_1) = w_1$.

We stress that the One and a Half Cycle Condition does not rule out the possibility of multiple cycles. In fact, if the One and a Half Cycle Condition is satisfied with various cycle and half cycle pairs, there may be multiple economies that induce $\succ$ in some state and yield multiple Bayesian Nash equilibrium outcomes.

We are now ready to state the first main result of this section.

Theorem 3.1 (Multiplicity of Equilibrium Outcomes)**.** *For any set of firms $\mathcal{F}$ and workers $\mathcal{W}$, let $\succ$ be a preference profile satisfying the One and a Half Cycle Condition. Then there exists an economy $(\mathcal{F}, \mathcal{W}, \{U(\theta)\}_{\theta \in \Theta}, \Theta, \Psi)$, where $U(\theta_1)$ represents $\succ$, that exhibits multiple Bayesian Nash equilibrium outcomes. Furthermore, there exists such an economy in which $|\Theta| = 2$.*

The details of the proof can be found in Appendix B. Below we discuss heuristically the main idea behind the proof, which generalizes the forces underlying the introductory example of section 3.2.

Consider $\succ$, with a unique stable matching μ, that satisfies the One and a Half Cycle Condition. A cycle as defined in the definition of this condition implies that there is a sub-market, involving the cycle's participants, that entails multiple stable matchings. In this sub-market, the implementation of the swap defined by the cycle would be beneficial for the workers in the cycle. However, such a swap would be blocked by worker $\bar{w}$, the spoiler, and some firm $f_{\bar{k}}$. Notice that absent $\bar{w}$, in the sub-market induced by the cycle, "truncation" by one of the workers, say by $w_{\bar{k}} = \mu(f_{\bar{k}})$, would allow the desirable swap by workers. How can we induce the spoiler $\bar{w}$ to drop his claim for firm $f_{\bar{k}}$? That is where the existence of a half cycle comes into play. We construct another state in which firms' preferences are such that at least some of the firms and workers corresponding to the half cycle form a full cycle. This imposes restrictions on preferences of firms that take part in the original half cycle as well as on preferences of other firms. Indeed, the constructed state has to exhibit a unique stable matching. Furthermore, in the cycle constructed in this new state, participating firms and workers need to be matched to one another through the unique stable matching. Importantly, since $\bar{w}$ and $f_{\bar{k}}$ take part in this cycle, and $f_{\bar{k}}$ is not $\bar{w}$'s favorite within the half cycle, preferences can be constructed so that $\bar{w}$ has an incentive to truncate his preferences in the corresponding sub-market. In particular, preferences can be constructed so that $\bar{w}$ has an incentive to drop the claim for firm $f_{\bar{k}}$. From uniqueness, there would then be spoilers in the new state and we would need to make sure they drop their claims for blocking partners. As it turns out, we can make sure the roles of $\bar{w}$ and $w_{\bar{k}} = \mu(f_{\bar{k}})$ are reversed in the two states. In particular, in the newly constructed state, $w_{\bar{k}} = \mu(f_{\bar{k}})$ is the unique spoiler whose incentives to drop his claims for blocking firms is given symmetrically through his participation in the cycle corresponding to our original preferences $\succ$.

Ultimately, the spoiler $\bar{w}$ in state θ_1, corresponding to our preference profile $\succ$, is compensated for dropping his claim for his blocking partner $f_{\bar{k}}$ through the improved match he receives in state θ_2. Symmetrically, the spoiler in state θ_2 is compensated for dropping his claim for his blocking partner in that state through state θ_1. That drop allows him to implement a beneficial swap in the cycle we started out with in state θ_1.

By appropriately choosing cardinal presentations of preferences that are consistent

with these implied preference orderings, we can make sure that spoilers best respond by dropping claims for their blocking partners.

Several points are worth stressing. First, at the root of our construction is a specification of ordinal preferences in the added state. Cardinal presentations of these preferences as well as the probabilities places on each state are important, as they make sure that the dropping strategies are, in expectation beneficial. We note that, in this case, where the construction involves only two states, θ_1 and θ_2, the precise specification of probabilities does not matter. Indeed, if the probability of θ_1 is p in this economy, we could construct another economy $(\mathcal{F}, \mathcal{W}, \left\{\tilde{U}(\theta)\right\}_{\theta\in\Theta}, \Theta, \Psi)$ with the same states that are equally likely, where $\tilde{U}(\theta_1) = pU(\theta_1)$ and $\tilde{U}(\theta_2) = (1-p)U(\theta_2)$. That economy would also generate multiple equilibrium outcomes. More importantly, the set of utilities and probability distributions that guarantee multiplicity of equilibrium outcomes is not knife-edge.[11]

Second, we assumed here that there is a cycle with a unique spoiler, who blocks the cycle with only one firm. We relax this assumption in the Online Appendix. When there are more spoilers or firms with which each spoiler blocks a candidate cycle, more states may need to be added to make sure spoilers drop their claims for each and every relevant blocking partner. The construction is similar to that described here, but notationally more cumbersome.

Last, there is a connection between the cycle we identify in the preference profile $\succ$ and the cycles defined in Ergin (2002) and Kesten (2006) in the context of assignment problems. For illustration, suppose the cycle we identify is between workers $\{w_1, w_2\}$ and firms $\{f_1, f_2\}$ and suppose $\mu(f_i) = w_i$, $i = 1, 2$. The existence of a cycle implies that $w_1 \succ_{f_1} w_2 \succ_{f_2} w_1$ and $f_1 \succ_{w_2} f_2 \succ_{w_1} f_1$. That is, workers w_1 and w_2 in this sub-market would like to swap stable partners and implement $\mu'(f_i) = w_{3-i}$, $i = 1, 2$. Since μ is the unique stable matching, there must be a spoiler, say w_3, who blocks this swap with one of the two "swapped" firms, say f_1. This implies that $w_1 \succ_{f_1} w_3 \succ_{f_1} w_2$. In addition, as stated, $w_2 \succ_{f_2} w_1$. This is precisely the Cycle Condition in Definition 1 of Ergin (2002), where objects or schools are interpreted as firms and agents or students are interpreted as workers. We stress that our definition of a cycle is far more restrictive, as it imposes restrictions on the preferences of the corresponding workers as well.

[11] Namely, it has a positive Lebesgue measure.

3.5 The Rural Hospital Theorem

The Rural Hospital Theorem (McVitie and Wilson, 1970) guarantees that unmatched individuals are identical across stable matchings. In the complete-information setting, when the proposing side of DA truthfully reveals, the set of equilibrium outcomes coincides with the set of stable matchings (Roth, 1984; Gale and Sotomayor, 1985). This implies that, with complete information, equilibrium selection in DA will not affect the set of matched individuals.

At the root of the Rural Hospital Theorem is the existence of a stable matching μ that is preferred by all members of one side of the market, say the firms, to any other stable matching and is the worst for all members of the other side of the market, the workers. Such a (complete-information) stable matching always exists. For instance, the firm-proposing DA generates a stable matching that is firm-optimal and worker-pessimal. Suppose μ leads to firms F and workers W being matched. Take any stable matching μ' that yields firms F' and workers W' being matched. Since μ is the firm-optimal stable matching and, in particular, individually rational, it must be that $F \supseteq F'$. Similarly, since μ' is preferable by the workers to μ and individually rational, it must be that $W' \supseteq W$. Now, $|F'| = |W'|$ and $|F| = |W|$, which implies that $F = F'$ and $W = W'$.

In the example of section 3.2, the two equilibrium outcomes we identify do not have the feature above—neither is optimal for one side of the market and pessimal for the other. That is where the Rural Hospital Theorem breaks down.

The example highlighted in section 3.2 is not special. As it turns out, whenever we start with an economy that exhibits multiplicity of equilibria, we can expand the market by adding one firm and one worker and generate a new economy in which the set of unmatched individuals differs across stable matchings. This addition can be fairly minimal in that the set of states and the distributions over them remains the same, as well as the match utilities derived from matching individuals in the economy we started out with.

Formally, for any economy $\mathcal{E} = (\mathcal{F}, \mathcal{W}, \{U(\theta)\}_{\theta\in\Theta}, \Theta, \Psi)$ and any firm $f \notin \mathcal{F}$ and worker $w \notin \mathcal{W}$, an *augmented economy of $\mathcal{E}$ with firm f and worker w* is an economy $\tilde{\mathcal{E}} = (\mathcal{F} \cup \{f\}, \mathcal{W} \cup \{w\}, \left\{\tilde{U}(\theta)\right\}_{\theta\in\Theta}, \Theta, \Psi)$ such that for any $f_i \in \mathcal{F}$ and $w_j \in \mathcal{W}$, for any state $\theta \in \Theta$,

$$\tilde{u}^f_{ij}(\theta) = u^f_{ij}(\theta), \tilde{u}^w_{ij}(\theta) = u^w_{ij}(\theta), \tilde{u}^f_{i\varnothing}(\theta) = u^f_{i\varnothing}(\theta), \tilde{u}^w_{\varnothing j}(\theta) = u^w_{\varnothing j}(\theta).$$

An augmented economy is restricted to satisfy our restriction that each realized market entails a unique stable matching.[12]

Theorem 3.2 (Unmatched Individuals in Equilibria)**.** *Take any economy $\mathcal{E}$ exhibiting multiplicity of Bayesian Nash equilibrium outcomes. Suppose there is one such equilibrium outcome in which, in some state, some worker is matched to a partner inferior to his stable match partner. Then there exists an augmentation with one firm and one worker that exhibits multiplicity of Bayesian Nash equilibrium outcomes and different distributions of unmatched individuals across these outcomes.*

The example in section 3.2 illustrated a case in which the unstable equilibrium outcome entails fewer matched individuals than the stable one. As it turns out, this is not generally the case. Unstable equilibrium outcomes may also correspond to a larger set of matched individuals relative to the stable outcome. We provide an example in the Online Appendix.

3.6 Economies with a Unique Stable Equilibrium Outcome

In previous sections we identified a class of markets in which, despite small cores in each realized state, multiple equilibrium outcomes emerged. At the root of that class of markets were cycles in preferences, at least in some states. In other words, in at least some states, a sub-market entailed multiple (complete-information) stable matchings. In the absence of such cycles, the construction we provide for generating an economy with multiple equilibria could not be used. In fact, in this section, we highlight economies that exhibit no such cycles and entail a unique equilibrium implementing the unique stable matching in each state. Specifically, we focus on markets in which participants on at least one side of the market are characterized by assortative preferences (*à la* Becker, 1973).

Assortative Preferences of Firms

Assume that firms share the same ranking over workers in each state. That is, in each state θ_i, there is a permutation $\pi_i : \{1, ..., m\} \to \{1, ...m\}$ such that all firms agree on the ranking

$$w_{\pi_i(1)} \succ w_{\pi_i(2)} \succ ... \succ w_{\pi_i(m)},$$

and, for ease of presentation, we assume that all workers are acceptable in each state.

[12] Notice that such an augmentation always exists. Indeed, one can always specify firm f and worker w as one another's only acceptable partners in each state.

Proposition 3.2. *If firms share the same ranking over workers state by state, then there is a unique Bayesian Nash equilibrium outcome corresponding to the unique (complete-information) stable outcome in each state.*

Intuitively, for any state θ, in the course of the firm-proposing DA, all firms that are active apply to the same worker. This fact is driven by the common preferences of firms and is independent of the workers' reports. Thus, the set of active firms constitutes a family of nested sets. Therefore, regardless of workers' reports, each worker is called to choose only once in the course of the firm-proposing DA and a worker's choice over the set of active firms is decisive. In particular, there is no sense in which misreporting can be beneficial to workers. The only potential impact of a worker misreporting is that he may forgo a firm he would otherwise match to and, consequently, be matched to an inferior firm, or no firm at all.

In terms of our construction above, assortative preferences for firms ensure that workers cannot generate a fictitious cycle through their reports, thus ensuring that it is a dominant strategy to report truthfully, and ruling out any Bayesian Nash Equilibrium that is (complete-information) unstable.

Assortative Preferences of Workers

Suppose now that workers share the same ranking of firms. Workers' preferences are identical across states and, without loss of generality, we can assume their ranking is given by:

$$f_1 \succ f_2 \succ ... \succ f_n.$$

Proposition 3.3. *If workers share the same ranking over firms, then there is a unique Bayesian Nash equilibrium outcome corresponding to the unique (complete-information) stable outcome in each state.*

The intuition underlying this Proposition is the following. Suppose $\mu(\theta)$ is the unique (complete-information) stable matching in state θ and that there exists an equilibrium yielding the matching $\lambda(\theta)$ in each state θ such that $\lambda(\theta) \neq \mu(\theta)$ for at least one state θ. Consider the smallest integer k such that $\lambda(\theta; f_k) \neq \mu(\theta; f_k)$ for some state θ.

It follows that worker $w = \mu(f_k)$ either reports f_k as unacceptable or reports $\lambda(w)$ as preferable to f_k. Suppose that, in this equilibrium, w reports $\succ^*$ such that

$$f_{\pi(k+1)} \succ^* f_{\pi(k+2)} \succ^* ... \succ^* w \succ^* f_{\pi(j)} \succ^* ... \succ^* f_{\pi(n)}$$

for some permutation $\pi : \{k+1, ..., n\} \rightarrow \{k+1, ..., n\}$. As it turns out, the following deviation is profitable for w : a report of $\succ'$ such that

$$f_1 \succ' f_2 \succ' ... \succ' f_k \succ' f_{\pi(k+1)} \succ' f_{\pi(k+2)} \succ' ... \succ' w \succ' f_{\pi(j)} \succ' ... \succ' f_{\pi(n)}.$$

In this deviation, the worker reports truthfully his preferences over his true top k firms and maintains the ranking used in the equilibrium yielding λ for all other firms. In any state in which the worker is supposed to match with f_j, $j > k$, under the stable matching μ, these reports assure that the worker either receives the partner he received under the original equilibrium, or a preferable firm (from the set $\{f_1, ...f_k\}$). In any state in which the worker is supposed to match with f_j,$j \leq k$, under the stable outcomes, the minimality of k comes into play. Intuitively, since k is chosen minimally, other workers' reports will not impede on the agent getting his stable match partner. In particular, in state θ, the worker would strictly improve on his match partner.

3.7 Conclusions

In this paper, we focus on centralized matching clearinghouses that are often used in applications. We highlight the fragility of several canonical results to the commonly employed complete information assumption. The literature has argued extensively that small cores resolve well-known incentive compatibility issues in such clearinghouses. In contrast, we show that small cores are no panacea. With arguably a small amount of incomplete information, equilibrium outcomes can change dramatically when market participants are not fully informed of all other participants' preferences.

In particular, there are four messages the paper suggests. First, incentive compatibility issues remain with small cores when incomplete information is present. Indeed, our Theorem 1 provides a characterization of settings in which small cores coexist with multiplicity of equilibrium outcomes that. Second, equilibrium outcomes corresponding to unstable matchings may actually be desirable for the receiving side in a DA clearinghouse, despite the traditional view of the proposing side as advantaged. Third, the set of matched individuals may differ across equilibrium outcomes and we offer a characterization of when that may happen. Unlike the complete information benchmark, this insight suggests that selection of equilibria can offer a useful instrument for affecting who gets matched. Last, we show that several specific technical simplifications that are used frequently when information

is complete, cease to hold when there is uncertainty about preferences. Specifically, best responses no longer necessarily take the form of truncation strategies.

Taken together, the results illustrate the importance of taking into account details pertaining to the information market participants have in centralized matching clearinghouses.

Turned on their head, our results offer caution for empirical studies of matching markets that use stability constraints to deduce participants' preferences over partners. Indeed, market features that guarantee small cores when information is complete, such as large volumes of participants and market imbalances, may be insufficient for guaranteeing unique equilibrium outcomes that are stable for the instantiation of the market a researcher observes.

Chapter 4

IMPLICATIONS OF OVERCONFIDENCE ON INFORMATION INVESTMENT

Overconfidence is a well documented behavioral bias by which an individual believes to have better information or perform at a task better than he actually does. On the other hand, a large body of literature has focused on understanding the decision to invest in information made by purely rational agents. This issue is relevant to several applications, such as bench trials, where a judge decides how much effort he is going to spend to collect information relevant to the case. Whether the results of this last strand of literature remain valid when agents are overconfident is an open question.

Moore and Healy (2008) classify overconfidence into three categories: overestimation, overplacement and overprecision. Overestimation and overplacement refer to cases where an individual thinks he has performed in a task better than he actually did or better than others, respectively. Overprecision is the case when an individual believes her information is more precise than it actually is. This paper deals with overprecision, the most prevalent and the least reversed phenomenon of the three (Moore and Healy, 2008; Benoît and Dubra, 2011; Ortoleva and Snowberg, 2015). In the reminder of the paper we refer to the overprecision-type bias as overconfidence.

The leading example we are using in this paper is about a judge who must decide whether to convict or acquit a defendant who can be either innocent or guilty. In this example, what the judge decides affects not only the judge but a lot of people, the defendant including. If we take the society as a whole, it cares about the quality of the verdict, that is, it wants to acquit innocent and convict guilty defendant. The question we address is how overconfidence in the judge's perception of information affects the probability that he reaches the correct verdict.

According to Moore and Healy (2008)'s informal definition, overconfidence is the "excessive precision in one's beliefs." In general, there are two types of beliefs — prior and posterior. Suppose the overconfidence occurs in the prior belief. For example, the judge thinks he knows a lot about the case while he actually knows less. Then his incentive to acquire additional information is lower than it should be.

So, the overconfidence in the prior belief leads to the lower quality of the verdict.

Now suppose the overconfidence occurs in the posterior belief. By definition, the posterior belief is the updated belief after acquiring information. So, the overconfidence in posterior means the excessive precision of that information. The reason for that might be that the judge overestimates his ability to analyze information.[1] If the judge thinks he gets information of higher quality, his incentives to acquire that information is higher. This means that the overconfidence in the information quality leads to the higher quality of the verdict.

Note that the notion of prior and posterior is relative, since today's posterior becomes tomorrow's prior. In this paper we assume that the judge starts with a uniform prior and consider a dynamic model of information collection. We model overconfidence as the distortion in the perceived precision of information flow.[2] With this model, we argue that the total effect that overconfidence has on the probability of the judge reaching the correct decision depends on the nature of the information collection process. More specifically, it depends on how much control the judge has over the amount of information he collects.

First, suppose an unbiased judge chooses whether to collect information or not, that is the information investment choice space is binary (section 4.3). For example, he decides whether to hold the trial or announce the verdict right away. Then overestimating the quality of information (or equivalently, the ability to perceive information) leads to higher willingness to pay for that information. This means that overconfidence has a positive effect on the probability of the correct verdict.

Now suppose the judge can decide how much information to collect (section 4.3). For example, he decides on the length of the trial. Then the effect of overconfidence

[1] Another interpretation is more literal, when the decision maker overestimates the quality of the information source (an expert, for example) he gets information from. In other words, perceiving information as being more precise than it actually is can be seen as a sign of excessive gullibility of the judge. In such interpretation, this phenomenon is the opposite of the most common formalization of overconfidence as overprecision of the prior. We have two arguments in response to that concern. First, one might think of the overconfidence ex-post, that is, at the time of the decision making. Then the posterior becomes the prior and the traditional definition of the overconfidence is applied. Moreover, in a dynamic model, — which is the main model in our paper, — this is no longer an issue due to prior and posterior being relative notions. Second, we can appeal to our first interpretation of this phenomenon as overestimation of one's ability to analyze information (similar to Heidhues, Koszegi, and Strack (2017) who model overconfidence as misperception of one's ability, which leads to misperception in the beliefs about the true state of the world). That essentially mixes two out of three notions of overconfidence offered by Moore and Healy (2008).

[2] By assuming no distortion in prior belief, we take as given the source of overconfidence (how much the judge overestimates the precision of the incoming information) and get the overconfidence in beliefs endogenously.

can be either positive or negative, depending on how much a *rational* judge would invest in information. If the rational judge invests very little (he holds a very short trial), then overconfidence increases information investment. On the other hand, if the rational judge invests a lot, overconfidence has the opposite effect. In general, the effect is shaped by two forces. The first is the only active force in the binary investment choice scenario. It comes from increasing the marginal benefit from each hour of trial. The second force comes from increasing the total benefit from a fixed investment. When the rational judge's investment is high, the second force prevails. Moreover, we show that there is an optimal level of belief distortion (either to over-precision or to under-precision) that maximizes the probability of the correct decision by balancing the two forces. This optimal level is higher when the judge does not care too much about choosing the correct sentence.

Finally, suppose the judge can decide how much information to collect dynamically (section 4.2). In contrast to the previous scenario, the judge does not have to decide upfront how long the trial would be and can stop the trial at any moment in time. In this setting we find the effect of overconfidence to be detrimental to the quality of the judgment. The dynamic nature of information collection in this scenario introduces a third force that pushes investment down. This third force describes an excessive sensitivity to the noise in information flow. By overestimating the quality of information, the judge treats unexpected noise as a meaningful signal and therefore his belief about the defendant's innocence reaches his desired standard of proof threshold belief sooner than he (ex-ante) expects. It turns out that under the assumptions of a normally distributed information flow and of a symmetric payoff, the net effect from all three forces is negative, meaning that having an underconfident judge is always better for the quality of the judgement. Intuitively, when the second force is weak, that is, when the accumulated information is low, the judge is very sensitive to noise, which makes the third force strong.

We can look at the forces from the prior-posterior perspective at a given moment in time. The first force corresponds to having excessive precision in posterior, as it comes from increasing the precision of information the decision maker is about to collect. The other two forces come from excessive precision in prior, as they both come from overestimating the quality of information already collected.

Since the dynamic setting introduces a third force that decreases the probability of the correct verdict, it is natural to conjecture that restricting the overconfident judge to commit upfront to the amount of information he is going to collect will increase

the probability of the correct verdict. However, that logic is wrong because it does not take into account the difference in the dynamic and the static benchmarks with the rational judge, where all forces are absent. The dynamic benchmark gives the judge additional flexibility and therefore the probability of the correct verdict is higher than in the static benchmark. In section 4.4, we formally show that there is a unique level of overconfidence such that below that level the dynamic settings leads to the higher probability of the correct verdict, while when the actual level is higher, its is socially optimal to restrict the overconfident judge to commit to the length of the trial.

So far, we considered the probability of the correct verdict as the welfare criterion, ignoring the cost of information collection. In section 4.4, we show that our conclusions are robust to that assumption. More precisely, we consider a principal-agent model, where a rational principal delegates the information collection and the decision to an overconfident agent. For example, the society hires a judge whose job is to held a trial and decide on a verdict. Now the judge himself does not care about the quality of the verdict, so the society has to pay him differently depending on the true state to create proper incentives. Under assumption that the payment cannot be negative (that is, the principal cannot take money from the agent), we show that there is a unique level of overconfidence such that below that level the optimal contract does not require the agent to commit to the amount of information he is going to collect, while this requirement is optimal above that level. Moreover, conditional on the set of contracts without commitment, increasing overconfidence is always bad, both from the principal's perspective and from the social welfare perspective (when the criterion is the sum of the principal's and the agent's objective expected utility). However, conditional on the set of contracts with commitment, there is an optimal level of overconfidence. This optimal level is higher from the principal's perspective than from the social welfare perspective, which reflects the principal's power to exploit the agent's irrationality to her own advantage.

Related Literature

The overconfidence phenomenon has been studied in many settings, including (but not limited to) financial markets (Scheinkman and Xiong, 2003; Kyle, Obizhaeva, and Wang, 2017), medicine (Berner and Graber, 2008), war (Johnson, 2009), political behavior (Ortoleva and Snowberg, 2015). Much evidence that people are prone to overconfidence has been documented in literature. Barber and Odean (2001); Chuang and Lee (2006); Goetzmann and Huang (2015) found empirical

support for overconfidence in financial environments. One of the earliest studies, Oskamp (1965), experimentally demonstrates overconfidence among actual judges when they are presented with information about published cases. Klayman, Soll, Gonzalez-Vallejo, and Barlas (1999) and Soll and Klayman (2004) provide more recent experimental evidence for judges' overconfidence. Using actual data on bail decisions made by judges in New York City between 2008 and 2013, Kleinberg, Lakkaraju, Leskovec, Ludwig, and Mullainathan (2017) show that judges appear to "respond to 'noise' as if it were signal." From the perspective of Moore and Healy (2008)'s classification of overconfidence, this observation can be interpreted as overprecision. By mixing the actual signal with noise, judges are effectively boosting the perceived precision of all incoming information as a whole.

An alternative way to model overconfidence as misperception of the quality of information is through correlation neglect. Ortoleva and Snowberg (2015) find theoretically and verify empirically that overconfidence (modeled as correlation neglect) leads to ideological extremeness, increased voter turnout and stronger partisan identification. Levy and Razin (2015) focus on information aggregation, as opposed to information investment, and find conditions under which correlation neglect can lead to increased information aggregation. In contrast Glaeser and Sunstein (2009) study a "credulous Bayesian" that neglects correlation in a context where there is no cost of information acquisition and find overconfidence to be detrimental to information aggregation. We depart from these studies in modeling overconfidence as a misperception of the precision parameter.[3] In contrast to correlation neglect, misperception of the precision can potentially occur in any setting, even if there is actually no correlation in the incoming information (conditional on the true state). In fact, we mostly focus on that case in this paper, though section 4.3 presents more general results as well.

Our discussion about an overconfident judge can also be applied to a jury room, where a juror misperceives the quality of her own information. This relates our paper to the literature studying information acquisition or investment by committees. Martinelli (2006) considers a setup when each committee member chooses how much to invest in the precision of a binary signal when costs are convex. This setup is very similar to our model in section 4.3. Chan, Lizzeri, Suen, and Yariv (2017) work with a dynamic setup that are close to our model in section 4.2. All papers

[3] Dubra (2004) defines overconfidence as an optimistic bias in prior beliefs. This interpretation of overconfidence is orthogonal to a more popular definition of overconfidence as underestimating the volatility (Alpert and Raiffa, 1982). We used the latter one.

assume that the jurors are rational.

Scheinkman and Xiong (2003) and Kyle, Obizhaeva, and Wang (2017) model overconfidence in a way similar to our model in section 4.2. Scheinkman and Xiong (2003) explain speculative bubbles using overconfidence. Overconfidence as misperception of information quality generates disagreement about fundamentals which results in a price bubble. Kyle, Obizhaeva, and Wang (2017) explain large trading volume and price dampening using that disagreement. However, both papers did not allow for the agents to choose whether to observe information flow or not.

4.1 Setup

Consider a single decision maker who has to decide between two actions. The payoff from these actions depends on the true state of the world. For example, suppose this decision maker is a judge who decides whether to acquit ($v = A$) or convict ($v = C$) a defendant. The defendant might be either innocent ($z = I$) or guilty ($z = G$). In this case z plays the role of the true state of the world.

We assume the decision maker gets a payoff $u(v, z)$ from action $v \in \{A, C\}$ if the true state is $z \in \{I, G\}$. At the beginning, the decision maker has some prior beliefs about the true state, $p_0 = \mathbb{P}(z = I)$. Given belief p, his expected utility from action v is $U(v, p) = pu(v, I) + (1 - p)u(v, G)$.

Naturally, we assume that when the defendant is innocent, it is better to acquit her, and when she is guilty, it is better to convict her. Moreover, for simplicity, we focus on the symmetric case:

Assumption 4.1. $u(A, I) = u(C, G) > u(C, I) = u(A, G)$.

Denote

$$Q = u(A, I) - u(C, I) = u(C, G) - u(A, G) > 0, \quad R = u(C, I) = u(A, G).$$

Thus, the judge gets utility R from the incorrect verdict and utility $Q + R$, $Q > 0$, from the correct verdict. Maximizing his expected utility, he gets

$$\max_{v \in \{A,C\}} U(v, p) = \begin{cases} pQ + R, & p \geq \frac{1}{2}, \\ (1 - p)Q + R, & p \leq \frac{1}{2}. \end{cases}$$

Before deciding on the action, the decision maker can collect more information about the true state. For simplicity, we restrict our analysis to a symmetric case when the decision maker has no initial bias in his belief:

Assumption 4.2 (Uniform prior belief)**.** $p_0 = 0.5$.

4.2 Dynamic Model

The decision maker collects information by observing the change in a Brownian motion process with state-dependent drift:

$$dX_t = \mu_z dt + \sigma dW_t, \quad \mu_z = \begin{cases} 1, & \text{if } z = I, \\ -1, & \text{if } z = G, \end{cases} \tag{4.1}$$

where W_t is the standard Wiener process.

Information collection is costly. In a dynamic setting, we can differentiate two types of costs, attention cost and time cost. The time cost is formalized through a discount factor $\delta \geq 0$. For simplicity, we focus on the no-discounting case here ($\delta = 0$). Appendix C.13 shows that the case $\delta > 0$ leads to the same conclusions.

The attention cost is proportional to the amount of time the decision maker spends on the information collection. Formally, the decision maker chooses a stopping time $\tau \geq 0$ (which is path-dependent, that is, whether or not the decision maker stops by t depends on X_t) and, upon stopping, a verdict $v \in \{A, C\}$ (which depends on X_τ). The utility he eventually gets is equal to $u(v, z) - \kappa\tau$, where $\kappa > 0$ is a parameter of the model.

The problem the decision maker faces is called an optimal stopping problem. It has already been studied in the literature, so we can take the solution off-the-shelf:

Theorem 4.1 (Shiryaev (2007), Chapter 4, Theorem 5, p.185)**.**

The optimal strategy exists and is given by

$$\tau = \inf\left\{t \geq 0 \colon p_t \notin (\lambda, 1-\lambda)\right\}, \quad v = \begin{cases} A, & p_\tau \geq 1 - \lambda, \\ C, & p_\tau \leq \lambda, \end{cases} \tag{4.2}$$

where p_t is the belief that the true state is I at time t. Threshold $\lambda \in (0, 0.5)$ is uniquely defined by

$$\frac{1-2\lambda}{2\lambda(1-\lambda)} - \log\left(\frac{\lambda}{1-\lambda}\right) = \frac{Q}{\kappa\sigma^2}. \tag{4.3}$$

For completeness, we include the proof in Appendix C.1.

When observing (4.1), the decision maker updates his belief about the state[4]

$$p_t = \mathbb{P}\left[z = I \mid X_t\right] = \frac{1}{1 + e^{-\frac{2X_t}{\sigma^2}}}, \tag{4.4}$$

which is equivalent to

$$X_t = \frac{\sigma^2}{2} \log\left(\frac{p_t}{1 - p_t}\right). \tag{4.5}$$

So, another way to write the optimal strategy (4.2) is

$$\tau = \inf\left\{t \geq 0 \colon X_t \notin (-\chi, \chi)\right\}, \quad v = \begin{cases} A, & X_\tau \geq \chi, \\ C, & X_\tau \leq -\chi, \end{cases} \tag{4.6}$$

where $\chi = \frac{\sigma^2}{2} \log\left(\frac{1-\lambda}{\lambda}\right) > 0$. The advantage of this representation is that it expresses the strategy in terms of an external (observable) variable X_t and not in terms of a mental quantity which is the decision maker's belief. This distinction is important to us since overconfidence introduces a distortion in the belief updating rule, so that an overconfident person would form a different belief than a rational one, given the same observed process X_t.

Definition 4.1. *An η-**type** decision maker updates his belief according to*

$$p_t = \frac{1}{1 + e^{-\frac{2X_t\eta}{\sigma^2}}} \tag{4.7}$$

while observing

$$dX_t = \mu_z dt + \sigma dW_t. \tag{4.8}$$

In other words, the η-type decision maker believes the variance is η times lower than it actually is. Parameter η captures the level of overconfidence, with $\eta = 1$ corresponding to the rational case. Thus, the η-type decision maker is overconfident when $\eta > 1$ and he is underconfident (he underestimates the precision of information) when $\eta < 1$.[5]

Given the observed process (4.8) and the strategy (4.6) with a fixed $\chi > 0$, the probability of the correct decision (the probability of acquittal if the defendant is

[4] We assume $X_0 = 0$.

[5] Throughout the paper, we assume that the decision maker is not aware of his overconfidence. If we allow the decision maker to learn the level of overconfidence dynamically, he would learn it immediately. Indeed, by the properties of the Brownian motion, any uncertainty about the variance is resolved on the spot.

innocent and of conviction if the defendant is guilty) is

$$\mathbb{P}\left(v=\begin{cases}A, & z=I\\ C, & z=G\end{cases}\right)=\mathbb{P}\left[v=A\mid z=I\right]=\mathbb{P}\left[v=C\mid z=G\right]=\frac{1}{1+e^{-\frac{2\chi}{\sigma^2}}}. \tag{4.9}$$

Indeed, this probability is equal to the probability that the decision is correct at the time when this decision is made.

Consider the η-type decision maker. His optimal strategy is (4.6) with threshold $\chi=\mathcal{X}\left(\frac{\sigma^2}{\eta}\right)$, where

$$\mathcal{X}\left(\sigma^2\right)=\frac{\sigma^2}{2}\log\left(\frac{1-\lambda\left(\sigma^2\right)}{\lambda\left(\sigma^2\right)}\right), \tag{4.10}$$

where $\lambda\left(\sigma^2\right)\in(0,0.5)$ is the solution to (4.3).

Note that the probability of the correct decision (4.9) is increasing in the threshold χ. Theorem 4.2 states that the higher the overconfidence level η, the lower the probability of the correct decision.

Theorem 4.2. *$\mathcal{X}\left(\sigma^2\right)$ defined by (4.10) is increasing in σ^2.*

See Appendix C.2 for the proof.

Intuitively, the expression $\chi=\frac{\sigma^2}{2}\log\left(\frac{1-\lambda}{\lambda}\right)$ shows that increasing σ^2 has two effects on χ. The direct effect increases χ. This effect comes from the attempt to keep the same standard of proof by collecting more information that is less precise. The indirect effect decreases χ through λ (λ is increasing in σ^2). This effect comes from the attempt to keep the same total cost of information by lowering the standard of proof $1-\lambda$. Theorem 4.2 states that the first effect always dominates the second.[6]

While being explicitly connected to the formula for $\mathcal{X}\left(\sigma^2\right)$, these two effects give an ambiguous prediction for the expected stopping time. On the one hand, increasing χ without changing the variance increases the expected stopping time. On the other hand, increasing the variance without changing the threshold decreases the expected stopping time. Thus, if the variance is actually changing, the first effect has an unclear prediction for whether the expected stopping time increases or decreases. If the variance is not actually changing, yet the decision maker thinks it increases, the second effect decreases the stopping time by lowering the perceived standard

[6] This result is not robust to relaxing Assumption 4.1, see Appendix C.14.

of proof and increases it by not updating aggressively enough. This distinction is important for us because it illuminates a commitment aspect of the information collection problem.

Suppose the decision maker has to decide ex-ante when to stop information collection, that is, he has to commit on the stopping time τ at time $t = 0$. Then there are two forces that shape the overall effect on τ from increasing σ^2. The first force comes from decreasing the benefit of the marginal information piece dX_t and therefore it lowers τ. The second force comes from decreasing the benefit of information already collected, X_t, and therefore it increases τ. Once we drop the commitment restriction, another force arises. This third force captures the discrepancy between what the decision maker expects to see (information flow with a high variance) and what he actually observes (information flow with a low variance). Though not changing his perception of the variance once observing X_t, the decision maker bases his stopping decision on the low variance information flow. Thus, the third force increases τ since the decision maker does not update enough thinking he observes more noise than he actually does.

We elaborate on the commitment model and the first two forces in the next section. We conclude this section by expanding the intuition for the third force.

Suppose that at time $t = 0$ the decision maker commits to stop collecting information at a certain time τ. The commitment prevents the decision maker to collect more information when $|X_\tau|$ is too small (which corresponds to low standard of proof). It also prevents him from stopping the information collection process earlier when $|X_\tau|$ is too large (high standard of proof). The optimal τ balances out these two events based on distribution X_τ. An η-type decision maker expects to observe $\mu_z\tau + \frac{\sigma}{\sqrt{\eta}}W_\tau$ distributed as $\mathcal{N}\left(\mu_z\tau, \frac{\sigma^2}{\eta}\right)$, while he actually observes $\mu_z\tau + \sigma W_\tau$ distributed as $\mathcal{N}(\mu_z\tau, \sigma^2)$. Thus, the η-type decision maker, $\eta > 1$, underestimates the probability $|X_\tau|$ being large. In other words, the η-type decision maker wants to stop the information collection before the committed stopping time with higher probability than he believes at $t = 0$. Thus, in the absence of commitment the η-type decision maker stops sooner. This effect is captured by the third force.

4.3 General Static Model

The decision maker collects information by acquiring a signal $S \in \mathcal{S}$ that has state-dependent distribution $F_z(\cdot)$. Upon observing $S = s$, the optimal verdict is $v = A$ if $dF_I(s) > dF_G(s)$ and $v = C$ otherwise. Denote set $\mathcal{S}_A = \{s\colon dF_I(s) >$

$dF_G(s)\}$ all realizations of S that leads to an acquittal decision. Similarly, denote $\mathcal{S}_C = \{s\colon dF_I(s) < dF_G(s)\}$ all realizations of S that leads to conviction. We assume that the measure of $\mathcal{S} \setminus (\mathcal{S}_A \cup \mathcal{S}_C)$ is zero under any state, so that the decision maker is (almost) never indifferent between the two verdicts. Denote $p_{z,v} = \mathbb{P}[S \in \mathcal{S}_v \mid z]$ the probability of making the decision $v \in \{A, C\}$, given state z. Then the probability of the correct decision is $\frac{1}{2}(p_{I,A} + p_{G,C})$ and therefore the expected utility from signal S is $\frac{1}{2}(p_{I,A} + p_{G,C})\,Q + R$. When the decision maker does not use the signal, his expected utility is $\frac{Q}{2} + R$. Thus, the quality of signal S can be summarized by

$$\frac{1}{2}(p_{I,A} + p_{G,C}) - \frac{1}{2}.$$

We assume that the decision maker can increase the quality by paying more for the signal. Formally, the quality of the signal is an increasing function of cost, $h(c)$. Thus, the expected utility is

$$\left(\frac{1}{2} + h(c)\right) Q + R - c$$

and the decision maker chooses cost $c > 0$. The first order condition is

$$h'(c)Q = 1. \tag{4.11}$$

To guarantee that the solution to (4.11) exists, is unique and maximizes the expected utility, we assume

Assumption 4.3. *$h\colon (0, +\infty) \to \left[0, \frac{1}{2}\right]$ is such that $h(0) = 0$, $\lim\limits_{c\to 0} h'(c) = +\infty$, $\lim\limits_{c\to+\infty} h'(c) = 0$, $h''(c) < 0$.*

Given that general model, we impose the following definition of overconfidence:

Definition 4.2. *An η-**type** decision maker perceives the quality of the signal being $h(\eta c)$ while paying c.*

Consider the η-type decision maker. His expected utility from signal S is

$$\left(\frac{1}{2} + h(\eta c)\right) Q + R - c. \tag{4.12}$$

Example (Normal distribution) When $S \equiv X_t \sim \mathcal{N}(\mu_z t, \sigma^2 t)$, its quality is equal to $f\left(\frac{t}{\sigma^2}\right)$, where

$$f(\rho) = \frac{1}{\sqrt{\pi}} \int_0^{\sqrt{\frac{\rho}{2}}} e^{-x^2} dx. \tag{4.13}$$

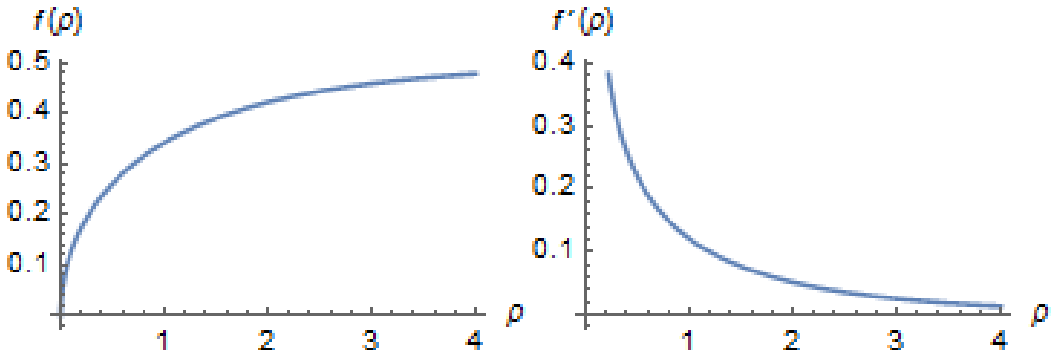


Figure 4.1: Function $f(\rho) = \frac{1}{\sqrt{\pi}} \int_0^{\sqrt{\frac{\rho}{2}}} e^{-x^2} dx$ and its derivative.

Then a linear cost function $c(t) = \kappa \cdot t$ implies $h(c) = f\left(\frac{c}{\kappa\sigma^2}\right)$. Figure 4.1 shows that $h(c)$ satisfies Assumption 4.3. Moreover, it is easy to see that Definition 4.2 is the analog of Definition 4.1 for the static case.

Optimizing (4.12) over $c > 0$, we get

$$c > 0: \ \eta h'(\eta c) Q = 1. \tag{4.14}$$

Treating the solution to (4.14) as a function of the overconfidence level η, we have

$$c'(\eta) = -\frac{h'(\eta c) + \eta c h''(\eta c)}{\eta^2 h''(\eta c)}. \tag{4.15}$$

A higher c means higher probability of the correct decision, $\frac{1}{2} + h(c)$. As we increase the level of overconfidence η, c increases if and only if $h'(\eta c) + \eta c h''(\eta c)$ is positive. From Assumption 4.3, $h'(\eta c)$ is always positive, while $\eta c h''(\eta c)$ is always negative. The term $h'(\eta c)$ corresponds to the **first force**: higher effective cost ηc increases the quality of the signal because h is increasing. The term $\eta c h''(\eta c)$ corresponds to the **second force**: higher effective cost ηc decreases the marginal benefit of information because h' is decreasing.

The total effect is captured by the behavior of function $xh'(x)$, which derivative is equal to $h'(x) + xh''(x)$. If it increases at point $x = \eta c(\eta)$, then increasing the level of overconfidence makes the final decision better. If it decreases, the final decision becomes worse with increased overconfidence.

Note the interpretation of $xh'(x)$, $x = \eta c$, as the marginal benefit of information. The first x corresponds to the first force, which increases function $xh'(x)$ as we

increase x through the level of overconfidence η. The second x corresponds to the second force, which decreases function $xh'(x)$.

Without making any additional assumptions, it is hard to say more about the behavior of $c(\eta)$. One interesting special case is when the following assumptions holds:

Assumption 4.4. *There exists $\hat{c} > 0$ such that $ch'(c)$ increases for $c < \hat{c}$ and it decreases for $c > \hat{c}$.*

This assumption says that when the amount of collected information is below some threshold, the second force is weaker than the first force, and vice versa. Recall that the second force comes from changing the benefit of already collected information. As we increase the amount of collected information is small, this force becomes stronger. On the other hand, the first force comes from changing the benefit of the marginal information piece and therefore it depends on the amount of collected information only through the non-stationary properties of the information flow. Assumption 4.4 says that the information flow is stationary enough to make sure that there is a unique threshold such that the second force prevails if and only if the amount of collected information is above that threshold.

Under Assumption 4.4, we can prove that there exists a *unique* optimal level of overconfidence η^* (which can be less than 1, which corresponds to underconfidence) such that more overconfidence is good below that level and it is bad for all η above η^*. Formally:

Theorem 4.3. *The probability of choosing the correct action is increasing in $\eta \in (0, \eta^*)$ and it is decreasing in $\eta \in (\eta^*, +\infty)$, where $\eta^* = \frac{1}{Qh'(\hat{c})}$.*

See Appendix C.3 for the proof.

Example (Normal distribution)

When $S \equiv X_t \sim \mathcal{N}(\mu_z t, \sigma^2 t)$, function $ch'(c) = \frac{c}{\kappa\sigma^2} f'\left(\frac{c}{\kappa\sigma^2}\right)$ satisfies Assumption 4.4 (see Figure 4.2). In that case $\eta^* = \frac{2\sqrt{2e\pi}\kappa\sigma^2}{Q}$.

Here is an example where Assumption 4.4 is violated.

Example (Binary distribution) Suppose $S \in \{I, G\}, \mathbb{P}[S = I \mid z = I] = \mathbb{P}[S = G \mid z = G] \geq \frac{1}{2}$. Then its quality $h = \mathbb{P}[S = z \mid z] - \frac{1}{2}$. This means that the binary distribution does not imply any specific form of function $h(c)$.

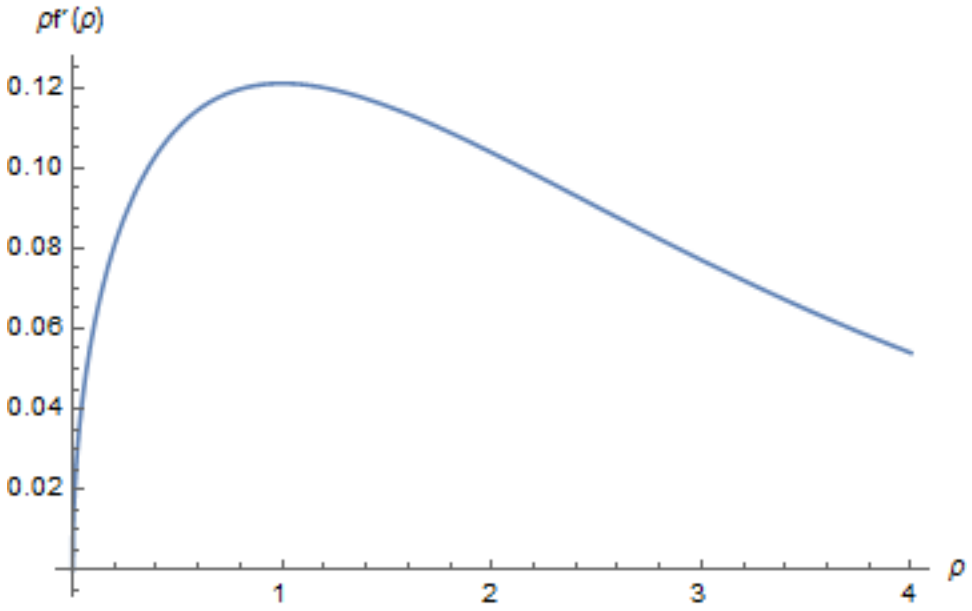


Figure 4.2: Function $\rho f'(\rho) = \frac{e^{-\frac{\rho}{2}}\sqrt{\rho}}{2\sqrt{2\pi}}$ defines how the marginal benefit of information changes with the level of overconfidence under normal distribution assumption.

1. Suppose the agent payment is some decreasing function of the variance of the random variable $\mathbf{1}(S = z)$, for example,

$$c = -\log\left(4\mathbb{P}\left[S = z \mid z\right]\left(1 - \mathbb{P}\left[S = z \mid z\right]\right)\right).$$

   Then $h(c) = \mathbb{P}\left[S = z \mid z\right] - \frac{1}{2} = \frac{1}{2}\sqrt{1 - e^{-c}}$ satisfies both Assumptions 4.3 and 4.4.

2. Function $h(c) = \left(2c + \frac{1}{c}\sin(c)\right)^{\frac{1}{4}} - 1$ satisfies Assumption 4.3 but not Assumption 4.4. See Figure 4.3.

Note that the optimal level of overconfidence is decreasing in Q. This means that overconfidence is bad when the benefit from choosing the correct action is high. Intuitively, when the benefit from choosing the correct action is very low, the rational agent collects very little information. A distortion in his incentives by increasing the perceived quality of information always has a positive effect. Indeed, an increase in the quality of already collected information (the second force) does not have a large effect since the amount of this information is small. So, overconfidence is good for low Q. On the other hand, when the benefit from choosing the correct action is very high, the rational agent collects a lot of information. This means that

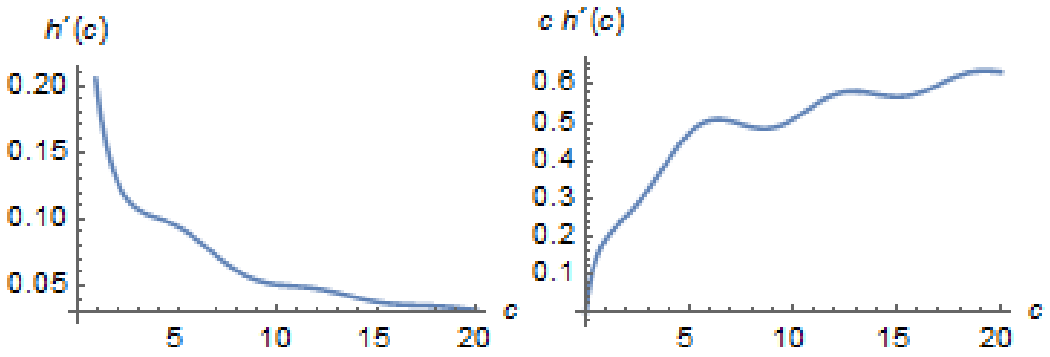


Figure 4.3: Functions $h'(c)$ and $ch'(c)$, where $h(c) = \left(2c + \frac{1}{c}\sin(c)\right)^{\frac{1}{4}} - 1$.

the second force has a lot of power as it works with a large amount of information. Consequently, overconfidence is bad in this case.

At the end of this section we give an example of a model when only the first force is active.

Binary Information Acquisition Decision

Suppose the decision maker can choose only between two values, $c \in \{0, \bar{c}\}$. This describes the situation when the agent simply has to decide whether to acquire information or not. In this case the optimality condition (4.14) should be changed to

$$c = \bar{c} \quad \Leftrightarrow \quad h(\eta\bar{c})Q > \bar{c}. \tag{4.16}$$

Definition 4.3. *The maximum cost the decision maker is ready to pay for the signal is called the* ***willingness to pay****.*

Given condition (4.16) and Assumption 4.3, the willingness to pay is equal to

$$c > 0: \quad h(\eta c)Q = c, \quad \eta h'(\eta c)Q < 1. \tag{4.17}$$

Treating the solution to (4.17) as a function of the overconfidence level η, we have $c'(\eta) > 0$.

Theorem 4.4. *The willingness to pay is increasing in* η*.*

See Appendix C.4 for the proof.

In this model, there is no "already collected information" since the choice is binary, either buy the signal or not. Thus, the second force is absent here and Theorem 4.4 result is driven by the first force.

4.4 Optimal Delegation

Welfare Implications of Commitment

So far we only compared the probability of the correct decision for different levels of overconfidence within the same model. In practice, the overconfidence level is hard to change, which means the comparison has a descriptive nature. In contrast to the overconfidence level, a model (an institution) can often be changed, which leads us to the following normative question. How does the probability of the correct decision change if we change the model, holding the level of overconfidence fixed? More precisely, in this section we compare the probability of the correct decision in the dynamic model with the one in the static model with normal distribution.

The probability of the correct decision in the dynamic model where the η-type decision maker observes (4.1) is given by

$$\Pi^D\left(\eta, \frac{Q}{\kappa\sigma^2}\right) = \frac{1}{1+\left(\frac{\lambda}{1-\lambda}\right)^{\frac{1}{\eta}}}, \tag{4.18}$$

where $\lambda \in (0, 0.5)$ solves

$$\frac{1-2\lambda}{2\lambda(1-\lambda)} - \log\left(\frac{\lambda}{1-\lambda}\right) = \frac{Q\eta}{\kappa\sigma^2}, \tag{4.19}$$

as follows from (4.9) and (4.10).

If the η-type decision maker commits to a stopping time at time 0, this probability is given by

$$\Pi^C\left(\eta, \frac{Q}{\kappa\sigma^2}\right) = \frac{1}{2} + \frac{1}{\sqrt{\pi}} \int_0^{\sqrt{\frac{\rho}{\eta}}} e^{-x^2} dx \tag{4.20}$$

where $\rho > 0$ solves

$$4e^{\rho}\sqrt{\pi\rho} = \frac{Q\eta}{\kappa\sigma^2}, \tag{4.21}$$

as follows from (4.13) and (4.14).

Note that in either model the probability of the correct decision depends only on two parameters, the overconfidence level η and the ratio $\frac{Q}{\kappa\sigma^2}$ (recall that Q is the bonus for the correct decision, κ is the per unit of time cost of information and σ^{-2} characterizes the objective precision of the information flow). The ratio $\frac{Q}{\kappa\sigma^2}$ characterizes how valuable the process X_t is for the rational decision maker. In other words, $\frac{Q}{\kappa\sigma^2}$ is the objective quality of the process X_t. In contrast, parameter η is a subjective characteristic of an overconfident decision maker.

Theorem 4.5 states that the commitment to a stopping time leads to a higher probability of the correct decision if and only if η is sufficiently high.

Theorem 4.5. *For any $\frac{Q}{\kappa\sigma^2}$, there is a unique level of overconfidence $\eta^{**} > 1$ such that $\Pi^D\left(\eta, \frac{Q}{\kappa\sigma^2}\right) > \Pi^C\left(\eta, \frac{Q}{\kappa\sigma^2}\right)$ for all $\eta \in (0, \eta^{**})$ and $\Pi^D\left(\eta, \frac{Q}{\kappa\sigma^2}\right) < \Pi^C\left(\eta, \frac{Q}{\kappa\sigma^2}\right)$ for all $\eta \in (\eta^{**}, +\infty)$.*

See Appendix C.5 for the proof.

The threshold η^{**} balances two effects. The first effect comes from the distortion that overconfidence brings to the model. In particular, it reflects the third force. As we discussed at the end of section 4.2, the higher the overconfidence level, the stronger the third force which makes the η-type decision maker stop sooner in the absence of commitment. Since this force decreases the probability of the correct decision, so as the absence of commitment. The second effect comes from the objective nature of commitment. From the η-type decision maker perspective, commitment is always bad since it limits his flexibility and therefore decreases his expected utility. In particular, commitment leads to a lower probability of the correct decision for the rational decision maker ($\eta = 1$). In sum, when we include commitment, the first effect increases the probability of the correct decision while the second effect lowers it. The higher η, the stronger the first effect, while the second effect says the same. Thus, the first effect overpowers the second effect for high levels of overconfidence.

Theorem 4.6 shows the comparative statics of the threshold η^{**} with respect to $\frac{Q}{\kappa\sigma^2}$.

Theorem 4.6. *The threshold $\eta^{**}\left(\frac{Q}{\kappa\sigma^2}\right)$ is decreasing in $\frac{Q}{\kappa\sigma^2} \in (0, +\infty)$ from $\frac{\pi^2}{4}$ to 1.*

See Appendix C.6 for the proof.

Parameter $\frac{Q}{\kappa\sigma^2}$ is connected with the second effect. The lower $\frac{Q}{\kappa\sigma^2}$, the more expensive information is (κ is higher), the more the rational decision maker values the flexibility that the dynamic model gives him, the stronger the second effect. Indeed, he does not lose much by committing to a higher stopping time when information is cheap. However, when information is expensive, increasing the stopping time becomes more expensive too, so as the commitment. Thus, as $\frac{Q}{\kappa\sigma^2}$ decreases, the region where commitment is good shrinks, or in other words, the threshold η^{**} increases.

Recall that $\eta^* = \frac{2\sqrt{2e\pi}\kappa\sigma^2}{Q}$ is the threshold such that the probability of choosing the correct action in the static model is increasing before it and it is increasing after (Theorem 4.3). Theorem 4.7 compares two thresholds, η^* and η^{**}, as functions of $\frac{Q}{\kappa\sigma^2}$.

Theorem 4.7. *There exists a unique $q \in (0, +\infty)$ such that $\eta^{**}\left(\frac{Q}{\kappa\sigma^2}\right) < \eta^*$ for all $\frac{Q}{\kappa\sigma^2} < q$ and $\eta^{**}\left(\frac{Q}{\kappa\sigma^2}\right) > \eta^*$ for all $\frac{Q}{\kappa\sigma^2} > q$.*

See Appendix C.7 for the proof.

When $\frac{Q}{\kappa\sigma^2}$ is high, the commitment is bad at the optimal level of overconfidence η^*. Intuitively, high Q means the overconfidence is bad, as discussed on page 64. Thus, the optimal level of overconfidence is low, which makes the first effect weak at this level. Despite the fact that a higher Q also weakens the second effect, it decreases the first effect even more because η^* reflects the best we can do in the static model. If we can control both the institute and the overconfidence level, commitment loses its appeal when overconfidence is not desirable. Thus, high Q means the dynamic model is better at η^*.

Optimal Contract

Section 4.4 assumes a binary control variable: we can choose either the model with commitment (static model) or without commitment (dynamic model). In this section we go further and assume we can also control the utility parameters, Q and R.

More precisely, we consider a principal-agent model, where a rational principal delegates the decision to an overconfident agent. We assume that the principal knows the agent's level of overconfidence η.[7] The principal's utility is characterized by the pair (Q_P, R_P), $Q_P > 0$, so that she gets $Q_P + R_P$ from a correct decision and R_P from an incorrect decision. Since R_P does not affect the principal's incentives, we can safely assume $R_P = 0$. For simplicity, we assume that the agent does not get anything from the decision other than what he gets from the principal. The principal offers the agent a contract (Q, R, M), $Q > 0$, $M \in \{0, 1\}$, where $M = 1$ corresponds to the dynamic information collection studied in section 4.2 and $M = 0$ corresponds to the static model with normal distribution where the agent has to commit to a stopping time.

[7] An asymmetric information case is out of the scope of this paper.

As in section 4.4, denote by $\Pi^D\left(\eta, \frac{Q}{\kappa\sigma^2}\right)$ the probability of the correct decision in the dynamic model and by $\Pi^C\left(\eta, \frac{Q}{\kappa\sigma^2}\right)$ the probability of the correct decision in the static model (see (4.18) and (4.20)).

Similarly, denote by Υ^D (Υ^C) the expected information collection cost in the dynamic (static) model.

For the dynamic model, we get

Lemma 4.1. $\Upsilon^D\left(\kappa\sigma^2, \eta, Q\right) = \frac{\kappa\sigma^2}{2}\frac{\left(\frac{1-\lambda}{\lambda}\right)^{\frac{1}{\eta}}-1}{\left(\frac{1-\lambda}{\lambda}\right)^{\frac{1}{\eta}}+1}\log\left(\left(\frac{1-\lambda}{\lambda}\right)^{\frac{1}{\eta}}\right)$, *where* $\lambda \in (0, 0.5)$ *solves (4.19).*

See Appendix C.8 for the proof.

From (4.13) and (4.14), we get $\Upsilon^C\left(\kappa\sigma^2, \eta, Q\right) = \frac{2\kappa\sigma^2\rho}{\eta}$, where $\rho > 0$ solves (4.21).

Denote

$$\Pi\left(\eta, \frac{Q}{\kappa\sigma^2}, M\right) = \Pi^D\left(\eta, \frac{Q}{\kappa\sigma^2}\right)\mathbf{1}\left(M=1\right) + \Pi^C\left(\eta, \frac{Q}{\kappa\sigma^2}\right)\mathbf{1}\left(M=0\right),$$

$$\Upsilon\left(\kappa\sigma^2, \eta, Q, M\right) = \Upsilon^D\left(\kappa\sigma^2, \eta, Q\right)\mathbf{1}\left(M=1\right) + \Upsilon^C\left(\kappa\sigma^2, \eta, Q\right)\mathbf{1}\left(M=0\right).$$

The contract (Q, R, M) is optimal if and only if it solves the following optimization problem:

$$\max_{Q>0, R, M\in\{0,1\}} \quad \Pi\left(\eta, \frac{Q}{\kappa\sigma^2}, M\right)\left(Q_P - Q\right) - R, \tag{4.22}$$

$$\text{s.t.} \quad \Pi\left(1, \frac{Q\eta}{\kappa\sigma^2}, M\right)Q + R - \Upsilon\left(\frac{\kappa\sigma^2}{\eta}, 1, Q, M\right) \geq 0. \tag{4.23}$$

Note that the constraint (4.23) follows from the definition of an η-type decision maker: he perceives the variance as being σ^2/η rather than σ^2 and he does not know he is overconfident.

Theorem 4.8. *For $\eta > 1$, the optimal contract (Q, R, M) is unique and has the following form: $M = 1$, $Q = +\infty$,*

$$R = \frac{\kappa\sigma^2}{2\eta}\left(1 - \log\left(\frac{\kappa\sigma^2}{2\eta}\right) + \log(Q)\right) - Q.$$

This contract gives $U^P = +\infty$ expected utility to the principal. The perceived expected utility of the agent is equal to his reservation utility, that is 0. The actual expected utility of the agent is $\Pi^D\left(\eta, \frac{Q}{\kappa\sigma^2}\right)Q + R - \Upsilon^D\left(\kappa\sigma^2, \eta, Q\right)$, which is equal to $U^A = -\infty$.

See Appendix C.9 for the proof. Underconfident agents ($0 < \eta < 1$) as well as the rational agent ($\eta = 1$) are not the main focus of our study; the optimal contract for these types can be found in Appendix C.10.

Theorem 4.8 shows that even a slight deviation from rationality to the direction of *over*confidence gives an enormous advantage to the principal who can exploit this little wedge between the perceived and the actual precision of information to the maximum extent, receiving an infinite expected utility. Intuitively, an overconfident agent thinks he has a better deal than he actually has because he overestimates the quality of the information source, and therefore, for a fixed effort, he thinks he has better chances of receiving Q than he actually has. By setting $Q = +\infty$, the principal basically eliminates the outcome when the agent does not discover the true state. However, the agent underestimates the expected time he spends collecting information, and the principal gets the difference. Given $Q = +\infty$, this difference is infinite, thus $U^P = +\infty$ and $U^A = -\infty$.

Note that the optimal contract sets $M = 1$, which means the principal does not want the agent to commit to the amount of information he collects ex-ante. However, if we constrain the set of feasible contracts by requiring $M = 0$, the principal can still find a contract that gives him $U^P = +\infty$ (see Appendix C.9 for the proof).

So far we assumed that the principal can take money from the agent, that is, R can be negative. This leads to a somewhat controversial result when by setting $R = -\infty$ and $Q = +\infty$, the principal can achive an infinite expected payoff. Now let's assume that R cannot be negative, so that contract (Q, R, M) is optimal if and only if it solves the following optimization problem:

$$\max_{Q\geq 0, R\geq 0, M\in\{0,1\}} \quad \Pi\left(\eta, \frac{Q}{\kappa\sigma^2}, M\right)(Q_P - Q) - R, \tag{4.24}$$

$$\text{s.t.} \quad \Pi\left(1, \frac{Q\eta}{\kappa\sigma^2}, M\right)Q + R - \Upsilon\left(\frac{\kappa\sigma^2}{\eta}, 1, Q, M\right) \geq 0. \tag{4.25}$$

Theorem 4.9. *For $\eta > 0$, the optimal contract (Q, R, M) is essentially unique*[8] *and*

[8] Non-uniqueness appears in the following cases: (1) when $Q = 0$ the value of M is irrelevant, (2) when $\frac{Q_P}{\kappa\sigma^2} = q(\eta)$ both $(Q^D, M = 1)$ and $(Q^C, M = 0)$ give exactly the same expected payoff for the principal.

has the following form[9]*:* $R = 0$*,*

$$Q = \begin{cases} 0, & \frac{Q_P}{\kappa\sigma^2} \leq \min\left\{4, \frac{2\pi}{\sqrt{\eta}}\right\}, \\ Q^D, & \min\left\{4, \frac{2\pi}{\sqrt{\eta}}\right\} < \frac{Q_P}{\kappa\sigma^2} < q(\eta), \\ Q^C, & q(\eta) < \frac{Q_P}{\kappa\sigma^2}, \end{cases} \quad M = \begin{cases} 1, & \min\left\{4, \frac{2\pi}{\sqrt{\eta}}\right\} < \frac{Q_P}{\kappa\sigma^2} < q(\eta), \\ 0, & q(\eta) < \frac{Q_P}{\kappa\sigma^2}, \end{cases}$$

where $q(\eta) = \begin{cases} \frac{2\pi}{\sqrt{\eta}}, & \eta \geq \frac{\pi^2}{4}, \\ \in \left(\frac{2\pi}{\sqrt{\eta}}, +\infty\right), & 1 < \eta < \frac{\pi^2}{4}, \\ +\infty, & 0 < \eta \leq 1 \end{cases}$ *is continuous and strictly decreasing*

for $\eta \geq 1$, $Q^D = \frac{\kappa\sigma^2}{\eta}\left(\frac{1-2\lambda}{2\lambda(1-\lambda)} - \log\left(\frac{\lambda}{1-\lambda}\right)\right)$ *and* $\lambda \in (0, 0.5)$ *solves*

$$\frac{\eta Q_P}{\kappa\sigma^2} = \frac{1-2\lambda}{2\lambda(1-\lambda)} - \log\left(\frac{\lambda}{1-\lambda}\right) + \frac{\eta}{2\lambda(1-\lambda)}\left(1 + \left(\frac{\lambda}{1-\lambda}\right)^{-\frac{1}{\eta}}\right), \quad (4.26)$$

$Q^C = \frac{\kappa\sigma^2}{\eta} 4e^{\rho}\sqrt{\pi\rho}$ *and* $\rho > 0$ *solves*

$$\frac{\eta Q_P}{\kappa\sigma^2} = 4e^{\rho}\sqrt{\pi}\left(\sqrt{\rho} + e^{\frac{\rho}{\eta}}\sqrt{\eta}(1+2\rho)\int_{-\infty}^{\sqrt{\rho/\eta}} e^{-x^2}dx\right). \quad (4.27)$$

See Appendix C.11 for the proof.

Basically, Theorem 4.9 formalizes the same idea as does Theorem 4.5: when η is low, the dynamic contract is better, while if η is high, the static contract is better. More precisely, Theorem 4.9 follows the following logic. When $R \geq 0$, the constraint (4.25) is not binding, and optimal (Q, M) maximize $\Pi\left(\eta, \frac{Q}{\kappa\sigma^2}, M\right)(Q_P - Q)$. By Theorem 4.5, $\Pi^D\left(\eta, \frac{Q}{\kappa\sigma^2}\right) > \Pi^C\left(\eta, \frac{Q}{\kappa\sigma^2}\right)$ if and only if $\eta < \eta^{**}\left(\frac{Q}{\kappa\sigma^2}\right)$. Thus, for a fixed Q, $M = 1$ is optimal if and only if $\eta < \eta^{**}\left(\frac{Q}{\kappa\sigma^2}\right)$. By Theorem 4.6, $\eta^{**}\left(\frac{Q}{\kappa\sigma^2}\right)$ is decreasing from $\frac{\pi^2}{4}$ to 1. Thus:

1. If $\eta < 1$, then $\eta < \eta^{**}\left(\frac{Q}{\kappa\sigma^2}\right)$ for all Q and therefore $M = 1$ is optimal.
2. If $\eta > \frac{\pi^2}{4}$, then $\eta > \eta^{**}\left(\frac{Q}{\kappa\sigma^2}\right)$ for all Q and therefore $M = 0$ is optimal.
3. If $1 < \eta < \frac{\pi^2}{4}$, then $M = 1$ is optimal for small Q and $M = 0$ is optimal for high Q. Since $\Pi\left(\eta, \frac{Q}{\kappa\sigma^2}, M\right)$ is increasing in Q, there is an optimal Q

[9] The definitions of Q^D and Q^C imply that the optimal Q is continuous at point $\frac{Q_P}{\kappa\sigma^2} = \min\left\{4, \frac{2\pi}{\sqrt{\eta}}\right\}$.

for each M (Q^D when $M = 1$ and Q^C when $M = 0$) that balances two terms, $\Pi\left(\eta, \frac{Q}{\kappa\sigma^2}, M\right)$ and $1 - \frac{Q}{Q_P}$. As Q_P is increasing, both Q^D and Q^C are increasing. Thus, for low Q_P, both Q^D and Q^C are low enough so that $M = 1$ is optimal, while $M = 0$ is optimal for high Q_P.

The threshold $q(\eta)$ is decreasing, which implies that $M = 1$ is optimal for low η, while $M = 0$ is optimal for high η.

Theorem 4.10. *When $M = 1$ is optimal, the principal's expected utility U^P and the social welfare ($U^P + U^A$, where U^A is the actual expected utility of the agent) are both decreasing in η. When $M = 0$ is optimal, U^P is increasing if $\eta < \hat{\eta}_P\left(\frac{Q_P}{\kappa\sigma^2}\right)$ and it is decreasing otherwise; $U^P + U^A$ is increasing if $\eta < \hat{\eta}_{PA}\left(\frac{Q_P}{\kappa\sigma^2}\right)$ and it is decreasing otherwise. Moreover, $\hat{\eta}_{PA}\left(\frac{Q_P}{\kappa\sigma^2}\right) < \hat{\eta}_P\left(\frac{Q_P}{\kappa\sigma^2}\right)$ for all $\frac{Q_P}{\kappa\sigma^2}$.*

See Appendix C.12 for the proof.

Theorem 4.10 echoes Theorems 4.2 and 4.3. Theorems 4.10 and 4.2 both state that when there is no commitment to the stopping time, increasing the level of overconfidence is always bad. Similarly, Theorems 4.10 and 4.3 both state that in the static model, there is an optimal level of overconfidence. The difference is in the criteria function for "good" and "bad" notions. While Theorems 4.2 and 4.3 imply maximization of the probability of the correct decision, disregarding both the cost of information and the benefit Q from the correct decision, Theorem 4.10 takes it all into account. More precisely, it considers the matter from the principal perspective and from the social welfare perspective (when both the principal and the agent's objective expected utility is taken into account). Notably, the optimal level of the agent's overconfidence is lower from the social welfare perspective than from the principal's perspective. The reason for this is similar to the intuition behind Theorem 4.8: the principal can exploit the wedge between the perceived and the actual precision of information. This exploitation mitigates the negative effect from the overconfidence.

4.5 Conclusion

We presented three forces that shape the effect that overconfidence has on the quality of the final decision, or equivalently, on the amount of information collected in equilibrium. Two aspects of the information collection process are crucial to understand the effect that misperception of information quality has on information investment in a particular scenario. First, whether the information investment choice

space is binary (collect or not) or continuous (how much to collect). In the latter case, there is a trade-off between increased quality of already collected information, which pushes the overconfident agent to collect less information, and increased quality of the marginal piece of information, which pushes him to collect more information. Second, if the information investment choice is continuous, whether information has been collected all at once or not. In the latter case, misperception of information quality creates a systematic bias between how much information the decision maker expects to collect and how much information he actually collects. An overconfident agent overestimates the expected amount of information he is going to collect in the future.

The normative implication from our analysis is the conditions under which the commitment to the amount of information the decision maker is going to collect is optimal. It is optimal when the level of overconfidence is high enough.

BIBLIOGRAPHY

Abdulkadiroğlu, A., N. Agarwal, and P. A. Pathak (2017): "The welfare effects of coordinated assignment: evidence from the New York City high school match," *American Economic Review*, 107(12), 3635–89. 3

Abdulkadiroğlu, A., P. A. Pathak, A. E. Roth, and T. Sönmez (2005): "The Boston public school match," *American Economic Review*, 95(2), 368–371. 3

Agarwal, N. (2015): "An Empirical Model of the Medical Match," *American Economic Review*, 105(7), 1939–1978. 32

Alcalde, J., and S. Barberà (1994): "Top dominance and the possibility of strategy-proof stable solutions to matching problems," *Economic Theory*, 4(3), 417–435. 26

Alpert, M., and H. Raiffa (1982): "A progress report on the training of probability assessors," in *Judgment under Uncertainty: Heuristics and Biases*, ed. by D. Kahneman, P. Slovic, and A. Tversky, pp. 294–305. Cambridge University Press, Cambridge. 54

Ashlagi, I., Y. Kanoria, and J. D. Leshno (2017): "Unbalanced random matching markets: The stark effect of competition," *Journal of Political Economy*, 125(1), 69–98. 28, 30

Balinski, M. L., and T. Sönmez (1999): "A tale of two mechanisms: student placement," *Journal of Economic Theory*, 84(1), 73–94. 26

Barber, B. M., and T. Odean (2001): "Boys will be boys: Gender, overconfidence, and common stock investment," *Quarterly Journal of Economics*, 116(1), 261–292. 53

Barberà, S., and B. Dutta (1995): "Protective behavior in matching models," *Games and Economic Behavior*, 8(2), 281–296. 8, 25, 26

Becker, G. S. (1973): "A theory of marriage: Part I," *Journal of Political Economy*, 81(4), 813–846. 30, 46

Bell, D. E. (1982): "Regret in decision making under uncertainty," *Operations Research*, 30(5), 961–981. 6, 12, 23

Benoît, J.-P., and J. Dubra (2011): "Apparent overconfidence," *Econometrica*, 79(5), 1591–1625. 50

Bergemann, D., and S. Morris (2008): "Ex-post implementation," *Games and Economic Behavior*, 63(2), 527–566. 25

Berner, E. S., and M. L. Graber (2008): "Overconfidence as a cause of diagnostic error in medicine," *The American Journal of Medicine*, 121(5), S2–S23. 53

Bikhchandani, S. (2017): "Stability with one-sided incomplete information," *Journal of Economic Theory*, 168, 372 – 399. 31

Bikhchandani, S., and U. Segal (2011): "Transitive regret," *Theoretical Economics*, 6(1), 95–108. 24

Birkhoff, G. (1967): *Lattice Theory*. American Mathematical Society. 86

Bronfman, S., A. Hassidim, A. Afek, A. Romm, R. Shreberk, A. Hassidim, and A. Massler (2015): "Assigning Israeli medical graduates to internships," *Israel Journal of Health Policy Research*, 4(6), 1–7. 3

Chakraborty, A., A. Citanna, and M. Ostrovsky (2010): "Two sided matching with interdependent values," *Journal of Economic Theory*, 145, 85–105. 31

Chan, J., A. Lizzeri, W. Suen, and L. Yariv (2017): "Deliberating collective decisions," *The Review of Economic Studies*, 85(2), 929–963. 54

Chen, P., M. Egesdal, M. Pycia, and M. B. Yenmez (2014): "Quantile Stable Mechanisms," *Working Paper*, pp. 1–22. 7, 10, 11, 26, 81

——— (2016): "Manipulability of Stable Mechanisms," *American Economic Journal: Microeconomics*, 8(2), 202–14. 7, 24

Chiappori, P.-A., and B. Salanié (2016): "The econometrics of matching models," *Journal of Economic Literature*, 54(3), 832–61. 32

Chuang, W.-I., and B.-S. Lee (2006): "An empirical evaluation of the overconfidence hypothesis," *Journal of Banking & Finance*, 30(9), 2489–2515. 53

Dénes, J., and A. D. Keedwell (1991): *Latin squares: New developments in the theory and applications*, vol. 46. Elsevier. 81

Dubins, L. E., and D. Freedman (1981): "Machiavelli and the Gale-Shapley Algorithm," *American Mathematical Monthly*, 88(7), 485–494. 24, 38

Dubra, J. (2004): "Optimism and overconfidence in search," *Review of Economic Dynamics*, 7(1), 198–218. 54

Echenique, F., and L. Yariv (2011): "An Experimental Study of Decentralized Matching," *Working Paper*. 7, 11

Ehlers, L. (2008): "Truncation Strategies in Matching Markets," *Mathematics of Operations Research*, 33(2), 327–335. 8, 25

Ehlers, L., and J. Massó (2007): "Incomplete information and singleton cores in matching markets," *Journal of Economic Theory*, 136(1), 587–600. 8, 31

——— (2015): "Matching markets under (in)complete information," *Journal of Economic Theory*, 157, 295–314. 8, 32

Ergin, H. (2002): "Efficient resource allocation on the basis of priorities," *Econometrica*, 70(6), 2489–2497. 40, 44

Gale, D., and L. S. Shapley (1962): "College admissions and the stability of marriage," *American Mathematical Monthly*, 69(1), 9–15. iii, 3, 9, 25, 27

Gale, D., and M. Sotomayor (1985): "Ms . Machiavelli and the Stable Matching," *American Mathematical Monthly*, 92(4), 261–268. 38, 45

Gilovich, T., and V. H. Medvec (1995): "The experience of regret: what, when, and why.," *Psychological Review*, 102(2), 379–395. 6, 12

Glaeser, E. L., and C. R. Sunstein (2009): "Extremism and social learning," *Journal of Legal Analysis*, 1(1), 263–324. 54

Goetzmann, W., and S. Huang (2015): "Momentum in Imperial Russia," Discussion paper, National Bureau of Economic Research. 53

Heidhues, P., B. Koszegi, and P. Strack (2017): "Unrealistic expectations and misguided learning," *Working Paper*. 51

Hsieh, Y.-W. (2011): "Understanding Mate Preferences from Two-Sided Matching Markets: Identification, Estimation and Policy Analysis," *Working Paper*, pp. 1–65. 32

Immorlica, N., and M. Mahdian (2005): "Marriage, honesty, and stability," in *Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 53–62. Society for Industrial and Applied Mathematics. 7, 25, 28, 30

Johnson, D. D. (2009): *Overconfidence and war*. Harvard University Press. 53

Kesten, O. (2006): "On two competing mechanisms for priority-based allocation problems," *Journal of Economic Theory*, 127(1), 155–171. 40, 44

Klaus, B., and F. Klijn (2006): "Median stable matching for college admissions," *International Journal of Game Theory*, 34(1), 1–11. 7, 10, 26

Klayman, J., J. B. Soll, C. Gonzalez-Vallejo, and S. Barlas (1999): "Overconfidence: It depends on how, what, and whom you ask," *Organizational Behavior and Human Decision Processes*, 79(3), 216–247. 54

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2017): "Human decisions and machine predictions," *Quarterly Journal of Economics*, 133(1), 237–293. 54

Kojima, F., and M. Manea (2010): "Axioms for Deferred Acceptance," *Econometrica*, 78(2), 633–653. 8, 25

Kojima, F., and P. A. Pathak (2009): "Incentives and stability in large two-sided markets," *American Economic Review*, 99(3), 608–627. 8, 25, 28, 30

Kyle, A. S., A. A. Obizhaeva, and Y. Wang (2017): "Smooth trading with overconfidence and market power," *The Review of Economic Studies*, 85(1), 611–662. 53, 55

Ledyard, J. O. (1977): "Incentive compatible behavior in core-selecting organizations," *Econometrica*, 45(7), 1607–1623. 25

Lee, S. (2017): "Incentive Compatibility of Large Centralized Matching Markets," *The Review of Economic Studies*, 84(1), 444–463. 8, 25, 30

Levy, G., and R. Razin (2015): "Correlation neglect, voting behavior, and information aggregation," *American Economic Review*, 105(4), 1634–45. 54

Liu, Q., G. J. Mailath, A. Postlewaite, and L. Samuelson (2014): "Stable Matching With Incomplete Information," *Econometrica*, 82(2), 541–587. 31, 37

Loomes, G., and R. Sugden (1982): "Regret Theory: An Alternative of Rational Choice Under Uncertainty," *The Economic Journal*, 92(368), 805–824. 6, 12

——— (1987): "Some implications of a more general form of regret theory," *Journal of Economic Theory*, 41(2), 270–287. 6, 12

Martinelli, C. (2006): "Would rational voters acquire costly information?," *Journal of Economic Theory*, 129(1), 225–251. 54

McVitie, D. G., and L. B. Wilson (1970): "Stable marriage assignment for unequal sets," *BIT Numerical Mathematics*, 10(3), 295–309. 10, 28, 35, 45

Moore, D. A., and P. J. Healy (2008): "The trouble with overconfidence.," *Psychological Review*, 115(2), 502. 50, 51, 54

Niederle, M., and L. Yariv (2009): "Decentralized matching with aligned preferences," *NBER Working paper series*, 14840, 41. 31

Ortoleva, P., and E. Snowberg (2015): "Overconfidence in political behavior," *American Economic Review*, 105(2), 504–35. 50, 53, 54

Oskamp, S. (1965): "Overconfidence in case-study judgments.," *Journal of Consulting Psychology*, 29(3), 261. 54

Pathak, P. A., and T. Sönmez (2013): "School admissions reform in Chicago and England: Comparing mechanisms by their vulnerability to manipulation," *American Economic Review*, 103(1), 80–106. 7, 24

Roth, A. E. (1982): "The Economics of Matching: Stability and Incentives," *Mathematics of Operations Research*, 7(4), 617–628. 15, 16, 24

——— (1984): "Misrepresentation and stability in the marriage problem," *Journal of Economic Theory*, 34(2), 383–387. 38, 45

——— (1989): "Two sided matching with incomplete information about others preferences," *Games and Economic Behavior*, 1, 191–209. 28, 31

——— (1991): "A Natural Experiment in the Organization of Entry-Level Labor Markets Regional markets for new physicians and surgeons in the United Kingdom," *American Economic Review*, 81(3), 415–440. 3

——— (2008): "Deferred acceptance algorithms: history, theory, practice, and open questions," *International Journal of Game Theory*, 36(3-4), 537–569. 3

Roth, A. E., and E. Peranson (1999): "The Redesign of the Matching Market for American Physicians: Some Engineering Aspects of Economic Design," *American Economic Review*, 89(4), 748–780. 7, 25, 29, 30

Roth, A. E., and U. G. Rothblum (1999): "Truncation strategies in matching markets - in search of advice for participants," *Econometrica*, 67(1), 21–43. 4, 8, 16, 25, 31

Roth, A. E., and M. Sotomayor (1990): *Two-sided matching: A study in game theoretic modeling and analysis*. Cambridge University Press. 8, 27, 30, 33, 80, 85

Roth, A. E., and J. H. Vande Vate (1991): "Incentives in Two-Sided Matching with Random Stable Mechanisms," *Economic Theory*, 1(1), 31–44. 16, 35

Scheinkman, J. A., and W. Xiong (2003): "Overconfidence and speculative bubbles," *Journal of Political Economy*, 111(6), 1183–1220. 53, 55

Shiryaev, A. N. (2007): *Optimal stopping rules*, vol. 8. Springer Science & Business Media. 56

Soll, J. B., and J. Klayman (2004): "Overconfidence in interval estimates.," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(2), 299. 54

Stoye, J. (2011): "Axioms for minimax regret choice correspondences," *Journal of Economic Theory*, 146(6), 2226–2251. 23

Teo, C.-P., and J. Sethuraman (1998): "The Geometry of Fractional Stable Matchings and Its Applications," *Mathematics of Operations Research*, 23(4), 874–891. 10, 26

Thurber, E. G. (2002): "Concerning the maximum number of stable matchings in the stable marriage problem," *Discrete Mathematics*, 248(1-3), 195–219. 81, 82

Wilson, R. (1987): "Game-Theoretic Analysis of Trading Processes," in *Advances in Economic Theory: Fifth World Congress*, ed. by T. Bewley, pp. 33–70. Cambridge University Press, Cambridge. 11, 25

A p p e n d i x A

APPENDICES TO CHAPTER 2

A.1 Proofs

Sufficiency

Theorem 2.1 (Sufficiency). In any non-extreme q-quantile matching mechanism truth-telling is not regret-free; that is $\forall q \in (0,1)$ we can find a market $(M, W, \succ) \in (2^M, 2^W, \mathcal{P})$ such that $\exists i \in N = M \cup W$ and a $\mu = \phi^q(\succ)$ where i regrets $\succ_i$ at μ through some $\succ'_i \in \mathcal{P}_i$, formally

$$(\forall \succ_{-i} \in \mathcal{I})\,[\phi^q(\succ'_i, \succ_{-i}) \succeq_i \mu] \tag{A.1}$$

$$(\exists \tilde{\succ}_{-i} \in \mathcal{I})\,[\phi^q(\succ'_i, \tilde{\succ}_{-i}) \succ_i \mu] \tag{A.2}$$

Proof. Fix $q \in (0,1)$. Without loss of generality, consider $i = m_1$. Define the function $\succ'_{m_1} : \mathcal{P}_i \to \mathcal{P}_i$

$$\succ'_{m_1} : \begin{cases} w \succ'_{m_1} w' \iff w \succ_{m_1} w' & \forall (w, w') \in W^2 \\ w \succeq'_{m_1} m_1 \iff w \succeq_{m_1} \phi^q(\succ_{m_1}, \succ_{-m_1})(m_1) & \forall w \in W \end{cases}$$

Although $\succ'_{m_1}$ is a function of $\succ$ we suppress it from the notation.

Remark. $A(\succ'_{m_1}) \subseteq A(\succ_{m_1})$

(Lemma 2.1). $S(\succ'_{m_1}, \succ_{-m_1}) = \{\mu \in S(\succ_{m_1}, \succ_{-m_1}) : \mu \succeq_{m_1} \phi^q(\succ)\} \quad \forall \succ \in \mathcal{P}$

Proof of Lemma. We prove it in two steps,

Claim A.1. $S(\succ'_{m_1}, \succ_{-m_1}) \subseteq S(\succ_{m_1}, \succ_{-m_1})$

Proof of claim. Suppose not, then there exists $\mu \in S(\succ'_{m_1}, \succ_{-m_1})$ and $\mu \notin S(\succ_{m_1}, \succ_{-m_1})$. It must be the case that μ is either blocked by a pair or by an individual. If μ is not individually rational under $(\succ_{m_1}, \succ_{-m_1})$ then it is not individually rational under $(\succ'_{m_1}, \succ_{-m_1})$. This is straightfoward since $\succ'_{-m_1} = \succ_{-m_1}$, and for m_1 it is a consequence of $A(\succ'_{m_1}) \subseteq A(\succ_{m_1})$ by construction of $\succ'_{m_1}$. Suppose it is blocked by a pair $(m_j, w) : m_j \neq m_1$, then since $\succ'_{-m_1} = \succ_{-m_1}$, (m_j, w) also blocks μ under preference profile $(\succ'_{m_1}, \succ_{-m_1})$. Then it has to be the case that the blocking pair involves m_1. If $\mu(m_1) = w'$, then it must hold that $w' \succ_{m_1} w$ and $w \succ'_{m_1} w'$, but

it contradicts the construction of $\succ'_{m_1}$ since it does not permute binary relations that do not involve alternative m_1, which denotes remaining single. Lastly, consider the case $\mu(m_1) = m_1$ then it means m_1 must be single in every stable matching under $(\succ'_{m_1}, \succ_{-m_1})$. If μ is blocked by a pair (m_1, w), then it must be that $\forall \mu'' \in S(\succ_{m_1}, \succ_{-m_1}) : \mu''(m_1) \in W$.[1] Next we note that $A(\succ'_{m_1}) \supseteq \{\mu_M(\succ_{m_1}, \succ_{-m_1})\}$ since by construction of $\succ'_{m_1}$ it follows that $\forall w \in W : w \succeq_{m_1} \phi_q(\succ_{m_1}, \succ_{-m_1})$ it holds that $w \in A(\succ'_{m_1})$ (for any quantile stable mechanism $q \in (0, 1)$). Suppose $\mu_M(\succ_{m_1}, \succ_{-m_1}) = m_1$ contradicts hypothesis (same argument as in the footnote). On the other hand, if $\mu_M(\succ_{m_1}, \succ_{-m_1}) \neq m_1$ then $\mu_M(\succ'_{m_1}, \succ_{-m_1}) \neq m_1$, this holds since the M-proposing DA algorithm follows the same steps; if at some point $\mu_M(\succ'_{m_1}, \succ_{-m_1})(m_1)$ rejected him at any step, then it should have rejected him in $\mu_M(\succ_{m_1}, \succ_{-m_1})(m_1)$. □

By the construction of $\succ'_{m_1}$ if follows that $S(\succ'_{m_1}, \succ_{-m_1}) \subseteq \{\mu \succeq_{m_1} \phi^q(\succ)\}$.

Claim A.2. *$\mu \in S(\succ_{m_1}, \succ_{-m_1})$ and $\mu \succeq_{m_1} \phi(\succ_{m_1}, \succ_{-m_1}) \implies \mu \in S(\succ'_{m_1}, \succ_{-m_1})$.*

Proof of claim. Suppose not, so $\mu \notin S(\succ'_{m_1})$, then either it is not individual rational or it is blocked by a pair. If $j \succ_j \mu(j)$ for $j \neq \{m_1\}$ then $\mu \notin S(\succ_{m_1})$, on the other hand if $m_1 \succ'_{m_1} \mu(i)$ then $\mu \not\succeq_{m_1} \phi^q(\succ) \succeq_{m_1} m_1$. If it is blocked by a pair (m_j, w) $j \neq \{1\}$ then they are also a blocking pair to μ under $(\succ_{m_1}, \succ_{-m_1})$. Lastly consider the blocking pairs (m_1, w). If $\mu(m_1) \neq m_1$, since by construction $\succ'_{m_1}$ does not change the binary relations not involving the alternative of being single m_1, they would also be a blocking pair to (m_1, w). Lastly, if $\mu(m_1) = m_1$, $w \succ'_{m_1} m_1$ and since $A(\succ'_{m_1}) \subseteq A(\succ_{m_1})$, $w \succ_{m_1} m_1$ which contradicts $\mu \in S(\succ_{m_1}, \succ_{-m_1})$. □

□

[1] Suppose not, so that $\mu''(m_1) = m_1 \; \forall \mu'' \in S(\succ_{m_1}, \succ_{-m_1})$; since (m_1, w) is the blocking pair, it follows that $w \succ_{m_1} m_1$ and $m_1 \succ_w \mu(w)$. Now let

$$\tilde{\mu} = \begin{cases} \mu(j) & \forall j \notin \{m_1, w, \mu(w)\} \\ w & \text{for } m_1 \\ m_1 & \text{for } w \\ \mu(w) & \text{for } \mu(w) \end{cases}$$

$\tilde{\mu}$ is an individually rational matching. Either, $\tilde{\mu}$ is stable, in which case $\tilde{\mu}(m_1) = w \in W$ which contradicts $\nexists \mu \in S(\succ_{m_1}, \succ_{-m_1}) : \mu(m_1) \neq m_1$, or $\tilde{\mu}$ is unstable. If the latter holds, still it must be individually rational (since it was stable for $(\succ'_{m_1}, \succ_{-m_1})$), then by the strong stability property (Roth and Sotomayor, 1990, Theorem 3.4, p. 56) there exists $\bar{\mu} \in S(\succ_{m_1}, \succ_{-m_1}) : \bar{\mu} \succeq_{m_1} \tilde{\mu}$ and $\bar{\mu} \succeq_w \tilde{\mu}$. Since $\tilde{\mu}(m_1) = w$ then $\bar{\mu}(m_1) \in W$ which is a contradiction.

Remark A.1. for any $q \in (0,1)$, $\phi^q(\succ'_{m_1}, \succ_{-m_1}) \succeq_{m_1} \phi^W(\succ'_{m_1}, \succ_{-m_1}) = \phi^q(\succ_{m_1}, \succ_{-m_1})$.

Theorem (Theorem 4 in Chen, Egesdal, Pycia, and Yenmez (2014))**.** *For any $q, q' \in (0,1] : q \neq q'$ there exists a matching market such that $\phi^q(\cdot)$ is different than $\phi^{q'}(\cdot)$.*

The key for this result is to find a market with k large enough, $k = |S(\succ)|$, such that $k(q' - q) > 1$, that is a market large enough (in terms of the number of stable matches) such that the non-extreme quantile mechanism and the extreme one result in different matches. Note that a priori we need a little more, since it could be the case that the matches are different but m_1 is matched to the same partner in both. Putting together the remark the theorem, and taking into account the construction of $\succ'_{m_1}$ we get the following corollary,

Corollary A.1. *For fixed $q \in (0,1)$, denote $k^\star(q) = \min\{k \in \mathbb{N} : k(1-q) \geq 1\}$. If $\exists(\succ'_{m_1}, \succ_{-m_1}) \in \mathcal{P} : |S(\succ'_{m_1}, \succ_{-m_1})| \geq k^\star(q)$ and $\mu(m_1) \neq \mu'(m_1)$ for all $\mu, \mu' \in S(\succ'_{m_1}, \succ_{-m_1})$ then $\phi^q(\succ'_{m_1}, \succ_{-m_1}) \succ_{m_1} \phi^W(\succ'_{m_1}, \succ_{-m_1}) = \phi^q(\succ_{m_1}, \succ_{-m_1})$*

Consider the following preferences: Let $\hat{k} : \inf\{k \in \mathbb{N} : \lceil kq \rceil \geq k^\star(q)\}$

$$
\begin{array}{ll}
m_1 : w_1 \succ w_2 \succ \cdots \succ w_{\hat{k}-1} \succ w_{\hat{k}} & w_{\hat{k}} : m_1 \succ m_2 \succ \cdots \succ m_{\hat{k}-1} \succ m_{\hat{k}} \quad (\star) \\
m_2 : w_2 \succ w_3 \succ \cdots \succ w_{\hat{k}} \succ w_1 & w_{\hat{k}-1} : m_{\hat{k}} \succ m_1 \succ \cdots \succ m_{\hat{k}-2} \succ m_{\hat{k}-1} \\
\vdots & \vdots \\
m_{\hat{k}} : w_{\hat{k}} \succ w_1 \succ \cdots \succ w_{\hat{k}-2} \succ w_{\hat{k}-1} & w_1 : m_2 \succ m_3 \succ \cdots \succ m_{\hat{k}} \succ m_1
\end{array}
$$

This is a fairly standard way of generating a matching with $|S(\succ_{m_1}, \succ_{-m_1})| = \hat{k}$ (see Thurber (2002) and Chen, Egesdal, Pycia, and Yenmez (2014)) namely, to construct preferences such that they form a Latin square marriage of order $\hat{k}$ (see Claim A.3).[2] But moreover, each individual gets a different partner in each stable matching. As a

[2]Dénes and Keedwell (1991): A Latin square of order n is an $n \times n$ matrix L whose entries are taken from a set S of n symbols and which has the property that every symbol from S occurs exactly once in each row and exactly once in each column.

consequence of Lemma 2.1, if $|S(\succ_{m_1}, \succ_{-m_1})| = \hat{k}(q)$ and $\{\hat{k} \in \mathbb{N} : \lceil \hat{k}q \rceil \geq k^\star(q)\}$ it follows that $|S(\succ'_{m_1}, \succ_{-m_1})| \geq k^\star(q)$, then corollary A.1 applies and we get that

$$\phi^q(\succ'_{m_1}, \succ_{-m_1}) \succ_{m_1} \phi^q(\succ_{m_1}, \succ_{-m_1})$$

which contradicts truth being regret-free.

Consequently, for any non-extreme ($q \in (0,1)$) $q-$quantile stable matching mechanism, we can find a market $(M, W, \succ)$ where an agent $i \in N$ regrets truth $\succ_i$ through some other report $\succ'_i$.[3,4] □

Claim A.3. *$|S(\succ)| = \hat{k}$ where preference profile $\succ$ is described in ($\star$).*

Proof. The preference profile described in ($\star$) yields an associated $\hat{k} \times \hat{k}$ ranking matrix A where each entry a_{ij} denotes man i's rank of woman j according to $\succ_i$, e.g. suppose man i's favorite woman is t then $a_{it} = 1$; given the construction in ($\star$) such entries are enough to characterize the preferences of both men and women, since they are perfectly opposed, that is the sum of man i and woman j about each other is exactly $\hat{k} + 1$.[5]

$m_1:$	w_1	w_2	w_3	$\dots$	w_{k-1}	w_k
$m_2:$	w_2	w_3	w_4	$\dots$	w_k	w_1
$\vdots$				$\dots$		
$m_k:$	w_k	w_1	w_2	$\dots$	w_{k-2}	w_{k-1}

Preferences of Men according to ($\star$)

	w_1	w_2	w_3	$\dots$	w_{k-1}	w_k
$m_1:$	1	2	3	$\dots$	$k-1$	k
$m_2:$	k	1	2	$\dots$	$k-2$	$k-1$
$\vdots$			$\dots$			
$m_k:$	2	3	4	$\dots$	k	1

Associated Ranking Matrix

Figure A.1

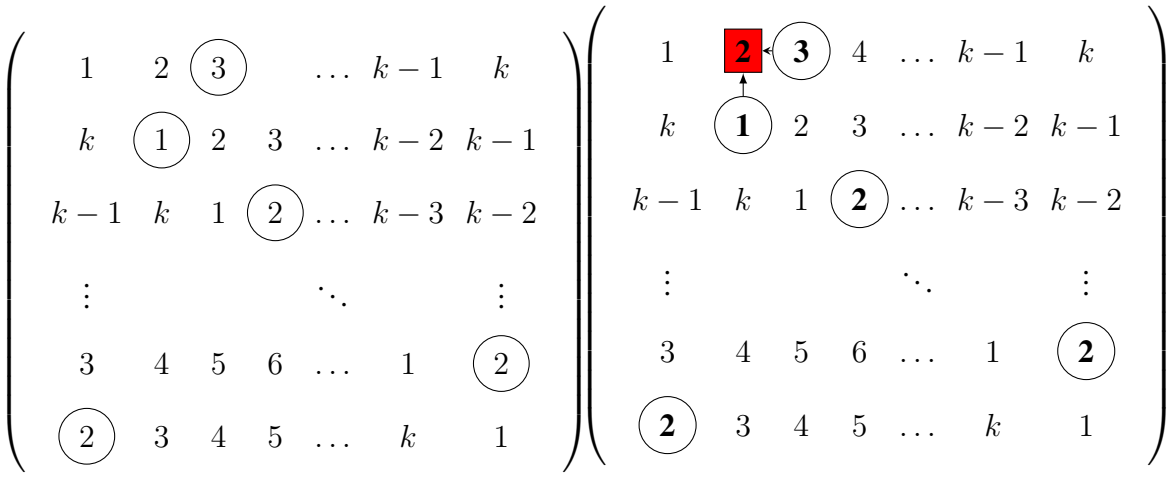
In such matrix, a matching corresponds to a selection of $\hat{k}$ cells, such that each column and row has only one cell selected. Suppose that (i, j) are a blocking pair

[3] The theorem holds for every $(M, W, \succ) : M \geq M^\star(q) = \hat{k}(q)$ and $W \geq W^\star(q) = \hat{k}(q)$ the reason being that it will be a Latin rectangle which can be completed into a Latin square, this is a consequence of Hall's theorem

[4] This is a maximal domain result; for any q it gives us an instance where someone regrets and tells us that for any instance greater than that it also will; however, this does not mean that it is the smallest instance at which an agent would regret truth in the $q-$quantile mechanism

[5] This is known as the Latin square subproblem of the stable marriage problem, see Thurber (2002).

to a given matching μ. Then $a_{i\mu(i)} > a_{ij} > a_{\mu(j)j}$, that is, man i prefers woman j to $\mu(i)$ which is represented by assigning a higher rank (lower number) and for woman j it is the case that larger numbers in her column correspond to more preferable matches, $a_{ij} > a_{\mu(j)j} \iff i \succ_j \mu(j)$. We refer to the entry in the ranking matrix corresponding to a blocking pair as a blocking cell. See figure A.2.



The circled cells represents a match: $\forall m \notin \{m_1, m_2\}$: $a_{m\mu(m)} = 2$.

$a_{12} = 2$ in red represents a blocking cell. m_1 would rather be matched to his second best choice than his third, while w_2 rather be matched to her $k-1$-th choice than to her k-th

Figure A.2: A matching represented in the associated ranking matrix.

Step 1) To see that the Latin square of order $\hat{k}$ generates at least $\hat{k}$ stable matches, note that there is always a permutation of rows such that the elements of the diagonal have the same value (and it is still a Latin square). For fixed $\ell \in \{1, \dots, \hat{k}\}$ let the matching be $a_{ii} = \ell$ for all $i \in M$ after the row-permutation and suppose it is blocked by some cell a_{ij}, then it has to be the case that the blocking cell satisfies: $a_{ii} > a_{ij} > a_{jj}$ which is impossible given that $a_{ii} = a_{jj} = \ell$. This means that for fixed ℓ, the matching μ with associated entries in the ranking matrix $a_{ii} = \ell \; \forall \in M$ is a stable matching. Since $\ell \in \{1, \dots, \hat{k}\}$ was arbitrary, this means there are $\hat{k}$ of such matchings, denote this set $\mathfrak{M}$. Note that $\mu^M, \mu^W \in \mathfrak{M}$, so the M-optimal and W-optimal matchings belong to this set. Moreover for any $\mu, \mu' \in \mathfrak{M}$ either $\mu \succ_M \mu'$ or $\mu' \succ_M \mu$, that is all matching in this set are strictly order for all men, each individual has $\hat{k}$ different stable partners, all acceptable partners are stable and attained in some matching in $\mathfrak{M}$. There is no stable matching outside of $\mathfrak{M}$ that is strictly ranked for all men, this is immediate from the previous observation.[6]

[6]Suppose $\mu \notin \mathfrak{M}$: for any $\mu' \in \mathfrak{M}$ either $\mu \succ_M \mu'$ or $\mu' \succ_M \mu$. Denote $\bar{\mu} = \{\mu' \in$

Step 2) Suppose $\exists \lambda \in \mathfrak{M}^c \cap S(\succ)$, since $\mu^M, \mu^W \in \mathfrak{M}$ then $\exists \bar{\mu}, \underline{\mu}$ where $\bar{\mu}(\lambda) = \inf_{\succeq_M}\{\mu \in \mathfrak{M} : \mu \succeq_M \lambda\}$ and $\underline{\mu}(\lambda) = \sup_{\succeq_M}\{\mu \in \mathfrak{M} : \lambda \succeq_M \mu\}$, transitivity ensures $\bar{\mu} \succeq_M \underline{\mu}$ and use $\bar{\mu}, \underline{\mu}$ for short. Consider the set $\mathbb{M}(\lambda) = P_{\mathfrak{M}}(\underline{\mu}) \cap Q_{\mathfrak{M}}(\bar{\mu})$ where $P_{\mathfrak{M}}(x), Q_{\mathfrak{M}}(x)$ denotes the set of predecessors and succesors of x in set $\mathfrak{M}$ under order $\succ_M$ respectively. Formally, $P_{\mathfrak{M}}(\underline{\mu}) = \{\mu \in \mathfrak{M} : \mu \succ_M \underline{\mu}\}$ and $Q_{\mathfrak{M}}(\bar{\mu}) = \{\mu \in \mathfrak{M} : \bar{\mu} \succ_M \mu\}$. If

1. $\mathbb{M} = \emptyset$

   Suppose $\bar{\mu} \succeq_M \lambda \succeq_M \underline{\mu},\ \lambda \in S(\succ) \setminus \{\underline{\mu}, \bar{\mu}\}$. If a man is getting his r-th best partner in $\underline{\mu}$, then he gets his $(r-1)$-th best partner at $\bar{\mu}$. By assumption there exists two nonempty sets of men M', M'' such that $M' = \{m \in M : \lambda \succ_{M'} \bar{\mu}\}$ and $M'' = \{m \in M : \lambda \sim_{M''} \bar{\mu}\}$. The matching λ must be such that for every $m \in M'$ it assigns him his $(r-1)$-th best partner, and for every $m \in M''$ assigns his r-th best best partner. Without loss of generality assume $m_1 \in M'$, then by construction $(\star)$ it follows that $\underline{\mu}(\mu(m_1)) = m_2$, that is his match under λ must be m_2's match under $\underline{\mu}$; in general $\underline{\mu}(\mu(m_t)) = m_{t+1}$ for $t \in \{1, \ldots, k-1\}$ and $\underline{\mu}(\mu(m_k)) = m_1$.So the improvement cycle by going from $\underline{\mu}$ to λ involves all men, then $M' = M$ and $M'' = \emptyset$, which is a contradiction.

2. $\mathbb{M} = \{\mu\}$

   Suppose $\bar{\mu} \succeq_M \lambda \succeq_M \underline{\mu},\ \lambda \in S(\succ) \setminus \{\underline{\mu}, \bar{\mu}, \mu\}$ and $\lambda, \underset{\not\succ_M}{\not\succeq_M} \mu$. Notice that if men are getting their r-t best stable partner in $\underline{\mu}$, then they are getting their $(r-1)$-th in μ and $(r-2)$-th in $\bar{\mu}$. The fact that λ and μ are not ordered according to $\succeq_M$ means there exists sets $\bar{M} \neq \emptyset, \tilde{M}, \underline{M} \neq \emptyset$ such that

$$\forall m \in \begin{cases} \bar{M} & \lambda \succ_{\bar{M}} \mu \\ \tilde{M} & \lambda \sim_{\tilde{M}} \mu \\ \underline{M} & \mu \succ_{\underline{M}} \lambda \end{cases}$$

   Notice that λ cannot assign any man anything worse than his r-th best stable choice since $\lambda \succeq_M \underline{\mu}$, similarly cannot assign anything better than his $(r-2)$-th best stable partner.

$\mathfrak{M} : \mu'$ is the $\succ_M$-greatest element for which $\mu' \succ_M \mu\}$, analogously define $\underline{\mu} = \{\mu' \in \mathfrak{M} : \mu'$ is the $\succ_M$-least element for which $\mu \succ_M \mu'\}$, so $\bar{\mu} \succ_M \mu \succ_M \underline{\mu}$. Both $\bar{\mu}$ and $\underline{\mu}$ are well defined since $\succ_M$ is a complete order on $\mathfrak{M} \cup \{\mu\}$. Suppose that in $\underline{\mu}$ each man is getting their r-th best stable partner, by construction in $\bar{\mu}$ each man gets its $(r-1)$-th best stable partner. Then $\mu(m)$ must give each man a strictly better stable partner than the r-th but strictly worst than his $r-1$ which is a contradiction.

Claim A.4. *$\underline{M} = M$ and $\bar{M} = \emptyset$*

Proof of claim. We proceed by induction. Suppose without loss of generality that $m_1 \in \underline{M}$, notice that $\mu(m_1) = w_{r-1} \succ_{m_1} w_r = \lambda(m_1) = \mu(m_2)$, so $\lambda(m_2) \neq \mu(m_2)$. Then for λ to be a matching a satisfy the case assumptions it has to be the case that $\lambda(m_2) \in \{\mu(m_1), \mu(m_3)\}$. Now given our construction ($\star$) if $\lambda \in S(\succ)$ then $\lambda(m_2) \neq \mu(m_1)$ since (m_1, w_r) would form a blocking pair. To see this notice that in the associated ranking matrix it holds that $a_{1r} = r > (r-1) = a_{1(r-1)} > (r-2) = a_{2(r-1)}$.

Next we show that if $\forall t < j, t, j \in \{1, \ldots, k\}$ it holds that $\lambda(m_t) = \mu(m_{t+1})$ then $\lambda(m_j) = \mu(m_{j+1})$, where the indexes are $\mod k$. By induction hypothesis $\lambda(m_j) \neq \mu(m_j) = \lambda(m_{j-1})$. Given the restrictions imposed by the existence of $\underline{\mu}$ and $\bar{\mu}$ it follows $\lambda(m_j) \in \{\mu(m_{j-1}), \mu(m_{j+1})\}$. Next we notice that if $\lambda(m_j) \in \mu(m_{j-1})$ then $(m_{j-1}, \mu(m_{j-1}))$ is a blocking pair, since the associated ranking matrix has entries: $a_{m_{j-1}\lambda(m_{j-1})} = r > (r-1) = a_{m_{j-1}\mu(m_{j-1})} > (r-2) = a_{m_j\mu(m_{j-1})}$. Consequently, $m_j \in \underline{M}$. The induction argument implies $\underline{M} = M$ and $\bar{M} = \emptyset$, which is a contradiction. □

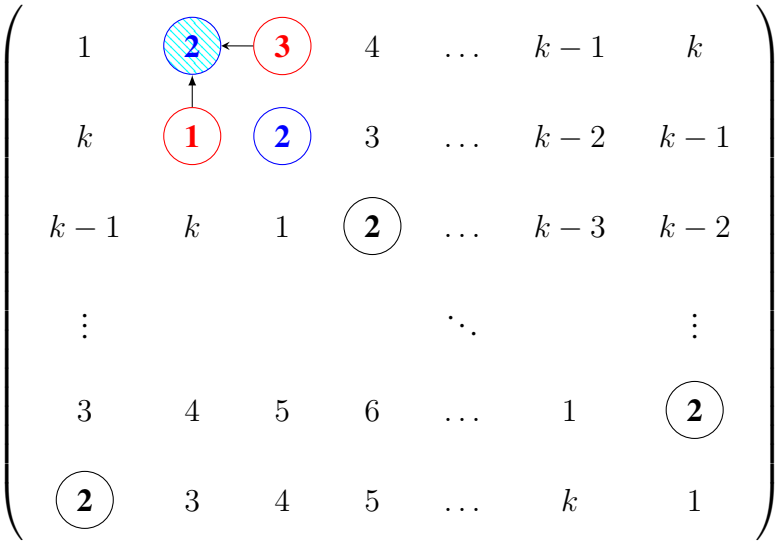
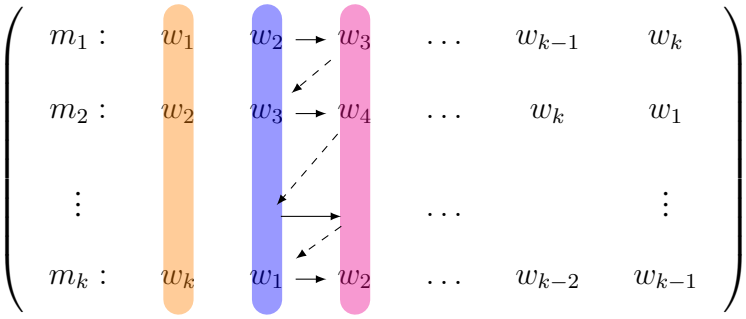


Figure A.3: The coordinates of the circled cells correspond to a matched couple in the associated ranking matrix.The black circles correspond to λ and μ. The blue circles correspond only to μ. The red circles correspond only to λ. The cyan cell is a blocking pair.

3. $|\mathbb{M}| \geq 2$,

Proof 1. Since $\mu, \mu', \lambda \in S(\succ)$ then $\mu \vee \lambda, \mu' \wedge (\mu \vee \lambda) \in S(\succ)$ since they are join and meet of stable matchings (Roth and Sotomayor, 1990, Theorem



Preferences of Men according to ($\star$)

Matching with $r = 3$,
$\bar{\mu}$: first column.
μ: second column.
$\underline{\mu}$: third column.

Trying to make one man worse than in μ yet weakly better than in $\underline{\mu}$ leads to a cycle that involves the whole set of men.

Figure A.4

5.31) Then

$$\begin{aligned}\mu \wedge (\mu' \wedge (\mu \vee \lambda)) =& (\mu \wedge \mu') \wedge (\mu \vee \lambda) \\ =& \mu \wedge (\mu \vee \lambda) \\ =& \mu\end{aligned}$$

where in the first equality we are using the associativity, in the second the fact that $\mu' \succ_M \mu$ and the absortion property of lattices in the third equality. This implies $\mu' \wedge (\mu \vee \lambda) \succeq_M \mu$ and by definition of $\wedge$, $\mu' \succeq_M \mu' \wedge (\mu \vee \lambda)$. This means there exists a stable matching $\hat{\lambda} = \mu' \wedge (\mu \vee \lambda)$ such that $\mu' \succeq_M \hat{\lambda} \succeq_M \mu$, whose existance we have already ruled out in case 1, therefore we have reached a contradiction. □

For the interested reader we provide a second proof that, though longer, makes use of a different technique.

Proof 2. We shall show that in this case, the alleged stable set contains a pentagon as a sublattice which implies the lattice is not distributive, therefore it cannot correspond to a stable set lattice. First we recall some basic definitions:

Definition (Birkhoff (1967) who attributes to Dedekind)**.** *A **lattice** is a poset P any two of whose elements have greatest lower bound or meet denoted $a \wedge b$*

and a least upper bound or join denoted $a \vee b$.[7]

A lattice L *is* ***complete*** *when each of its subsets X has a g.l.b. and a l.u.b. in L.*

A lattice L *is* ***distributive*** *if for every* $a, b, c \in L$ *the following equalities hold:* $a \wedge (b \vee c) = (a \wedge b) \vee (a \wedge c)$ *and* $a \vee (b \wedge c) = (a \vee b) \wedge (a \vee c)$. *Notice that not all lattices satisfy the distributive property.*

A ***sublattice*** *of a lattice* L *is a subset* X *of* L *such that* $a \in X, b \in X \Rightarrow a \wedge b \in X$ *and* $a \vee b \in X$.

A sublattice X *of* L *is called a* ***pentagon*** *if* X *is isomorphic to* $\mathfrak{N}_5 = \{i, o, a, b, c : b \vee c = i, a \wedge c = o, a > b\}$.

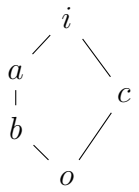


Figure A.5: Pentagon $\mathfrak{N}_5$

Theorem (attributed to Conway)**.** *When preferences are strict, the set of stable matchings is a distributive lattice under* $\succeq_M$.

Theorem. *If a lattice* L *is distributive then it does not contain a sublattice isomorphic to* $\mathfrak{N}_5$ *(pentagon).*

By assumption there must exist $\mu, \mu' \in \mathbb{M}$ such that $x \overset{\succ}{\underset{\not\succ}{\not\succeq}} \lambda$ for $x \in \{\mu, \mu'\}$, wlog assume $\mu' \succ \mu$ (notice this is well defined since $\mu, \mu \in \mathfrak{M}$), μ' being the immediate predecessor of μ according to $\succ_M$ (this hold by transitivity). Let $\mu^\star = \mu \vee \lambda$ and $\mu_\star = \mu \wedge \lambda$. Analogous definitions for $\mu'_\star$ and $\mu'^\star$. Notice $(\mu^\star, \mu_\star, \mu'_\star \mu'^\star)$ are all stable matchings since they are join and meet of stable matchings. One of the following cases must hold

a) $\mu^\star \in \mathfrak{M}$, $\mu_\star \in \mathfrak{M}$.

b) $\mu^\star \notin \mathfrak{M}$, $\mu_\star \in \mathfrak{M}$.

[7] We recall some basic yet useful properties of $\wedge$ and $\vee$.

a) Associativity: $a \wedge (b \wedge c) = (a \wedge b) \wedge c$ and $a \vee (b \vee c) = (a \vee b) \vee c$

b) Commutative: $a \wedge b = b \wedge a$ and $a \vee b = b \vee a$

c) Absortion: $a \wedge (a \vee b) = a$ and $a \vee (a \wedge b) = a$

d) Idempotency: $a \wedge a = a$ and $a \vee a = a$

c) $\mu^\star \in \mathfrak{M}$, $\mu_\star \notin \mathfrak{M}$.

d) $\mu^\star \notin \mathfrak{M}$, $\mu_\star \notin \mathfrak{M}$.

We show the contradiction for the first case, similar constructions work for the other cases.

Claim A.5. *$(\mu, \mu', \mu^\star, \mu_\star, \lambda)$ is a sublattice of $S(\succ)$ isomorphic to $\mathfrak{N}_5$*

Proof of claim. First we show that the set $\{\mu, \mu', \mu^\star, \mu_\star, \lambda\}$ constitutes a sublattice of $S(\succ)$ under the assumptions. To do so first we show: $\mu^\star = \mu'^\star$. Notice $\mu^\star \succeq_M \bar{\mu}$ by definition of $\bar{\mu}$, $\mu' \in Q_{\mathfrak{M}}(\bar{\mu}^0)$ implies $\mu^\star \succ_M \mu'$ and $\mu^\star \succeq_M \lambda$, which imply $\mu^\star$ is an upper bound to $\{\mu', \lambda\}$. Suppose it is not the least upper bound, then $\exists \mu'' : \mu^\star \succ_M \mu''$ such that μ'' is also an upper bound to $\{\mu', \lambda\}$, and given that $\mu' \succ_M \mu$, then $\mu^\star$ cannot be the least upper bound on $\{\mu, \lambda\}$, which contradicts the definition of $\mu^\star = \mu \vee \lambda$. Similar reasoning shows $\mu_\star = \mu'_\star$. Consequently meets and joins of every pair of elements are by construction inside the set $(\mu, \mu', \mu^\star, \mu_\star, \lambda)$ Then the following $\phi : \{\mu, \mu', \mu^\star, \mu_\star, \lambda\} \to \mathfrak{N}_5 = \{i, o, a, b, c : b \vee c = i, a \wedge c = o, a > b\}$ is an isomorphism to $\mathfrak{N}_5$:

$$\phi(\cdot) : \begin{cases} \mu^\star & \longmapsto i \\ \mu' & \longmapsto a \\ \mu & \longmapsto b \\ \lambda & \longmapsto c \\ \mu_\star & \longmapsto o \end{cases}$$

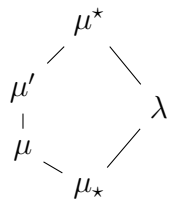


Figure A.6: Sublattice

Consequently, we have shown that the set cannot correspond to a stable matching lattice. □

□

□

Necessity

Existence

Theorem 2.1 (Necessity). $\forall i \in N = M \cup W$, $\succ_i$ is regret-free in $\phi(\cdot) \in \{\phi^M(\cdot), \phi^W(\cdot)\}$

The result is established through four claims: Claim A.6 shows that truth-telling is regret-free for the proposing side, as a consequence of dominant strategy incentive compatibility. Consequently the rest of the proof focuses only on the receiving side. Claim A.7, A.8 and A.9 show that there does not exist a report through which i regrets telling the truth. Each claim deals with reports that differ from the truth in a specific manner. Claim A.7 shows that changing the order of alternatives that are preferred to the observed match will not change the resulting matching. Claim A.8 and A.9 show that any other report that differs from the truth in an essential manner is never a safe deviation compared to telling the truth, in the sense that whenever it may result in a more preferable match it may also result in a less preferable as well. Claim A.8 deals with those deviations where an alternative which is less preferred to the observed match by the true preference profile is reported as preferred to the said match. Claim A.9 deals with deviations where the relative order between two alternatives that are less preferred to the observed match is reversed. All reports that differ from the truth in an essential way have to fit into the conditions of (at least) one of these claims.

Proof. Let $(M, W, \succ)$ be a private information matching market where $\phi(\succ) = \mu_M(\succ)\ \forall \succ \in \mathcal{P}$, that is the M-proposing deferred acceptance algorithm. Take an arbitrary agent i and fix an arbitrary $\mu = \phi(\succ_i, \succ_{-i})$ for some $\succ_{-i} \in \mathcal{P}_{-i}$. Define

$$UC^{\succ_i}_{\phi(\succ)(i)} = \{j \in J : j \succ_i \phi(\succ)(i)\}$$

$$LC^{\succ_i}_{\phi(\succ)(i)} = \{j \in J : \phi(\succ)(i) \succ_i j\}$$

that is, $UC^{\succ_i}_{\phi(\succ)(i)}$ denotes the upper contour set with respect to $\phi(\succ)(i)$ under $\succ_i$, that is the set of partners that player i considers strictly preferable (according to his/her true preference) to the partner under $\phi(\succ)$; analogously interpret $LC^{\succ_i}_{\phi(\succ)(i)}$.

Claim A.6. *Truth-telling is regret-free for every agent in the proposing side.*

Proof of Claim. Follows directly from strategy-proofness for men. □

Consequently we can focus on the receiving side (so $i \in W$ from now on). We must show that for an arbitrary agent on the receiving side and for an arbitrary matching that may result from her reporting her true preference, there is no alternative report through which that agent regrets truth-telling. We start by showing (Claim A.7) that a report that differs from truth only in the way that it orders elements that are preferred to the observed matching cannot yield a better matching for the agent.

Claim A.7. *For any* $\mu = \phi(\succ_i, \succ_{-i})$ *for some* $\succ_{-i} \in \mathcal{P}_i$, *if* $\succ_i'$: $UC^{\succ_i}_{\phi(\succ)(i)} = UC^{\succ_i'}_{\phi(\succ)(i)}$ *and* $a \succ_i' b \Leftrightarrow a \succ_i b \quad \forall a, b \in LC^{\succ_i}_{\phi(\succ)(i)} \cup \{\phi(\succ)(i)\} \Rightarrow \phi(\succ_i, \succ_{-i}) = \phi(\succ_i', \succ_{-i})$.

Proof of Claim. The player does not reject an offer under $\succ_i'$ that was accepted under $\succ_i$, she cannot affect the set of offers that are made to her, and consequently cannot affect the outcome favorably.[8] □

Claim A.8. *Suppose* $\exists(\succ_i', \hat{\succ}_{-i})$ *such that*

(i) $\phi(\succ_i, \hat{\succ}_{-i}) = \mu$

(ii) $\succ_i'$: $\exists \tilde{j} \in LC^{\succ_i}_{\phi(\succ)(i)}$ *and* $\tilde{j} \in UC^{\succ_i'}_{\phi(\succ)(i)}$

(iii) $\phi(\succ_i', \hat{\succ}_{-i}) \succ_i \phi(\succ_i, \hat{\succ}_{-i})$

then $\exists \tilde{\succ}_{-i}$ *such that* $\phi(\succ_i, \tilde{\succ}_{-i}) = \mu$ *and* $\phi(\succ_i, \tilde{\succ}_{-i}) \succ_i \phi(\succ_i', \tilde{\succ}_{-i})$.

Proof of claim. **Case 1. Truncation** $\tilde{j} = i$ It is enough to consider the following preference profile to see that the truncation can leave the agent worse off than telling the truth

$$\tilde{\succ}_k : \phi(\succ)(k) \tilde{\succ}_k k \tilde{\succ}_k \dots \ \forall k \neq i$$

Notice that the profile is consistent with the observed matching since everyone considers their assigned partner as their unique acceptable partner. However under $\succ_i'$ that match is no longer acceptable for i. Since $\phi(\cdot)$ is stable with respect to the reported preferences (in particular individually rational) it leaves i unmatched

[8] Player i did not receive any offer from a member of $UC^{\succ_i}_{\phi(\succ)(i)}$ in $\phi(\succ)$ by construction of the DA. Since everyone else's preference is the same as before, the first round of offers does not change; since player i's offers are the same (and belong to $LC^{\succ_i}_{\phi(\succ)(i)} \cup \{\phi(\succ)(i)\}$) and her preferences over those alternatives did not change, the set of active players in the next round is the same. The set of active players in the second round is the same, again she faces choices on $LC^{\succ_i}_{\phi(\succ)(i)} \cup \{\phi(\succ)(i)\}$, and makes the same choice. An inductive argument finishes the proof.

under $\succ_i'$, which is a strictly worse off situation from the point of view of i's the true preference profile.

Case 2. A non-truncation ($\tilde{j} \neq i$). Similar to the truncation case it is enough to consider:

$$\tilde{\succ}_{\tilde{j}} : i \; \tilde{\succ}_{\tilde{j}} \; \phi(\succ)(\tilde{j}) \; \tilde{\succ}_{\tilde{j}} \; \tilde{j} \; \tilde{\succ}_{\tilde{j}} \dots$$
$$\tilde{\succ}_{k} : \phi(\succ)(k) \; \tilde{\succ}_{k} \; k \; \tilde{\succ}_{k} \dots \; \forall k \neq \{\tilde{j}, i\}$$

It is straightforward to verify that the preference profile is consistent with the observed matching. Under the alternative report, everyone except i has the same preference profile as before, so first round proposals are the same. However, under $\succ_i'$ agent $\tilde{j}$ is declared as preferred to $\phi(\succ)(i)$, which means that agent i accepts $\tilde{j}$'s offer. Since there are no more active players the algorithm stops and matches agent i to $\tilde{j}$ which is a strictly worse outcome under i's true preference profile, that is $\phi(\succ_i, \tilde{\succ}_{-i})(i) = \phi(\succ)(i) \succ_i \tilde{j} = \phi(\succ_i', \tilde{\succ}_{-i})(i)$. Consequently, i cannot regret truth-telling ($\succ_i$) through a deviation ($\succ_i'$) consistent with the claim at the observed matching (μ) in DA. □

Claim A.9. *Suppose* $\exists(\succ_i', \hat{\succ}_{-i})$ *such that*

(i) $\phi(\succ_i, \hat{\succ}_{-i}) = \mu$

(ii) $\succ_i'$: $\exists u, v \in LC^{\succ_i}_{\phi(\succ)(i)}$ *such that* $u \succ_i v$ *and* $v \succ_i' u$

(iii) $\phi(\succ_i', \hat{\succ}_{-i}) \succ_i \phi(\succ_i, \hat{\succ}_{-i})$

then $\exists \tilde{\succ}_{-i}$ *such that* $\phi(\succ_i, \tilde{\succ}_{-i}) = \mu$ *and* $\phi(\succ_i, \tilde{\succ}_{-i}) \succ_i \phi(\succ_i', \tilde{\succ}_{-i})$.

Proof of Claim. The general structure of the proof is the same as in the previous claim so we just described the preference profile that we need to take into consideration. Let $\tilde{\succ}_{-i}$ be the following:

$$\tilde{\succ}_{v} : i \; \tilde{\succ}_{v} \; \phi(\succ)(v) \; \tilde{\succ}_{v} \dots$$
$$\tilde{\succ}_{u} : i \; \tilde{\succ}_{u} \; \phi(\succ)(u) \; \tilde{\succ}_{u} \dots$$
$$\tilde{\succ}_{\phi(\succ)(v)} : v \; \tilde{\succ}_{\phi(\succ)(v)} \; \phi(\succ)(i) \; \tilde{\succ}_{\phi(\succ)(v)} \; \dots$$
$$\tilde{\succ}_{\phi(\succ)(i)} : \phi(\succ)(v) \; \tilde{\succ}_{\phi(\succ)(i)} \; i \; \tilde{\succ}_{\phi(\succ)(i)} \; \dots$$
$$\tilde{\succ}_{k} : \phi(\succ)(k) \; \tilde{\succ}_{k} \; k \; \tilde{\succ}_{k} \; \dots \; \forall k \neq \{i, u, v, \phi(\succ)(i), \phi(\succ)(v)\}$$

This preference profile is consistent with the observed matching. Where i to report $\succ_i'$ instead,

	w_1	$\dots$	i	$\phi(\succ)(v)$	$\phi(\succ)(u)$	$\dots$	$w_{\lvert W\rvert}$
1st	$\boldsymbol{\phi(\succ)(\mathbf{w_1})}$	$\dots$	$u, \mathbf{v}$	$\boldsymbol{\phi(\succ)(\mathbf{i})}$		$\dots$	$\boldsymbol{\phi(\succ)(\mathbf{w_{\lvert W\rvert}})}$
2nd					$\mathbf{u}$		
$\phi(\succ_i', \tilde{\succ}_{-i})$	$\boldsymbol{\phi(\succ)(\mathbf{w_1})}$	$\dots$	$\mathbf{v}$	$\boldsymbol{\phi(\succ)(\mathbf{i})}$	$\mathbf{u}$	$\dots$	$\boldsymbol{\phi(\succ)(\mathbf{w_{\lvert W\rvert}})}$

The resulting allocation matches i to v, which is a worse outcome for agent i according to her true preference profile, that is $\phi(\succ_i, \tilde{\succ}_{-i})(i) = \phi(\succ)(i) \succ_i v = \phi(\succ_i', \tilde{\succ}_{-i})(i)$. Consequently, i cannot regret truth-telling ($\succ_i$) through a deviation ($\succ_i'$) consistent with the claim at the observed matching (μ) in DA. □

The analysis shows that there is no report through which agent i can regret telling the truth at μ. Since this was done for an arbitrary $\mu \in \mathcal{M}|_{\succ_i}$, it holds for all such matchings that may result from telling the truth. Consequently, there is no μ at which i regrets truth-telling which means truth-telling is regret-free for agent i. Since this conclusion holds for an arbitrary agent in either side of the market (proposing or receiving), putting together the previous claims the proposition is proven. □

Let $T_i = \{\succ_i'' \in \mathcal{P}_i : A_i(\succ_i'') = A_i(\succ_i)$ and $a \succ_i'' b \Leftrightarrow a \succ_i b \; \forall a, b \in A_i(\succ_i) \cup \{i\}\}$ denote the set of all preferences for i that only differ from the true one in how they rank unacceptable choices between themselves. Note that by construction of the DA the matching generated by truth and by a report in T_i is the same, which means i cannot regret telling the truth through an element in T_i. However this reports differ from the truth only in an inessential manner.

Corollary A.2. *Any report that differs from the truth only in how it ranks the elements of the unacceptable set among themselves ($\succ_i' \in T_i$) is also regret-free.*

Proof. Since the DA does not take into account the relative ranking among alternatives in the unacceptable set $\forall \succ_i' \in T_i, \forall \tilde{\succ}_{-i} \in \mathcal{P}_{-i} \quad \phi(\succ_i, \tilde{\succ}_{-i}) = \phi(\succ_i', \tilde{\succ}_{-i})$, then any $\succ_i' \in T_i$ is regret-free. □

Uniqueness

Proposition 2.1. Truth is the essentially unique regret-free report in the DA mechanism. Moreover, i regrets any other report *through truth*.

Proof of Proposition. Any $\succ_i' \in \mathcal{P}_i \setminus T_i$ must belong to one of the following cases:

1. $\succ_i' \in \mathcal{P}_i$ such that $\exists k : k \in A(\succ_i)$ and $U(\succ_i')$.
2. $\succ_i' \in \mathcal{P}_i$ such that $\exists j \in U_i(\succ_i)$ and $j \in A_i(\succ_i')$.
3. $\succ_i'$ involves a permutation among the acceptable set.

Cases 1 and 2 are tackled below in A.2.

Here we tackle case 3.

We develop an algorithm to find a $\mu \in \mathcal{M}|_{\succ_i'}$ at which i regrets reporting $\succ_i'$ through $\succ_i$ for an arbitrary i in the receiving side. An analogous argument works for an agent in the proposing side.[9]

Let $|J|$ denote the cardinality of the agents on the proposing side that are acceptable to i with respect to her true preference profile. Relabel agents such that their index reflects their position according to $\succ_i$, that is j_1 is the $\succ_i$-maximal element (agent) on $A_{i,1}(\succ_i) = A_i(\succ_i)$, j_2 the $\succ_i$-maximal element on $A_{i,2}(\succ_i) = A_{i,1}(\succ_i) \setminus \{j_1\}$, etc.

> *Step 1.* If the index of the $\succ_i'$-maximal element on $A_{i,|J|-1}(\succ_i)$ is smaller than the index of the $\succ_i'$-maximal element on $A_{i,|J|}(\succ_i)$, then go to step 2.
> Otherwise, set $\mu \in \mathcal{M}|_{\succ_i'} : \mu(i) = \{\succ_i'\text{-maximal element on } A_{i,|J|-1}(\succ_i)\}$ and $\mu(k) = k \; \forall k \neq \{i, \mu(i)\}$.[10] ***Break.***

> *Step* $k \in \{2, \ldots, |J| - 1\}$. If the index of the $\succ_i'$-maximal element on $A_{i,|J|-k}(\succ_i)$ is smaller than the index of the $\succ_i'$-maximal element on $A_{i,|J|-(k-1)}(\succ_i$), then go to step $k+1$.
> Otherwise, set $\mu \in \mathcal{M}|_{\succ_i'} : \mu(i) = \{\succ_i'\text{-maximal element on } A_{i,|J|-k}(\succ_i)\}$ and $\mu(k) = k \; \forall k \neq \{i, \mu(i)\}$. ***Break.***

[9] In the case of the receiving side described in the text the algorithm looks for the first switch in the preference relation from least to most preferred acceptable alternative. In the case of the proposing side the search is done from most to least preferred acceptable partner.

[10] The essential part of the matching found by the algorithm is to whom agent i is matched, the choice of leaving everyone else unmatched is arbitrary and not unique.

Given that $\succ_i'$ is a permutation of $\succ_i$ on $A_i(\succ_i)$ it cannot be the case that $\forall j \in A_i(\succ_i)$ $j_k \succ_i' j_l$ whenever $k < l$. Therefore the algorithm necessarily sets a μ. Next we explain why at such μ i regrets $\succ_i'$ through $\succ_i$.

First, consider a case where the algorithm stops after step 1, setting $\mu(i) = x = \{\succ_i'$-maximal element on $A_{i,|J|-1}(\succ_i)\}$. By construction of DA, i can only have received offers from x and $y = \{\succ_i'$-maximal element on $A_{i,|J|}(\succ_i)\}$, necessarily so from x since it is matched to him under the observed matching. The preference profiles $\tilde{\succ}_{-i} \in \mathcal{M}|_{\succ_i'}$ are divided into those cases in which i received an offer from y and those in which it did not; there always exist preference profiles that satisfy each condition. If she did not, then she only observed an offer from x and consequently, $\phi(\succ_i, \tilde{\succ}_{-i}) = \phi(\succ_i', \tilde{\succ}_{-i}) = x$ since i does not reject or accept any offer differently under $\succ_i'$ than under $\succ_i$. On the other hand, if i received an offer from y it means at some point she decided between y and x in favor of x. However, since the algorithm stopped to produce μ it means that $y \succ_i x$, consequently $\phi(\succ_i, \tilde{\succ}_{-i}) \succ_i \phi(\succ_i', \tilde{\succ}_{-i})$.

The same logic extends to the case where the algorithm stops at a step k: i cannot have received offers from any $z \succ_i' \mu(i)$. For any $s, t \in J : \mu(i) \succ_i' s$ and $\mu(i) \succ_i' t$ it is the case that $s \succ_i' t \iff s \succ_i t$. That is, the binary relation between the options that can potentially have made an offer to i is the same under $\succ_i'$ than under $\succ_i$, which means that any offer that did not involve $\mu(i)$ is accepted or rejected in the same manner under both $\succ_i'$ and $\succ_i$. The only cases in which they differ are in those where $\mu(i)$ was chosen over some $s \in J : \mu(i) \succ_i' s$ and $s \succ_i \mu(i)$. Consequently $\phi(\succ_i, \tilde{\succ}_{-i}) \succ_i \phi(\succ_i', \tilde{\succ}_{-i})$.

□

A.2 Remarks

Existence and Uniqueness not guaranteed in general

Remark 2.1. There exists a mechanism ϕ such that no agent has a regret-free report, namely the Boston Mechanism. In contrast, every report is regret-free in a constant mechanism.

Proof. By way of example,

Constant Mechanism

Consider the following trivial mechanism, $\phi(\succ)(i) = i \quad \forall i \in M \cup W$. Trivially every report yields the same outcome, therefore no report can ex-post dominate

another and consequently multiple and (essentially) different reports can be regret-free at the same time; at least if we do not restrict the mechanism to be stable. Particularly note that both a truncation strategy and truth-telling are both being regret-free at the same time.

Boston Mechanism

We will show that in the Boston mechanism, for a specific market size and preference profile of a player, this player has no regret-free report at his disposal. Let $|M| = |W| = 3$ and $m_1 : w_1 \succ w_2 \succ w_3$. He is allowed to make any of the following reports:[11]

$1)123\emptyset$	$2)132\emptyset$	$3)1\emptyset23$	$4)1\emptyset32$	$5)13\emptyset2$	$6)12\emptyset3$
$7)312\emptyset$	$8)321\emptyset$	$9)3\emptyset12$	$10)3\emptyset21$	$11)32\emptyset1$	$12)31\emptyset2$
$13)2\emptyset13$	$14)2\emptyset31$	$15)23\emptyset1$	$16)231\emptyset$	$17)21\emptyset3$	$18)213\emptyset$

m_1 regrets $\succ'_{m_1} \in \{3, 4, 5, 6, 9, 10, 11, 12, 13, 14, 15, 17\}$ at a matching where $\phi(\succ'_{m_1}, \succ_{-m_1})(m_1) = m_1$ through $\succ = \{1\}$, that is the true preference profile. First we note that the Boston Mechanism is individually rational, therefore m_1 cannot do worst than being single by reporting the truth, no matter what the true preference profile of others is. Note that any $\succ'_{m_1} \neq \succ$ involves declaring some acceptable alternative w_j as unacceptable. Then we note that when $\succ_{w_j}$: $m_1 \succ \phi(\succ'_{m_1}, \succ_{-m_1})(w_j) \overset{\succ}{(\sim)} w_j$ and $\succ_k$: $\phi(\succ'_{m_1}, \succ_{-m_1}) \overset{\succ}{(\sim)} k$ for $k \in M \cup W \setminus \{m_1, w_j\}$ the Boston mechanism allocates $\phi(\succ_{m_1}, \succ_{-m_1})(m_1) = w_j$ which is strictly preferred by m_1 according to his true preference profile than remaining single.

m_1 regrets $\succ'_{m_1} \in \{2, 7, 8\}$ and $\succ''_{m_1} \in \{16, 18\}$ at a matching $\phi(\succ'_{m_1}, \succ_{-m_1}) = \left(\begin{smallmatrix} w_1 & w_2 & w_3 & \cdot \\ m_3 & m_2 & \cdot & m_1 \end{smallmatrix}\right)$ and $\phi(\succ''_{m_1}, \succ_{-m_1}) = \left(\begin{smallmatrix} w_1 & w_2 & w_3 & \cdot \\ m_3 & \cdot & m_2 & m_1 \end{smallmatrix}\right)$ respectively through $\succ = \{1\}$. By individual rationality of BM he cannot do worst by telling the truth and there is always the possibility that the reason he ends up single is in fact trying to compete for what is reported as his first choice, and therefore losing his potential seat in the remaining ones, which would not happen (for at least some preference profile) by telling the truth.

[11] With exception of declaring every alternative as unacceptable.

m_1 regrets $\succ'_{m_1} = \{2\}$ at matching $\phi(\succ'_{m_1}, \succ_{-m_1}) = \left(\begin{smallmatrix} w_1 & w_2 & w_3 & \cdot \\ m_3 & m_2 & \cdot & m_1 \end{smallmatrix}\right)$ through $\succ = \{1\}$, when the profile is

w_1 :312 $\quad$ m_1 :$\cdot$

w_2 :123 $\quad$ m_2 :123

w_3 :23$\emptyset$ $\quad$ m_3 :132

m_1 regrets $\succ'_{m_1} \in \{7, 8\}$ at matching $\phi(\succ'_{m_1}, \succ_{-m_1}) = \left(\begin{smallmatrix} w_1 & w_2 & w_3 & \cdot \\ m_3 & \cdot & m_2 & m_1 \end{smallmatrix}\right)$ through $\succ = \{1\}$, when the profile is

w_1 :132 $\quad$ m_1 :$\cdot$

w_2 :23$\emptyset$ $\quad$ m_2 :312

w_3 :213 $\quad$ m_3 :123

m_1 regrets $\succ'_{m_1} \in \{16, 18\}$ at a matching where $\phi(\succ'_{m_1}, \succ_{-m_1}) = \left(\begin{smallmatrix} w_1 & w_2 & w_3 & \cdot \\ m_3 & \cdot & m_2 & m_1 \end{smallmatrix}\right)$ through $\succ = \{1\}$, when the profile is

w_1 :132 $\quad$ m_1 :$\cdot$

w_2 :213 $\quad$ m_2 :213

w_3 :23$\emptyset$ $\quad$ m_3 :132

□

Stability and regret-free reports

Remark. If a mechanism is stable then (i) no report that declares some acceptable partner as unacceptable can be regret-free. Similarly, (ii) no report that declares an unacceptable partner as acceptable can be regret-free. Moreover in each of these cases the individual regrets the report *through truth*.

Proof. Let $\succ'_i \in \mathcal{P}_i$ be such that $\exists k : k \in A(\succ_i), U(\succ'_i)$. Take the matching

$$\phi(\succ'_i, \succ_{-i}) = \begin{pmatrix} \cdot & i & \cdot \\ k & \cdot & \cdot \end{pmatrix}$$

That is, i remains single.[12] Consider the preference profile consistent with such matching

$$\begin{aligned} i &: \cdots \succ' i \succ' k \\ k &: i \succ'' \phi(\succ', \succ_{-i})(k) \overset{\succ''}{(\sim)} k \\ j &: \phi(\succ', \succ_{-i})(j) \overset{\succ''}{(\sim)} j (\succ'' i) \end{aligned}$$

By individual rationality, $\phi(\cdot, \succ_{-i}) \succeq_i i$. And we note that for $\succ''_{-i}$, $\phi(\succ_i, \succ''_{-i}) = k \succ_i i$ because of stability (not IR) since otherwise (i, k) constitute a blocking pair. This establishes (i).

To establish (ii), suppose $\succ'_i$ is such that $\exists j \in U_i(\succ_i)$ and $j \in A_i(\succ'_i)$. Then, $\exists \mu \in \mathcal{M}|_{\succ'_i} : \mu(i) = j$. Since $\phi(\cdot)$ is individually rational, $\phi(\succ_i, \tilde{\succ}_{-i}) \succ_i j$. □

[12] To see this note that if you run the DA with i's side proposing it remains single, then by the rural hospital theorem, i must be single in any stable matching, consequently, no matter which stable $\phi(\cdot)$ selects, it will still have i as single.

Independence of stability and regret-free truth-telling

Stability and regret-free truth-telling are independent properties in the sense that neither is implied by the other.

ϕ	Stable	Unstable
RFTT	DA	Serial Dictatorship
Not RFTT	Median Stable Mech.	Boston Mechanism

Table A.1: Independence of stability and regret-free truth-telling.

where RFTT stands for Regret-Free Truth-Telling. DA was shown to be stable and RFTT in section 2.4. The median stable mechanism was shown to fail regret-free truth-telling in section 2.4. For Boston Mechanism, see section A.2. Lastly, the serial dictatorship is known to be an unstable mechanism. It trivially satisfies RFTT since it is strategy-proof mechanism.

A.3 Switching DA: Stable, not RFTT

Consider the following mechanism: Let $\phi(\succ) = \phi_W(\succ) \quad \forall \succ \neq \{\succ'\}$ and $\phi(\succ') = \phi_M(\succ')$ where $\succ'$ is the following:

$$
\begin{aligned}
m_1 &: w_1 \succ' w_2 \succ' \emptyset & \qquad w_1 &: m_2 \succ' m_1 \succ' \emptyset \\
m_2 &: w_2 \succ' w_1 \succ' \emptyset & \qquad w_2 &: m_1 \succ' m_2 \succ' \emptyset \\
m_3 &: \emptyset & \qquad w_3 &: \emptyset
\end{aligned}
$$

Claim A.10. *w_1 regrets $w_1 : m_2 \succ' m_1 \succ' \emptyset$ at $\mu = \begin{pmatrix} m_1 & m_2 & m_3 & \cdot \\ w_1 & w_2 & \cdot & w_3 \end{pmatrix}$ through $w_1 : m_2 \succ m_1 \succ m_3 \succ \emptyset$.*

Proof. First, if $(\succ'_{w_1}, \succ'_{-w_1})$ then

$$\phi(\succ') = \phi_M(\succ') = \begin{pmatrix} m_1 & m_2 & m_3 & \cdot \\ w_1 & w_2 & \cdot & w_3 \end{pmatrix}$$

if instead w_1 had reported $w_1 : m_2 \succ m_1 \succ m_3 \succ \emptyset$ then $\phi(\succ_{w_1}, \succ'_{-w_1}) = \phi_W(\succ_{w_1}, \succ'_{-w_1})$, then

$$\phi_W(\succ_{w_1}, \succ'_{-w_1})(w_1) = m_2 \succ'_{w_1} m_1 = \phi_M(\succ')(w_1)$$

Now we need to show that $\forall \succ_{-w_1} \in \mathcal{P} : \phi(\succ'_{w_1}, \succ_{-w_1}) = \mu$ it holds that $\phi(\succ_{w_1}, \succ_{-w_1}) \succeq'_{w_1} \phi(\succ'_{w_1}, \succ_{-w_1})$.

If $\succ_{-w_1} \neq \succ'_{-w_1}$ then $\mu = \phi_W(\succ'_{w_1}, \succ_{-w_1})$, in which case it means w_1 proposed to m_2 - potentially got tentatively accepted and then rejected, then proposed to m_1 which accepted and never rejected w_1. Since the $\succ_{-w_1}$ is fixed and the order of proposals made by w_1 did not change it has to be the case that w_1 is still matched to m_1 (so it is not worse-off). $\square$

Appendix B

APPENDIX TO CHAPTER 3

Proof of Theorem 3.1. We start with a preference profile with a unique stable matching μ that satisfies the One and a Half Cycle Condition. That preference will ultimately correspond to one state in the economy, denoted θ_1. Suppose the cycle guaranteed by the One and a Half Cycle Condition has a spoiler $\bar{w}$ that blocks the swaps determined by the cycle with firm $f_{\bar{k}}$, who is to be matched with $w_{\bar{k}} = \mu(f_{\bar{k}})$ under the unique stable matching μ. As mentioned in the body of the text, for the sake of expositional transparency, we restrict attention to the case where there is a unique spoiler and a unique firm that would block the swap for at least one cycle.[1] Suppose, then, that $\bar{w}$ is a unique spoiler, who blocks the swap determined by the cycle only with $f_{\bar{k}}$.

We construct a preference profile that will ultimately define another state in the economy, denoted θ_2. To do so, we take the half cycle identified for the original preferences and define preferences for firms that turn it into a full cycle. This implies that the sub-market involving the corresponding workers and firms has multiple stable matchings. In addition, we make sure that the constructed preference profile belongs to the domain of preference profiles generating a unique stable matching. This implies that there must be a spoiler corresponding to the proper cycle in the constructed preferences. The spoiler is (intentionally) chosen to be $w_{\bar{k}}$. In this way, the roles of spoiler and spoiled are reversed across the two states. We construct preferences so that workers maintain their match preferences from the original preference ordering, while firms' preferences are specified as follows:[2]

$$\begin{aligned} \text{for any } f'_{k'} \in F' \setminus f_{\bar{k}} \qquad & \succ_{f'_{k'}}\colon w'_{k'}, w'_{k'-1}, \varnothing, \dots \\ \text{for } f_{\bar{k}} \qquad & \succ_{f_{\bar{k}}}\colon \bar{w}, w_{\bar{k}}, w'_{K'}, \varnothing, \dots \\ \text{for any } f \notin F' \qquad & \succ_{f}\colon \varnothing, \dots \end{aligned}$$

By design, the firm-proposing DA in market θ_2 generates the matching:

[1] This restriction is weakened in the Online Appendix.

[2] The preferences described are certainly very specific and are by no means the only ones that would generate the multiplicity claimed in the Theorem. Their choice is design to simplify our arguments.

$$\begin{aligned} \text{for any } f'_{k'} \in F' \setminus f_{\bar{k}} \quad & \mu(\theta_2)(f'_{k'}) = w'_{k'} \\ \text{for } f_{\bar{k}} \quad & \mu(\theta_2)(f_{\bar{k}}) = \bar{w} \\ \text{for any } f \notin F' \quad & \mu(\theta_2)(f) = \varnothing \end{aligned}$$

In fact, $\mu(\theta_2)$ is the unique stable matching in θ_2. To see this, consider any firm or worker f outside the half cycle. By the Rural Hospital Theorem, they must be unmatched in every stable matching. Thus, if there is any multiplicity of stable matchings, it must come from rematching the workers and firms in the original half cycle. There is only one other candidate matching to be stable in θ_2, which is the matching that results by implementing the swap desired by the workers in the generated cycle of θ_2. However, the matching resulting from such a swap would be blocked by firm $f_{\bar{k}}$ and worker $w_{\bar{k}}$. Therefore $\mu(\theta_2)$ is the unique stable matching.[3]

With respect to this generated cycle in state θ_2, notice that worker $w_{\bar{k}}$ plays the role of spoiler in state θ_2. Moreover, the firm and worker pair, $w_{\bar{k}}$ and $f_{\bar{k}}$, form the ***unique*** blocking pair to the matching resulting from the swap. Thus, in θ_2, the sub-market without worker $w_{\bar{k}}$ presents multiple stable matchings.

Now, consider the sub-market in state θ_2 that excludes $w_{\bar{k}}$. Worker $\bar{w}$ has two stable partners, f'_2 and $f_{\bar{k}}$ in this sub-market. In fact, that sub-market has two stable matchings. Thus, faced with the firm-proposing DA, if all other agents in the sub-market are truth-telling, $\bar{w}$ has an incentive to misrepresent his preferences. In particular, the best he can do through any strategy is to achieve his most preferred stable partner in the sub-market, firm f'_2. Consider then the following dropping strategy for $\bar{w}$: declare $f_{\bar{k}}$ as unacceptable, but report preferences truthfully otherwise. If $\bar{w}$ plays this dropping strategy, he can force the firm-proposing DA to implement the matching resulting from the swap, and thus obtain his most preferred stable partner in the sub-market (which is the best he can do).[4]

One way in which $\bar{w}$ effectively faces the sub-market that excludes $w_{\bar{k}}$ is if $w_{\bar{k}}$ is himself playing a dropping strategy in which he declares $f_{\bar{k}}$ as unacceptable,

[3]Naturally, we could have shown uniqueness more directly by illustrating that the worker-proposing DA generates $\mu(\theta_2)$ as well. We spell out these arguments as they will be useful in what follows.

[4]To see this, notice that if $\bar{w}$ uses a dropping strategy instead of being truthful, then $\bar{w}$ cannot be matched to $f_{\bar{k}}$ as result of the firm-proposing DA. Consequently, the matching resulting from the swap in the sub-market must be implemented. Notice also that this outcome cannot be achieved by any strategy where $\bar{w}$ declares $f_{\bar{k}}$ as acceptable.

and ranks all other firms truthfully, while all other agents report their preferences truthfully. This means that conditional on $w_{\bar{k}}$ playing a dropping strategy, it is a best response (in state θ_2) for $\bar{w}$ to play a dropping strategy.

The dropping strategy carries a cost for $\bar{w}$ in state θ_1 since, if $w_{\bar{k}}$ plays the dropping strategy, $\bar{w}$ would be effectively giving up $f_{\bar{k}}$ for a less desirable firm. If the utility of matching with f_2' is high enough, however, then playing the dropping strategy is a best response for $\bar{w}$, as it maximizes his expected utility when $w_{\bar{k}}$ uses the above dropping strategy, and all other workers and firms are truthful.

To show this profile constitutes an equilibrium we have to argue that it is also a best response for $w_{\bar{k}}$ to play a dropping strategy, and for the remaining agents to be truthful. Since the roles of $\bar{w}$ and $w_{\bar{k}}$ are reversed across states, the argument that a dropping strategy is optimal for $w_{\bar{k}}$, for appropriate choices of cardinal representation of preferences, is analogous. Indeed, if the state is θ_1 and $\bar{w}$ plays a dropping strategy, then $w_{\bar{k}}$ is facing a sub-market with multiple stable partners, in which case he has an incentive to misrepresent his preferences by playing a dropping strategy. If the state is θ_2, and $\bar{w}$ is playing the dropping strategy then $w_{\bar{k}}$ is giving up $f_{\bar{k}}$ for a less preferred alternative. If the utility of matching with $f_{\bar{k}+1}$ is high enough, then $w_{\bar{k}}$ playing the dropping strategy is a best response as it maximizes his expected utility given that $\bar{w}$ uses a dropping strategy, and all other workers and firms truthful.

Last, notice that being truthful is a weakly dominant strategy for firms in the firm-proposing DA, which we assumed they follow. For workers other than the spoiler and spoiled, reporting preferences truthfully is also a best-response given the profile of strategies under consideration. To see this, take an arbitrary worker $w \notin \{\bar{w}, w_{\bar{k}}\}$, and consider his incentives when $\bar{w}$ and $w_{\bar{k}}$ play the dropping strategies described above and the remaining workers are truthful. In each state, if w is honest, there is a unique stable matching. Consequently, w has no incentive to deviate from truthful reporting.

The profile we describe, in which $w_{\bar{k}}$ or $\bar{w}$ both use dropping strategies, therefore constitutes an equilibrium. In this equilibrium, $w_{\bar{k}}$ or $\bar{w}$ cannot be matched with their true stable partners in either θ_1 or θ_2, respectively, since they declare them unacceptable. Thus, the outcome of this equilibrium is necessarily (complete-information) unstable in each state. ■

Proof of Proposition 3.2. Towards a contradiction, assume there exists an equilibrium that induces an unstable matching $\lambda(\theta)$ for some state of the world θ. Then

there exists a firm $f_{i^*} \in \mathcal{F}$ and a worker $w_{j*} \in \mathcal{W}$ that block $\lambda(\theta)$. That is,

$$U^f_{i^*j^*}(\theta) > U^f_{i^*\lambda(\theta;i^*)}(\theta) \quad \text{and} \quad U^w_{i^*j^*}(\theta) > U^w_{\lambda(\theta;j^*)j^*}.$$

Since $U^f_{i^*j^*}(\theta) > U^f_{i^*\lambda(\theta;i^*)}(\theta)$, it follows that $w_{\lambda(\theta;i^*)}$ has a lower priority than w_{j^*} in the common preference of firms. This, in turn, means that firm i^* proposed to w_{j^*} and, in equilibrium, he rejected rejected f_{i^*} for a less desirable offer. Had he instead reported truthfully, he would have matched to someone ranked at least as highly as f_{i^*} in state θ.

Moreover, since state by state w_{j^*}'s option set (the set of firms he can match to) is independent of his report, in each of those states truthful reporting would achieve at least a good a partner as any other strategy. Therefore, w_{j^*} is not best responding, in contradiction. ■

Proof of Proposition 3.3. Let $\mu(\theta)$ denote the unique (complete-information) stable matching in state θ. Suppose there exists an equilibrium yielding the matching $\lambda(\theta)$ in each state θ such that $\lambda(\theta) \neq \mu(\theta)$ for at least one state θ. Consider the smallest integer k such that $\lambda(\theta; f_k) \neq \mu(\theta; f_k)$ for some state θ.

It follows that worker $w = \mu(f_k)$ either reports f_k as unacceptable or reports $\lambda(w)$ as preferable to f_k. Suppose that, in this equilibrium, w reports $\succ^*$ such that

$$f_{\pi(k+1)} \succ^* f_{\pi(k+2)} \succ^* ... \succ^* w \succ^* f_{\pi(j)} \succ^* ... \succ^* f_{\pi(n)}$$

for some permutation $\pi : \{k+1, ..., n\} \to \{k+1, ..., n\}$. We claim that the following deviation is profitable for w : a report of $\succ'$ such that

$$f_1 \succ' f_2 \succ' ... \succ' f_k \succ' f_{\pi(k+1)} \succ' f_{\pi(k+2)} \succ' ... \succ' w \succ' f_{\pi(j)} \succ' ... \succ' f_{\pi(n)}.$$

Let $\lambda'(\theta)$ be the resulting stable matching in each state θ.

Claim B.1. *In any state $\tilde{\theta}$ in which $\mu(\tilde{\theta}; w) = f_j$ for $j \in \{1, ..., k\}$, under worker w $\lambda'(\tilde{\theta}; w) \succeq_w \lambda(\tilde{\theta}; w) = f_j$.*

Proof of Claim B.1. First, suppose that $j = 1$. Under the reported preferences, w and f_1 are one another's favorites, and so $\lambda'(\tilde{\theta}; w) = f_1 \succeq_w \lambda(\tilde{\theta}; w)$.

Suppose now that $j > 1$. Without loss of generality, suppose that in state $\tilde{\theta}$, $\mu(\tilde{\theta}; f_i) = w_i$ for all i, so that $w = w_j$. Notice that, by construction, w_1 is f_1's

favorite, f_2 either prefers w_1 to w_2 or has w_2 as her favorite, f_3 can only prefer either w_1 or w_2 to w_3, and so on.

If $\lambda'(\tilde{\theta}; w_j) \prec_w \lambda(\tilde{\theta}; w_j)$, from stability of $\lambda'(\tilde{\theta})$ for the reported preferences, it must be the case that $\lambda'(\tilde{\theta}; f_j) = w_{j^{(1)}} \succ_{f_j} w_j$ and, therefore, $j^{(1)} < j$. By the minimality of k, $\lambda(\tilde{\theta}; w_{j^{(1)}}) = f_{j^{(1)}}$ and therefore, under the reported preferences, $f_{j^{(1)}} \succ_{w_{j^{(1)}}} f_j$ (else, $(f_j, w_{j^{(1)}})$ would block $\lambda(\tilde{\theta})$). From stability of $\lambda'(\tilde{\theta})$ it then follows that $\lambda'(\tilde{\theta}; f_{j^{(1)}}) = w_{j^{(2)}}$, where $j^{(2)} < j^{(1)}$. We can continue recursively till we reach $j^{(m)} = 1$. If, under the reported preferences, $f_{j^{(m-1)}} \succ_{w_1} f_1$ then $(f_{j^{(m-1)}}, w_1)$ block $\lambda(\tilde{\theta})$ under the reported preferences. f_1 has to be acceptable to w_1 under the reported preferences for $\lambda(\tilde{\theta})$ to be individually rational. It follows that $f_1 \succ_{w_1} f_{j^{(m-1)}}$, in which case (f_1, w_1) block $\lambda'(\tilde{\theta})$, achieving our contradiction. Figure 2 below describes the process, where arrows pointing to the right describe preferences of firms (the target node being the more preferred worker) derived from the stability of $\lambda'(\tilde{\theta})$ and arrows pointing to the left describe preferences of workers (again, the target node being the more preferred firm) derived from the stability of $\lambda(\tilde{\theta})$.

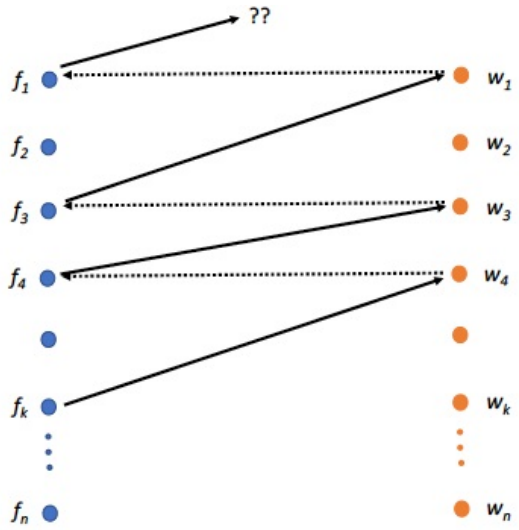


Figure B.1: Idea of Proof of Proposition 3.3

Claim B.2. *In any state $\bar{\theta}$ in which $\mu(\bar{\theta}; w) = f_j$ with $j > k$, under worker w's the original preferences, $\lambda'(\bar{\theta}; w) \succeq_w \lambda(\bar{\theta}; w)$.*

Proof of Claim B.2. From the definition of k and the structure of the deferred acceptance algorithm, $\lambda'(\bar{\theta}; f_i) = \lambda(\bar{\theta}; f_i) = \mu(\bar{\theta}; f_i)$ for $i = 1, ..., k-1$. Further-

more, $f_1, ..., f_{k-1}$ will not form a blocking pair with any worker in $\{\mu(\bar{\theta}; f_i)\}_{i=k}^n$, regardless of the preferences those workers report. Consider then the sub-market with firms $\{f_i\}_{i=k}^n$ and workers $\{\mu(\bar{\theta}; f_i)\}_{i=k}^n$, with preferences induced by the full market. It suffices to look at the firm-optimal stable matching in that sub-market. Consider then the deferred acceptance algorithm on this sub-market. If f_k makes an offer to w, $\lambda'(\bar{\theta}; w) = f_k \succ_w \lambda(\bar{\theta}; w)$. Otherwise, the deferred acceptance algorithm coincides with that corresponding to the original preferences in this sub-market and $\lambda'(\bar{\theta}; w) = \lambda(\bar{\theta}; w)$.

Last, in the state θ in which $\lambda(\theta; f_k) \neq \mu(\theta; f_k)$, under the reported preferences, $\lambda'(\theta; w) = f_k \succ_w \lambda(\theta; f_k)$. The proposition then follows. ■

Appendix C

APPENDICES TO CHAPTER 4

C.1 Proof for Theorem 4.1

The Hamilton-Jacobi-Bellman equation

$$\min\left\{\mathcal{L}(p), V(p) - \max_{v\in\{G,I\}} U(v,p)\right\} = 0, \tag{C.1}$$

where

$$\mathcal{L}(p) = \kappa - \frac{2}{\sigma^2}p^2(1-p)^2V''(p),$$

gives the sufficient condition for a continuously differentiable function $V\colon [0,1] \to \mathbb{R}$ to be the value function

$$V(p) = \sup_{(\tau,v)} \mathbb{E}\left[U(v,p_\tau) - \kappa\tau \mid p_0 = p\right]. \tag{C.2}$$

Differential equation $\mathcal{L}(p) = 0$ has the following solution:

$$V(p) = C_1 + pC_2 + \kappa\sigma^2 f(p),$$

where $f(p) = \left(p - \frac{1}{2}\right)\log\left(\frac{p}{1-p}\right)$ and C_1 and C_2 are some constants.

Consider the following class of functions defined on $p \in [0,1]$ parameterized with $\lambda \in (0, 0.5]$:

$$V_\lambda(p) = \begin{cases} pQ + R, & p \geq 1-\lambda, \\ (1-p)Q + R, & p \leq \lambda, \\ (1-\lambda)Q + R + \kappa\sigma^2\left(f(p) - f(\lambda)\right), & \text{otherwise.} \end{cases}$$

Note that these functions are continuous, symmetric around $p = 0.5$, that is $V_\lambda(p) = V_\lambda(1-p)$, and satisfy $\mathcal{L}(p) = 0$ for $\lambda < p < 1-\lambda$. Moreover, function $V_\lambda(p)$ is continuously differentiable if and only if

$$\lim_{p\to\lambda+0} V_\lambda'(p) = -Q,$$

which is equivalent to (4.3). Note that the left hand side of (4.3) is a decreasing function of $\lambda \in (0, 0.5]$ from $+\infty$ to 0. Thus, the solution to (4.3) always exists and unique.

Finally, note that

1. $V_\lambda(p) \geq \max\limits_{v\in\{G,I\}} U(v,p)$ for $\lambda < p < 1-\lambda$ since $V_\lambda(p)$ is convex for $\lambda < p < 1-\lambda$,

2. $\mathcal{L}(p) \geq 0$ for $p \geq 1-\lambda$ and $p \leq \lambda$ since the utility function is linear over the belief.

Thus, $V_\lambda(p)$ is the value function and therefore the strategy (4.2) is the unique optimal one.

C.2 Proof for Theorem 4.2

First, we calculate $\lambda'\left(\sigma^2\right)$ from (4.3) holding κ, Q, R fixed:

$$\lambda'\left(\sigma^2\right) = \frac{(1-\lambda)\lambda}{\sigma^2}\left(1-2\lambda-2(1-\lambda)\lambda\log\left(\frac{\lambda}{1-\lambda}\right)\right). \tag{C.3}$$

Substituting (C.3) into

$$\mathcal{X}'\left(\sigma^2\right) = \frac{1}{2}\left(\log\left(\frac{1-\lambda}{\lambda}\right) - \frac{\sigma^2}{(1-\lambda)\lambda}\lambda'\left(\sigma^2\right)\right), \tag{C.4}$$

we get $\mathcal{X}'\left(\sigma^2\right) = g\left(\lambda\left(\sigma^2\right)\right)$, where for any $\lambda \in (0, 0.5)$ function $g(\lambda)$ is defined as

$$g(\lambda) = \lambda - \frac{1}{2} + \left(\frac{1}{2} - (1-\lambda)\lambda\right)\log\left(\frac{1-\lambda}{\lambda}\right).$$

Function $g(\lambda)$ is decreasing in $\lambda \in (0, 0.5)$ from $+\infty$ to 0. Thus, it is always positive and therefore $\mathcal{X}'\left(\sigma^2\right) > 0$.

C.3 Proof for Theorem 4.3

Consider function $g(\eta) = \eta c(\eta) - \hat{c}$. $c'(\eta) > 0$ if $g(\eta) < 0$ and $c'(\eta) < 0$ if $g(\eta) > 0$. Since $g'(\eta)|_{g(\eta)=0} = c(\eta) > 0$, function $g(\eta)$ can cross 0 only once and only from below. From (4.14), if the solution to $g(\eta^*) = 0$ exists, it is equal to $\eta^* = \frac{1}{Qh'(\hat{c})}$. Substituting $\eta^* = \frac{1}{Qh'(\hat{c})}$ to (4.14), we get $c(\eta^*) = \frac{\hat{c}}{\eta^*}$ and thus $g(\eta^*) = 0$.

C.4 Proof for Theorem 4.4

Differentiating (4.17), we get $c'(\eta) = \frac{cQh'(\eta c)}{1-\eta h'(\eta c)Q} > 0$.

C.5 Proof for Theorem 4.5

For $\lambda \in (0, 0.5)$, denote by $\rho(\lambda) \in (0, +\infty)$ the unique solution to

$$4e^\rho\sqrt{\pi\rho} = \frac{1-2\lambda}{2\lambda(1-\lambda)} - \log\left(\frac{\lambda}{1-\lambda}\right). \tag{C.5}$$

For $\eta > 0$ and $\lambda \in (0, 0.5)$, denote

$$Y(\eta, \lambda) = \frac{1}{1 + \left(\frac{\lambda}{1-\lambda}\right)^{\frac{1}{\eta}}} - \left(\frac{1}{2} + \frac{1}{\sqrt{\pi}} \int_0^{\sqrt{\frac{\rho(\lambda)}{\eta}}} e^{-x^2} dx \right).$$

It is equal to zero if and only if $\rho(\lambda) = f(\eta, \lambda)$, where

$$f(\eta, \lambda) = \eta \left(\text{erf}^{-1} \left(\frac{2}{1 + \left(\frac{\lambda}{1-\lambda}\right)^{\frac{1}{\eta}}} - 1 \right) \right)^2,$$

where $\text{erf}^{-1}(\cdot)$ is the inverse error function.

Lemma C.1. *For any $\lambda \in (0, 0.5)$, function $f(\eta, \lambda)$ is decreasing in $\eta \in (0, +\infty)$ from $-\log\left(\frac{\lambda}{1-\lambda}\right)$ to 0.*

Proof. $\frac{\partial f(\eta,\lambda)}{\partial \eta} = g\left(\left(\frac{\lambda}{1-\lambda}\right)^{\frac{1}{\eta}}\right)$, where

$$g(x) = \text{erf}^{-1}\left(\frac{2}{1+x} - 1\right)\left(\text{erf}^{-1}\left(\frac{2}{1+x} - 1\right) + \frac{2\sqrt{\pi} x \log(x) e^{\left(\text{erf}^{-1}\left(\frac{2}{1+x}-1\right)\right)^2}}{(1+x)^2}\right).$$

Note that since $\lambda \in (0, 0.5)$ and $\eta > 0$, we must have $\left(\frac{\lambda}{1-\lambda}\right)^{\frac{1}{\eta}} \in (0, 1)$. As we can see from Figure C.1, $g(x) < 0$ for all $x \in (0, 1)$. Thus, $f(\eta, \lambda)$ is decreasing in $\eta \in (0, +\infty)$. ■

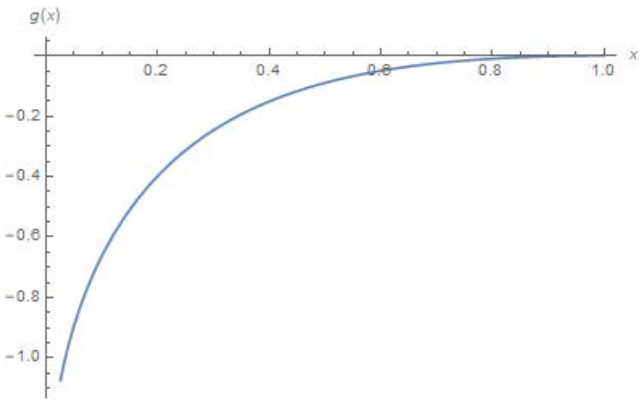


Figure C.1: Function $g(x)$.

For $\lambda \in (0, 0.5)$, denote

$$l(\lambda) = -\log\left(\frac{\lambda}{1-\lambda}\right) - \rho(\lambda).$$

Figure C.2 shows that function $l(\lambda)$ is always positive (it is decreasing from $+\infty$ to 0). Thus, the solution $\eta(\lambda) \in (0, +\infty)$ to $\rho(\lambda) = f(\eta, \lambda)$ always exists and it is unique.

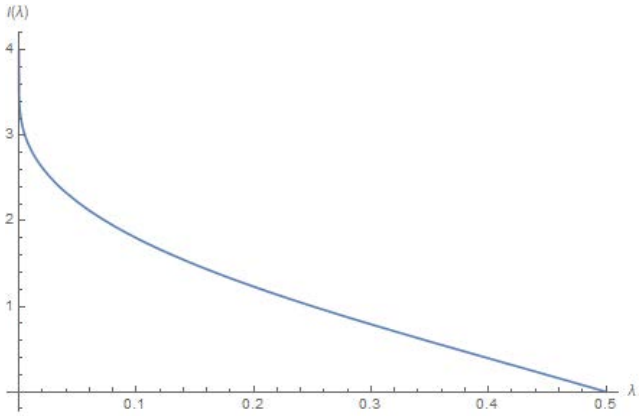


Figure C.2: Function $l(\lambda)$.

For $\lambda \in (0, 0.5)$, denote by $\eta(\lambda) \in (0, +\infty)$ the unique solution to $\rho(\lambda) = f(\eta, \lambda)$.

Lemma C.2. *Function $\eta(\lambda)$ is increasing from 1 to $\frac{\pi^2}{4}$.*

Proof.

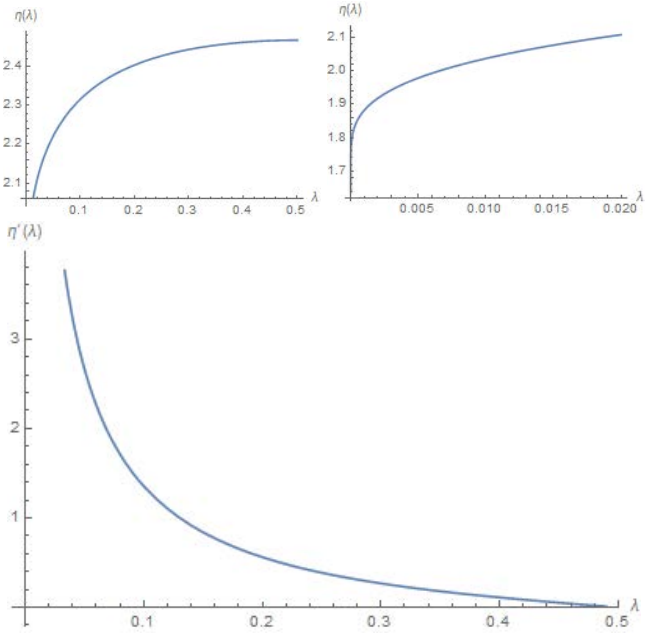


Figure C.3: Function $\eta(\lambda)$ and its derivative.

Figure C.3 shows that function $\eta(\lambda)$ is increasing.

From (C.5) we get $\lim_{\lambda\to 0}\rho(\lambda)=+\infty$, $\lim_{\lambda\to 0}\left(8e^{\rho(\lambda)}\sqrt{\pi\rho(\lambda)}\lambda\right)=1$ and therefore

$$\lim_{\lambda\to 0}\frac{\rho(\lambda)+\log(\lambda)}{\log\log\left(\frac{1}{\lambda}\right)}=-\frac{1}{2}. \tag{C.6}$$

Since $\eta(\lambda)$ is increasing, we have $\lim_{\lambda\to 0}\eta(\lambda)<+\infty$ and therefore $\lim_{\lambda\to 0}\left(\frac{\lambda}{1-\lambda}\right)^{\frac{1}{\eta(\lambda)}}=0$. Thus,

$$\lim_{\lambda\to 0}\frac{f(\eta(\lambda),\lambda)+\log(\lambda)}{\log\log\left(\frac{1}{\lambda}\right)}=-\frac{1}{2}\lim_{\lambda\to 0}\eta(\lambda). \tag{C.7}$$

(C.6) and (C.7) together imply $\lim_{\lambda\to 0}\eta(\lambda)=1$.

From (C.5) we get $\lim_{\lambda\to 0.5}\rho(\lambda)=0$, $\lim_{\lambda\to 0.5}\frac{e^{\rho(\lambda)}\sqrt{\pi\rho(\lambda)}}{1-2\lambda}=1$ and therefore

$$\lim_{\lambda\to 0.5}\frac{\rho(\lambda)}{(1-2\lambda)^2}=\frac{1}{\pi}. \tag{C.8}$$

Since $\eta(\lambda)$ is increasing, we have $\lim_{\lambda\to 0.5}\eta(\lambda)>0$ and therefore $\lim_{\lambda\to 0.5}\left(\frac{\lambda}{1-\lambda}\right)^{\frac{1}{\eta(\lambda)}}=1$. Thus,

$$\lim_{\lambda\to 0.5}\frac{f(\eta(\lambda),\lambda)}{(1-2\lambda)^2}=\frac{\pi}{4\lim_{\lambda\to 0.5}\eta(\lambda)}. \tag{C.9}$$

(C.8) and (C.9) together imply $\lim_{\lambda\to 0}\eta(\lambda)=\frac{\pi^2}{4}$. ■

For $\lambda\in(0,0.5)$, denote

$$F(\lambda)=\frac{1-2\lambda}{2\lambda(1-\lambda)}-\log\left(\frac{\lambda}{1-\lambda}\right).$$

It is easy to see that function $F(\lambda)$ is decreasing for $\lambda\in(0,0.5)$ from $+\infty$ to 0. Thus, since $\eta(\lambda)$ is increasing, there exists a unique solution $\lambda\left(\frac{Q}{\kappa\sigma^2}\right)\in(0,0.5)$ to $F(\lambda)=\frac{Q\eta(\lambda)}{\kappa\sigma^2}$.

Consider function $y\left(\eta,\frac{Q}{\kappa\sigma^2}\right)=\Pi^D\left(\eta,\frac{Q}{\kappa\sigma^2}\right)-\Pi^C\left(\eta,\frac{Q}{\kappa\sigma^2}\right)$. It is equal to zero if and only if $\eta=\eta\left(\lambda\left(\frac{Q}{\kappa\sigma^2}\right)\right)$. Moreover, from (4.19) and (4.21) we have $\lim_{\eta\to 0}y\left(\eta,\frac{Q}{\kappa\sigma^2}\right)=\frac{1}{1+e^{-\frac{Q}{2\kappa\sigma^2}}}-\frac{1}{2}>0$. So, $y\left(\eta,\frac{Q}{\kappa\sigma^2}\right)$ is positive for $\eta<\eta\left(\lambda\left(\frac{Q}{\kappa\sigma^2}\right)\right)$ and negative for $\eta>\eta\left(\lambda\left(\frac{Q}{\kappa\sigma^2}\right)\right)$.

C.6 Proof for Theorem 4.6

Since $\eta(\lambda)$ is increasing (Lemma C.2) and $F(\lambda)$ is decreasing, we conclude that $\lambda\left(\frac{Q}{\kappa\sigma^2}\right)$ is decreasing and therefore the threshold $\eta\left(\lambda\left(\frac{Q}{\kappa\sigma^2}\right)\right)$ is decreasing. It is easy to see that $\lim_{\frac{Q}{\kappa\sigma^2}\to 0}\lambda\left(\frac{Q}{\kappa\sigma^2}\right)=0.5$ and $\lim_{\frac{Q}{\kappa\sigma^2}\to+\infty}\lambda\left(\frac{Q}{\kappa\sigma^2}\right)=0$. Thus, $\lim_{\frac{Q}{\kappa\sigma^2}\to 0}\eta\left(\lambda\left(\frac{Q}{\kappa\sigma^2}\right)\right)=\frac{\pi^2}{4}$ and $\lim_{\frac{Q}{\kappa\sigma^2}\to+\infty}\eta\left(\lambda\left(\frac{Q}{\kappa\sigma^2}\right)\right)=1$.

C.7 Proof for Theorem 4.7

$\eta^{**}\left(\frac{Q}{\kappa\sigma^2}\right)>\eta^*$ if and only if

$$\eta\left(\lambda\left(\frac{Q}{\kappa\sigma^2}\right)\right)>\frac{2\sqrt{2e\pi}\kappa\sigma^2}{Q}. \tag{C.10}$$

Since $F(\lambda)=\frac{Q\eta(\lambda)}{\kappa\sigma^2}$, (C.10) is equivalent to

$$F\left(\lambda\left(\frac{Q}{\kappa\sigma^2}\right)\right)>2\sqrt{2e\pi}. \tag{C.11}$$

As $F(\lambda)$ is decreasing in $\lambda\in(0,0.5)$ from $+\infty$ to 0 and $\lambda\left(\frac{Q}{\kappa\sigma^2}\right)$ is decreasing in $\frac{Q}{\kappa\sigma^2}\in(0,+\infty)$ from 0.5 to 0, we conclude that $F\left(\lambda\left(\frac{Q}{\kappa\sigma^2}\right)\right)$ is increasing in $\frac{Q}{\kappa\sigma^2}\in(0,+\infty)$ from 0 to $+\infty$. Thus, there exists a unique solution $q\in(0,+\infty)$ such that (C.11) is equivalent to $\frac{Q}{\kappa\sigma^2}>q$.

C.8 Proof for Lemma 4.1

For a given threshold χ, strategy (4.6) leads to

$$\frac{\kappa\chi\left(e^{\frac{2\chi}{\sigma^2}}-1\right)}{e^{\frac{2\chi}{\sigma^2}}+1} \tag{C.12}$$

For the η-type decision maker, the optimal threshold is $\chi=\mathcal{X}\left(\frac{\sigma^2}{\eta}\right)$. Substituting this threshold into (C.12), we get the statement.

C.9 Proof for Theorem 4.8

Obviously, constraint (4.23) is binding, so that

$$R=\Upsilon\left(\frac{\kappa\sigma^2}{\eta},1,Q,M\right)-\Pi\left(1,\frac{Q\eta}{\kappa\sigma^2},M\right)Q. \tag{C.13}$$

Substituting (C.13) into (4.22), we get

$$\max_{Q,M}F(Q,M,\eta,\kappa\sigma^2,Q_P), \tag{C.14}$$

where

$$F(Q, M, \eta, \kappa\sigma^2, Q_P) = \\ \Pi\left(\eta, \frac{Q}{\kappa\sigma^2}, M\right)(Q_P - Q) + \Pi\left(1, \frac{Q\eta}{\kappa\sigma^2}, M\right)Q - \Upsilon\left(\frac{\kappa\sigma^2}{\eta}, 1, Q, M\right).$$

First, we optimize over Q for a given M.

Since

$$\lim_{Q\to+\infty} F(Q, M, \eta, \kappa\sigma^2, Q_P) = +\infty, \tag{C.15}$$

we conclude that $Q = +\infty$ is optimal for $\eta > 1$.

Moreover, when $\eta > 1$, we have

$$\lim_{Q\to+\infty} \frac{F(Q, 1, \eta, \kappa\sigma^2, Q_P)}{Q^{\frac{\eta-1}{\eta}}} = \left(\frac{\kappa\sigma^2}{2\eta}\right)^{\frac{1}{\eta}}, \tag{C.16}$$

$$\lim_{Q\to+\infty} \frac{F(Q, 0, \eta, \kappa\sigma^2, Q_P)}{\left(\frac{Q}{\sqrt{\log Q}}\right)^{\frac{\eta-1}{\eta}}} = \frac{\sqrt{\eta}}{2\sqrt{\pi}}\left(\frac{4\sqrt{\pi}\kappa\sigma^2}{\eta}\right)^{\frac{1}{\eta}}. \tag{C.17}$$

Comparing (C.16) and (C.17), we conclude that $M = 1$ is optimal for $\eta > 1$.

Substituting $Q \to +\infty$ and $M = 1$ into (C.13), we get $R \to -\infty$. More precisely,

$$\lim_{Q\to+\infty} \Upsilon\left(\frac{\kappa\sigma^2}{\eta}, 1, Q, M\right) - \Pi\left(1, \frac{Q\eta}{\kappa\sigma^2}, M\right)Q + Q - \frac{\kappa\sigma^2}{2\eta}\log(Q) \\ = \frac{\kappa\sigma^2}{2\eta} - \frac{\kappa\sigma^2}{2\eta}\log\left(\frac{\kappa\sigma^2}{2\eta}\right), \tag{C.18}$$

which implies $R = \left(\frac{\kappa\sigma^2}{2\eta} - \frac{\kappa\sigma^2}{2\eta}\log\left(\frac{\kappa\sigma^2}{2\eta}\right)\right) - \left(Q - \frac{\kappa\sigma^2}{2\eta}\log(Q)\right)$ in the optimum for $\eta > 1$.

C.10 Optimal Contract for $\eta \in (0, 1]$

Theorem 4.8 covers only the overconfidence case, $\eta > 1$. Here, we focus on $0 < \eta \leq 1$.

Theorem C.1. *For $0 < \eta \leq 1$, the optimal contract (Q, R, M) is unique and has the following form: $M = 1$,*

$$R\left(\eta, \kappa\sigma^2, Q_P\right) = \frac{\kappa\sigma^2}{2\eta}\left(2 - \frac{1}{\lambda} + \log\left(\frac{\lambda}{1-\lambda}\right)\right),$$

$$Q\left(\eta, \kappa\sigma^2, Q_P\right) = \frac{\kappa\sigma^2}{\eta}\left(\frac{1-2\lambda}{2\lambda(1-\lambda)} - \log\left(\frac{\lambda}{1-\lambda}\right)\right),$$

where $\lambda \in (0, 0.5)$ *is uniquely defined from*

$$\frac{\eta Q_P}{\kappa\sigma^2} = \frac{1-2\lambda}{2\lambda(1-\lambda)} - \log\left(\frac{\lambda}{1-\lambda}\right) + \frac{\eta}{2\lambda}\left(\left(\frac{\lambda}{1-\lambda}\right)^{1-\frac{1}{\eta}} - 1\right)\left(1 + \left(\frac{\lambda}{1-\lambda}\right)^{\frac{1}{\eta}}\right). \tag{C.19}$$

This contract gives the following expected utility to the principal:

$$U^P\left(\eta, \kappa\sigma^2, Q_P\right) = \frac{\kappa\sigma^2}{2\eta}\left(\frac{\eta}{\lambda}\left(\left(\frac{\lambda}{1-\lambda}\right)^{1-\frac{1}{\eta}} - 1\right) - 2 + \frac{1}{\lambda} - \log\left(\frac{\lambda}{1-\lambda}\right)\right).$$

The perceived expected utility of the agent is equal to his reservation utility, that is 0. The actual expected utility of the agent is $\Pi^D\left(\eta, \frac{Q}{\kappa\sigma^2}\right)Q + R - \Upsilon^D\left(\kappa\sigma^2, \eta, Q\right)$*, which is equal to*

$$U^A\left(\eta, \kappa\sigma^2, Q_P\right) = \frac{\kappa\sigma^2}{2\eta}\frac{1-2\lambda}{1-\lambda}\left(1 - \frac{1}{\lambda\left(1 + \left(\frac{1-\lambda}{\lambda}\right)^{\frac{1}{\eta}}\right)}\right).$$

Proof. We are going to use the notation $F(Q, M, \eta, \kappa\sigma^2, Q_P)$ from the proof of Theorem 4.8.

Since

$$\lim_{Q\to+\infty} F(Q, M, \eta, \kappa\sigma^2, Q_P) = -\infty,$$

we conclude that $Q = +\infty$ is *not* optimal for $\eta \leq 1$.

To optimize for Q when $\eta \leq 1$, we consider

$$\frac{\partial F(Q, M, \eta, \kappa\sigma^2, Q_P)}{\partial Q} = \begin{cases} \frac{2(1-\lambda)\lambda\left(\frac{\lambda}{1-\lambda}\right)^{\frac{1}{\eta}}}{\eta\left(1+\left(\frac{\lambda}{1-\lambda}\right)^{\frac{1}{\eta}}\right)^2}\left(\frac{\eta Q_P}{\kappa\sigma^2} - g(\lambda,\eta)\right), & M = 1, \\ \frac{e^{-\left(1+\frac{1}{\eta}\right)\rho}}{4\pi\sqrt{\eta}(1+2\rho)}\left(\frac{\eta Q_P}{\kappa\sigma^2} - h(\rho,\eta)\right), & M = 0, \end{cases} \tag{C.20}$$

where $\lambda \in (0, 0.5)$ solves (4.19), $\rho > 0$ solves (4.21) and

$$g(\lambda, \eta) = \frac{1-2\lambda}{2\lambda(1-\lambda)} - \log\left(\frac{\lambda}{1-\lambda}\right) - \frac{\eta}{2\lambda}\left(1 - \left(\frac{\lambda}{1-\lambda}\right)^{1-\frac{1}{\eta}}\right)\left(1 + \left(\frac{\lambda}{1-\lambda}\right)^{\frac{1}{\eta}}\right),$$

$$h(\rho, \eta) = 4e^{\rho}\sqrt{\pi}\left(\sqrt{\rho} + e^{\frac{\rho}{\eta}}\sqrt{\eta}(1+2\rho)\int_{\sqrt{\rho}}^{\sqrt{\rho/\eta}} e^{-x^2}dx\right).$$

Lemma C.3. *If $0 < \eta \leq 1$, then $g(\lambda, \eta)$ is strictly decreasing in $\lambda \in (0, 0.5)$ from $+\infty$ to 0.*

Proof. $\lim_{\lambda \to 0} g(\lambda, \eta) = +\infty$, $g(0.5, \eta) = 0$, $\frac{\partial g(\lambda,\eta)}{\partial \lambda} = \frac{\left(1+\left(\frac{\lambda}{1-\lambda}\right)^{-\frac{1}{\eta}}\right)(1-\eta\lambda)}{2(1-\lambda)^2\lambda} g_1(\lambda, \eta)$, where $g_1(\lambda, \eta) = \left(\frac{\lambda}{1-\lambda}\right)^{\frac{1}{\eta}-1} \frac{\eta(1-\lambda)-1}{1-\eta\lambda} - 1$.

Since $\frac{\partial g_1(\lambda,\eta)}{\partial \lambda} = -\frac{\left(\frac{\lambda}{1-\lambda}\right)^{\frac{1}{\eta}}\eta}{\lambda^2(1-\eta\lambda)^2}\left(\frac{(\eta-1)^2}{\eta^2} - (2\eta-3)(1-\lambda)\lambda\right) < 0$, $\lim_{\lambda \to 0} g_1(\lambda, \eta) = -1$ and $g_1(0.5, \eta) = -2$, we get that $g_1(\lambda, \eta)$ is strictly decreasing in $\lambda \in (0, 0.5)$ from -1 to -2.

■

Lemma C.4. *If $0 < \eta \leq 1$, then $h(\rho, \eta)$ is strictly increasing in $\rho > 0$ from 0 to $+\infty$.*

Proof. $\lim_{\rho \to 0} h(\rho, \eta) = 0$, $\lim_{\rho \to +\infty} h(\rho, \eta) = +\infty$, $\frac{\partial h(\rho,\eta)}{\partial \rho} = \frac{4e^{\rho+\frac{\rho}{\eta}}\sqrt{\pi}(1+3\eta+2(1+\eta)\rho)}{\sqrt{\eta}} h_1(\rho, \eta)$, where $h_1(\rho, \eta) = \frac{\frac{e^\rho}{e^{\rho/\eta}\sqrt{\eta}} - \frac{1}{2}}{e^\rho\sqrt{\rho}\left(\frac{1}{\eta}+1+\frac{2}{1+2\rho}\right)} + \int_{\sqrt{\rho}}^{\sqrt{\rho/\eta}} e^{-x^2} dx$.

$h_1(\rho, \eta) > 0$ for all $0 < \eta \leq 1$, $\rho > 0$ since:

- if $\frac{e^\rho}{e^{\rho/\eta}\sqrt{\eta}} > \frac{1}{2}$, then $h_1(\rho, \eta) > 0$;
- $\lim_{\rho \to 0} h_1(\rho, \eta) = +\infty$, $\lim_{\rho \to +\infty} h_1(\rho, \eta) = 0$;
- $\frac{\partial h_1(\rho,\eta)}{\partial \rho} = -\frac{\left(\frac{3-\eta}{2}+\frac{\eta}{1+2\rho}\right)\left(1+\eta+\frac{2\eta}{1+2\rho}\right)+\frac{\rho+\eta-\eta^2\left(\frac{4(\rho-1)}{(1+2\rho)^2}+1+\rho\right)}{\rho}\left(\frac{e^\rho}{e^{\rho/\eta}\sqrt{\eta}}-\frac{1}{2}\right)}{2e^\rho\sqrt{\rho}\left(1+\eta+\frac{2\eta}{1+2\rho}\right)^2}$;
- if $\frac{e^\rho}{e^{\rho/\eta}\sqrt{\eta}} \leq \frac{1}{2}$ and $\rho + \eta - \eta^2\left(\frac{4(\rho-1)}{(1+2\rho)^2} + 1 + \rho\right) \leq 0$, then $\frac{\partial h_1(\rho,\eta)}{\partial \rho} < 0$;
- if $\rho+\eta-\eta^2\left(\frac{4(\rho-1)}{(1+2\rho)^2} + 1 + \rho\right) > 0$, then $\frac{\partial h_1(\rho,\eta)}{\partial \rho} < -\frac{h_2(\rho,\eta)}{2e^\rho\sqrt{\rho}\left(1+\eta+\frac{2\eta}{1+2\rho}\right)^2}$, where $h_2(\rho, \eta) = 1 - \frac{2\eta^2\left(3-\left(\rho+\frac{3}{2}\right)^2\right)}{\rho(1+2\rho)^2} + \eta\left(1 - \frac{1}{2\rho} + \frac{4}{1+2\rho}\right)$;
- if $0 < \eta < 2\rho$, then $h_2(\rho, \eta) > 0$;
- if $\frac{e^\rho}{e^{\rho/\eta}\sqrt{\eta}} \leq \frac{1}{2}$, then $0 < \eta < 2\rho$.

■

Lemma C.3 and derivative (C.20) imply that if $0 < \eta \leq 1$, then $F(Q, 1, \eta, \kappa\sigma^2, Q_P)$ is strictly increasing in $Q \in (0, Q^D)$ and it is strictly decreasing in $Q \in (Q^D, +\infty)$, where $Q^D > 0$ is defined from

$$\frac{\eta Q^D}{\kappa\sigma^2} = \frac{1-2\lambda}{2\lambda(1-\lambda)} - \log\left(\frac{\lambda}{1-\lambda}\right), \tag{C.21}$$

where $\lambda \in (0, 0.5)$ is uniquely defined from

$$\frac{\eta Q_P}{\kappa\sigma^2} = g(\lambda, \eta). \tag{C.22}$$

(C.21) and (C.22) imply $Q_P \geq Q^D$ for $0 < \eta \leq 1$.

Lemma C.4 and derivative (C.20) imply that if $0 < \eta \leq 1$, then $F(Q, 0, \eta, \kappa\sigma^2, Q_P)$ is strictly increasing in $Q \in (0, Q^C)$ and it is strictly decreasing in $Q \in (Q^C, +\infty)$, where $Q^C > 0$ is defined from

$$\frac{\eta Q^C}{\kappa\sigma^2} = 4e^{\rho}\sqrt{\pi\rho}, \tag{C.23}$$

where $\rho > 0$ is uniquely defined from

$$\frac{\eta Q_P}{\kappa\sigma^2} = h(\rho, \eta). \tag{C.24}$$

(C.23) and (C.24) imply $Q_P \geq Q^C$ for $0 < \eta \leq 1$.

To maximize over M, we compare $F(Q, 1, \eta, \kappa\sigma^2, Q_P)$ and $F(Q, 0, \eta, \kappa\sigma^2, Q_P)$ for $0 < \eta \leq 1$ and $0 < Q \leq Q_P$.

Lemma C.5. *$F(Q, 1, \eta, \kappa\sigma^2, Q_P) > F(Q, 0, \eta, \kappa\sigma^2, Q_P)$ for all $0 < \eta \leq 1$ and $0 < Q \leq Q_P$.*

Proof. By Theorem 4.5, we have $\Pi^D\left(\eta, \frac{Q}{\kappa\sigma^2}\right) > \Pi^C\left(\eta, \frac{Q}{\kappa\sigma^2}\right)$ for all $0 < \eta \leq 1$.

The agent's expected utility from contract (Q, R, M) is

$$\Pi\left(1, \frac{Q\eta}{\kappa\sigma^2}, M\right) Q + R - \Upsilon\left(\frac{\kappa\sigma^2}{\eta}, 1, Q, M\right).$$

Obviously, the agent strictly prefers $M = 1$ over $M = 0$ since he thinks he is better off with more flexibility. Thus,

$$\Pi^D\left(1, \frac{Q\eta}{\kappa\sigma^2}\right) Q - \Upsilon^D\left(\frac{\kappa\sigma^2}{\eta}, 1, Q\right) > \Pi^C\left(1, \frac{Q\eta}{\kappa\sigma^2}\right) Q - \Upsilon^C\left(\frac{\kappa\sigma^2}{\eta}, 1, Q\right).$$

■

Using $Q_P \geq Q^D$ and $Q_P \geq Q^C$ for $0 < \eta \leq 1$ and Lemma C.5, we have

$$\begin{aligned} F(Q^D, 1, \eta, \kappa\sigma^2, Q_P) - F(Q^C, 0, \eta, \kappa\sigma^2, Q_P) &= F(Q^D, 1, \eta, \kappa\sigma^2, Q_P) \\ - F(Q^C, 1, \eta, \kappa\sigma^2, Q_P) + F(Q^C, 1, \eta, \kappa\sigma^2, Q_P) &- F(Q^C, 0, \eta, \kappa\sigma^2, Q_P) \\ \geq F(Q^C, 1, \eta, \kappa\sigma^2, Q_P) - F(Q^C, 0, \eta, \kappa\sigma^2, Q_P) &> 0. \end{aligned}$$

Thus, $M = 1$ is optimal for $\eta \leq 1$.

Substituting Q^D and $M = 1$ into (C.13), we get

$$R = \frac{\kappa\sigma^2}{2\eta}\left(2 - \frac{1}{\lambda} + \log\left(\frac{\lambda}{1-\lambda}\right)\right),$$

where $\lambda \in (0, 0.5)$ is uniquely defined from (C.22). Note that $R < 0$.

■

To give some intuition for the optimal contract provided in Theorem C.1, we provide the comparative statics results. Corollary C.1 covers all cases when a variable of interest is monotone with respect to a parameter.

Corollary C.1. *For* $0 < \eta < 1$,

- $R\left(\eta, \kappa\sigma^2, Q_P\right)$ *is decreasing in* η *and* Q_P*;*
- $Q\left(\eta, \kappa\sigma^2, Q_P\right)$ *is increasing in* η*,* Q_P *and* $\kappa\sigma^2$*;*
- $Q\left(\eta, \kappa\sigma^2, Q_P\right) + R\left(\eta, \kappa\sigma^2, Q_P\right)$ *is increasing in* Q_P *and* $\kappa\sigma^2$*;*
- $U^P\left(\eta, \kappa\sigma^2, Q_P\right)$ *is increasing in* η *and* Q_P *but decreasing in* $\kappa\sigma^2$*;*
- $U^A\left(\eta, \kappa\sigma^2, Q_P\right)$ *is decreasing in* η *but increasing in* Q_P*;*
- $U^P\left(\eta, \kappa\sigma^2, Q_P\right) + U^A\left(\eta, \kappa\sigma^2, Q_P\right)$ *is increasing in* Q_P *but decreasing in* $\kappa\sigma^2$*.*

Proof. $-\frac{1}{\kappa\sigma^2}\frac{\partial R\left(\eta,\kappa\sigma^2,Q_P\right)}{\partial\eta}, -\frac{\partial R\left(\eta,\kappa\sigma^2,Q_P\right)}{\partial Q_P}, \frac{1}{\kappa\sigma^2}\frac{\partial Q\left(\eta,\kappa\sigma^2,Q_P\right)}{\partial\eta}, \frac{\partial Q\left(\eta,\kappa\sigma^2,Q_P\right)}{\partial(\kappa\sigma^2)}, \frac{\partial Q\left(\eta,\kappa\sigma^2,Q_P\right)}{\partial Q_P},$

$\frac{\partial\left(Q\left(\eta,\kappa\sigma^2,Q_P\right)+R\left(\eta,\kappa\sigma^2,Q_P\right)\right)}{\partial(\kappa\sigma^2)}, \frac{\partial\left(Q\left(\eta,\kappa\sigma^2,Q_P\right)+R\left(\eta,\kappa\sigma^2,Q_P\right)\right)}{\partial Q_P},$

$\frac{1}{\kappa\sigma^2}\frac{\partial U^P\left(\eta,\kappa\sigma^2,Q_P\right)}{\partial\eta}, -\frac{\partial U^P\left(\eta,\kappa\sigma^2,Q_P\right)}{\partial(\kappa\sigma^2)}, \frac{\partial U^P\left(\eta,\kappa\sigma^2,Q_P\right)}{\partial Q_P},$

$-\frac{1}{\kappa\sigma^2}\frac{\partial U^A\left(\eta,\kappa\sigma^2,Q_P\right)}{\partial\eta}, \frac{\partial U^A\left(\eta,\kappa\sigma^2,Q_P\right)}{\partial Q_P},$

$-\frac{\partial\left(U^P\left(\eta,\kappa\sigma^2,Q_P\right)+U^A\left(\eta,\kappa\sigma^2,Q_P\right)\right)}{\partial(\kappa\sigma^2)}$, $\frac{\partial\left(U^P\left(\eta,\kappa\sigma^2,Q_P\right)+U^A\left(\eta,\kappa\sigma^2,Q_P\right)\right)}{\partial Q_P}$ can be expressed as functions of η and $\lambda \in (0, 0.5)$ that solves (C.19). Each of these functions are positive for all $\eta \in (0, 1)$ and $\lambda \in (0, 0.5)$.

■

The comparative statics with respect to the underconfidence parameter η is the most interesting one. As the agent become more underconfident (η is decreasing), he has to be compensated more to agree to take the risk of the contract (R becomes larger), as he thinks he has to spend a lot of effort collecting what he thinks being very noisy signals. At the same time, his benefit compensation Q is decreasing, as the agent's strategy becomes less sensitive to it (in the extreme case, when $\eta \to 0$, the agent does not collect any information at all, no matter how large Q is). This irrationality of the agent decreases the principal's expected payoff from the contract (in the extreme case, when $\eta = 1$, the agent gets 0 and collects the optimal amount of information from the rational person point of view, which gives the highest utility to the principal).

The most interesting result is that the agent "wins" from his irrationality: a more underconfident agent gets higher actual expected utility. This happens because the principal has to persuade the agent to agree to a deal that has a higher return than the agent thinks.

Whether it is socially optimal for the agent to be more underconfident (that is, whether $U^P + U^A$ is decreasing in η) depends on the parameters' values. Figure C.4 demonstrates this point. For most values, the social welfare is decreasing as the agent becomes more underconfident. However, when it is socially optimal to collect a lot of information (that is when $\frac{Q_P}{\kappa\sigma^2}$ is high), increasing the level of underconfidence might increase the total expected utility.

C.11 Proof for Theorem 4.9

By Theorems 4.8 and C.1, the constraint $R \geq 0$ is binding. Thus, $R = 0$ and the optimization problem (4.24)-(4.25) becomes

$$\max_{Q\geq 0, M\in\{0,1\}} \quad \Pi\left(\eta, \frac{Q}{\kappa\sigma^2}, M\right)(Q_P - Q) \equiv F(Q, M, \eta, \kappa\sigma^2, Q_P), \quad \text{(C.25)}$$

$$\text{s.t.} \quad \Pi\left(1, \frac{Q\eta}{\kappa\sigma^2}, M\right)Q - \Upsilon\left(\frac{\kappa\sigma^2}{\eta}, 1, Q, M\right) \geq 0. \quad \text{(C.26)}$$

Constraint (C.26) says that the maximum subjective expected utility the agent can get cannot be less than 0. Since he gets $R = 0$ from the wrong decision and $R + Q \geq 0$

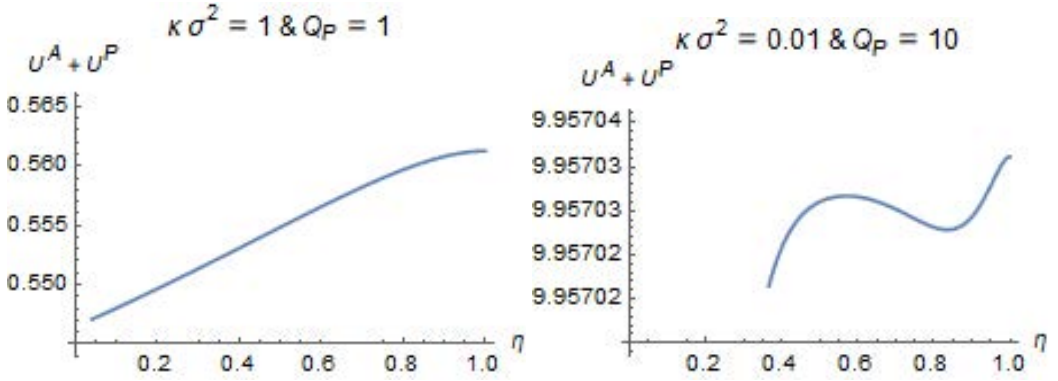


Figure C.4: Social welfare $U^P(\eta, \kappa\sigma^2, Q_P) + U^A(\eta, \kappa\sigma^2, Q_P)$ as a function of η.

from the correct decision, this expected utility must be greater or equal to zero for any $Q \geq 0$ and $M \in \{0, 1\}$. Thus, the optimization problem (4.24)-(4.25) becomes (C.25), without any constraints.

Optimizing over $Q \geq 0$ for a given M, we find the first order condition for (C.25):

$$\frac{\partial F(Q, M, \eta, \kappa\sigma^2, Q_P)}{\partial Q} = \begin{cases} \frac{2(1-\lambda)\lambda\left(\frac{\lambda}{1-\lambda}\right)^{\frac{1}{\eta}}}{\eta\left(1+\left(\frac{\lambda}{1-\lambda}\right)^{\frac{1}{\eta}}\right)^2}\left(\frac{\eta Q_P}{\kappa\sigma^2} - g(\lambda,\eta)\right), & M = 1, \\ \frac{e^{-\left(1+\frac{1}{\eta}\right)\rho}}{4\pi\sqrt{\eta}(1+2\rho)}\left(\frac{\eta Q_P}{\kappa\sigma^2} - h(\rho,\eta)\right), & M = 0, \end{cases}$$

where $\lambda \in (0, 0.5)$ solves (4.19), $\rho > 0$ solves (4.21) and $g(\lambda, \eta)$ is the right hand side of (4.26), while $h(\rho, \eta)$ is the right hand side of (4.27).

Lemma C.6. *Function $g(\lambda, \eta)$ is strictly decreasing in $\lambda \in (0, 0.5)$ from $+\infty$ to 4η. Function $h(\rho, \eta)$ is strictly increasing in $\rho > 0$ from $2\pi\sqrt{\eta}$ to $+\infty$.*

Proof. $\frac{\partial g(\lambda,\eta)}{\partial\lambda} = -\frac{1+\eta(1-2\lambda)}{2(1-\lambda)^2\lambda^2}\left(1 + \left(\frac{\lambda}{1-\lambda}\right)^{-\frac{1}{\eta}}\right) < 0,$

$$\frac{\partial h(\rho,\eta)}{\partial\rho} = 4e^{\rho}\sqrt{\pi}\left(\frac{1+2\rho}{\sqrt{\rho}} + e^{\frac{\rho}{\eta}}\frac{1+3\eta+2(1+\eta)\rho}{\sqrt{\eta}}\int_{-\infty}^{\sqrt{\rho/\eta}} e^{-x^2}dx\right) > 0. \qquad \blacksquare$$

Summarizing, we have:

- if $\frac{\eta Q_P}{\kappa\sigma^2} \leq \min\left\{4\eta, 2\pi\sqrt{\eta}\right\}$, then $F(Q, M, \eta, \kappa\sigma^2, Q_P)$ is strictly decreasing in $Q \Rightarrow$ the optimal Q is zero (the value of M is irrelevant in this case since there will be no information collection);
- if $4\eta < \frac{\eta Q_P}{\kappa\sigma^2} \leq 2\pi\sqrt{\eta}$, then $F(Q, 0, \eta, \kappa\sigma^2, Q_P)$ is strictly decreasing in Q while $F(Q, 1, \eta, \kappa\sigma^2, Q_P)$ is strictly increasing in $Q \in (0, Q^D)$ and it is

strictly decreasing in $Q \in (Q^D, +\infty) \Rightarrow$ the optimal $M = 1$ and $Q = Q^D$ (note that $F(0, M, \eta, \kappa\sigma^2, Q_P) = \frac{Q_P}{2}$ does not depend on M);

- if $2\pi\sqrt{\eta} < \frac{\eta Q_P}{\kappa\sigma^2} \leq 4\eta$, then, by similar reasoning, the optimal $M = 0$ and $Q = Q^C$;

- if $\max\left\{4\eta, 2\pi\sqrt{\eta}\right\} < \frac{\eta Q_P}{\kappa\sigma^2}$, then the optimal Q given M is Q^D if $M = 1$ and Q^C if $M = 0$ (note that $Q^D < Q_P$ and $Q^C < Q_P$), and it is not obvious whether $M = 1$ or $M = 0$ is optimal. To find it out, we need to find the sign of the function

$$f\left(\eta, \kappa\sigma^2, Q_P\right) = \max_{Q \geq 0} F(Q, 1, \eta, \kappa\sigma^2, Q_P) - \max_{Q \geq 0} F(Q, 0, \eta, \kappa\sigma^2, Q_P)$$
$$= \frac{\kappa\sigma^2}{2\lambda(1-\lambda)}\left(\frac{\lambda}{1-\lambda}\right)^{-\frac{1}{\eta}} - \frac{4e^{\rho+\frac{\rho}{\eta}}(1+2\rho)\kappa\sigma^2}{\sqrt{\eta}}\left(\int_{-\infty}^{\sqrt{\rho/\eta}} e^{-x^2}dx\right)^2,$$

where $\lambda \in (0, 0.5)$ solves (4.26) and $\rho > 0$ solves (4.27).

Lemma C.7. *If $\eta \geq \frac{\pi^2}{4}$, then*

- *if $\frac{\eta Q_P}{\kappa\sigma^2} \leq 2\pi\sqrt{\eta}$, then the optimal Q is zero;*

- *if $\frac{\eta Q_P}{\kappa\sigma^2} > 2\pi\sqrt{\eta}$, then the optimal $M = 0$ and $Q = Q^C$.*

Proof. First of all, note that $\eta \geq \frac{\pi^2}{4}$ is equivalent to $2\pi\sqrt{\eta} \leq 4\eta$. Thus, for $\frac{\eta Q_P}{\kappa\sigma^2} \leq 4\eta$, the statement follows from the summary above.

From Theorems 4.5 and 4.6 it follows that $\Pi^D\left(\eta, \frac{Q}{\kappa\sigma^2}\right) < \Pi^C\left(\eta, \frac{Q}{\kappa\sigma^2}\right)$ for all $\eta \geq \frac{\pi^2}{4}$ and $Q > 0$. Since $Q^D < Q_P$, this implies

$$f\left(\eta, \kappa\sigma^2, Q_P\right) = F(Q^D, 1, \eta, \kappa\sigma^2, Q_P) - F(Q^C, 0, \eta, \kappa\sigma^2, Q_P)$$
$$= F(Q^D, 1, \eta, \kappa\sigma^2, Q_P) - F(Q^D, 0, \eta, \kappa\sigma^2, Q_P) + F(Q^D, 0, \eta, \kappa\sigma^2, Q_P)$$
$$- F(Q^C, 0, \eta, \kappa\sigma^2, Q_P) \leq F(Q^D, 1, \eta, \kappa\sigma^2, Q_P) - F(Q^D, 0, \eta, \kappa\sigma^2, Q_P)$$
$$= \left(\Pi^D\left(\eta, \frac{Q^D}{\kappa\sigma^2}\right) - \Pi^C\left(\eta, \frac{Q^D}{\kappa\sigma^2}\right)\right)(Q_P - Q^D) < 0.$$

■

Lemma C.8. *If $\eta \leq 1$, then*

- *if* $\frac{\eta Q_P}{\kappa\sigma^2} \leq 4\eta$*, then the optimal* Q *is zero;*
- *if* $\frac{\eta Q_P}{\kappa\sigma^2} > 4\eta$*, then the optimal* $M = 1$ *and* $Q = Q^D$*.*

The proof of Lemma C.8 is similar to the proof of Lemma C.7.

Lemma C.9. *If* $1 < \eta < \frac{\pi^2}{4}$*, then there exists* $q(\eta) > \frac{2\pi}{\sqrt{\eta}}$ *such that*

- *if* $\frac{\eta Q_P}{\kappa\sigma^2} \leq 4\eta$*, then the optimal* Q *is zero;*
- *if* $4\eta < \frac{\eta Q_P}{\kappa\sigma^2} < \eta q(\eta)$*, then the optimal* $M = 1$ *and* $Q = Q^D$*;*
- *if* $\frac{\eta Q_P}{\kappa\sigma^2} > \eta q(\eta)$*, then the optimal* $M = 0$ *and* $Q = Q^C$*.*

Proof. For $\frac{\eta Q_P}{\kappa\sigma^2} \leq 2\pi\sqrt{\eta}$, the statement follows from the summary above.

Thus, when $\frac{\eta Q_P}{\kappa\sigma^2} = 2\pi\sqrt{\eta}$, we have $f\left(\eta, \kappa\sigma^2, Q_P\right) > 0$. Moreover, it is not hard to proof that $\lim_{Q_P \to +\infty} f\left(\eta, \kappa\sigma^2, Q_P\right) = -\infty$. Thus, function $f\left(\eta, \kappa\sigma^2, Q_P\right)$ crosses 0 at least once on $Q_P > \frac{2\pi\kappa\sigma^2}{\sqrt{\eta}}$, and the first crossing must be from above (from positive values). To prove that $f\left(\eta, \kappa\sigma^2, Q_P\right)$ crosses 0 exactly once, it is sufficient to show that all local optima with respect to Q_P are local maximums: if $\frac{\partial f(\eta, \kappa\sigma^2, Q_P)}{\partial Q_P} = 0$, then $\frac{\partial^2 f(\eta, \kappa\sigma^2, Q_P)}{\partial Q_P^2} < 0$.

$$\frac{\partial f(\eta, \kappa\sigma^2, Q_P)}{\partial Q_P} = \Pi^D\left(\eta, \frac{Q^D}{\kappa\sigma^2}\right) - \Pi^C\left(\eta, \frac{Q^C}{\kappa\sigma^2}\right)$$

If $\frac{\partial f(\eta, \kappa\sigma^2, Q_P)}{\partial Q_P} = 0$, then

1. $\lambda = \Lambda(\rho, \eta) \equiv \frac{1}{1 + \left(\frac{2}{\text{erfc}\left(\sqrt{\frac{\rho}{\eta}}\right)} - 1\right)^{\eta}}$, where $\text{erfc}\,(z) = 1 - \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt$ is the complementary error function, $\rho > 0$ solves (4.27) and $\lambda \in (0, 0.5)$ solves (4.26).
2. $L(\rho, \eta) = 0$, where $\rho > 0$ solves (4.27) and function $L(\rho, \eta)$ is defined as follows:

$$L(\rho, \eta) = \frac{1 - \sqrt{1 + \left(l(\rho, \eta) - \frac{4\eta}{\text{erfc}\left(\sqrt{\frac{\rho}{\eta}}\right)}\right) l(\rho, \eta)}}{2l(\rho, \eta)} + \frac{1}{2} - \Lambda(\rho, \eta),$$

$$l(\rho, \eta) = h(\rho, \eta) + \eta \log\left(\frac{2}{2 - \text{erfc}\left(\sqrt{\frac{\rho}{\eta}}\right)} - 1\right).$$

One can check that function $L(\rho,\eta)$ is strictly increasing in $\eta \in \left(1, \frac{\pi^2}{4}\right)$ from $L(\rho,1) < 0$ to $L\left(\rho, \frac{\pi^2}{4}\right) > 0$. Thus, there exists a unique solution $\tilde{\eta}(\rho) \in \left(1, \frac{\pi^2}{4}\right)$ to $L(\rho,\eta) = 0$.

The second order derivative $\frac{\partial^2 f(\eta,\kappa\sigma^2,Q_P)}{\partial Q_P^2}$ can be written as $\frac{1}{\kappa\sigma^2} f_2(\rho,\lambda,\eta)$, where $\rho > 0$ solves (4.27) and $\lambda \in (0, 0.5)$ solves (4.26). If $\frac{\partial f(\eta,\kappa\sigma^2,Q_P)}{\partial Q_P} = 0$, then this derivative is equal to $\frac{1}{\kappa\sigma^2} f_2(\rho,\Lambda(\rho,\eta\tilde{\eta}(\rho)),\tilde{\eta}(\rho)) \equiv \frac{1}{\kappa\sigma^2} F_2(\rho)$ for a certain $\rho > 0$. Figure C.5 shows that $F_2(\rho) < 0$ for all $\rho > 0$.

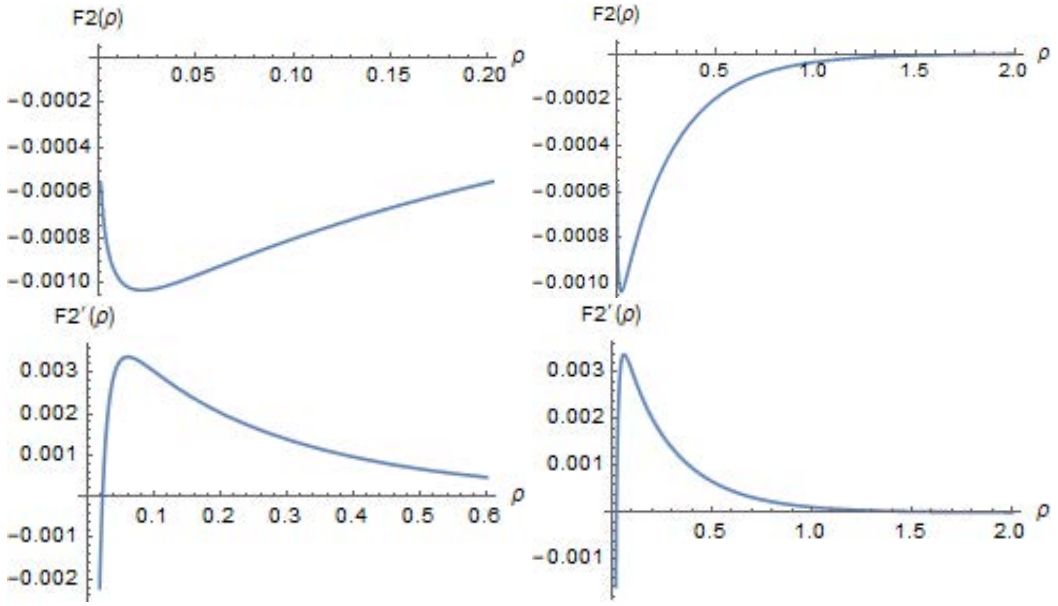


Figure C.5: Function $F_2(\rho)$ and its derivative for $\rho > 0$. Note that $\lim_{\rho\to 0} F_2(\rho) = 0$ and $\lim_{\rho\to+\infty} F_2(\rho) = 0$

That proves that if there is a local optimum of function $f(\eta,\kappa\sigma^2,Q_P)$ with respect to $Q_P > 0$, then this optimum must be a local maximum. Thus, we conclude that there exists a unique $Q_P > \frac{2\pi\kappa\sigma^2}{\sqrt{\eta}}$ such that $f(\eta,\kappa\sigma^2,Q_P) = 0$. From the definition of function $f(\eta,\kappa\sigma^2,Q_P)$, it immediately follows that such Q_P is equal to $\kappa\sigma^2$, times some function that depends only on η. Denote this function as $q(\eta)$.

■

Lemma C.10. *Function $q(\eta)$ is strictly decreasing in $1 < \eta < \frac{\pi^2}{4}$ from $+\infty$ to $\frac{2\pi}{\sqrt{\eta}}$.*

Proof.

$$q'(\eta) = \frac{\Pi^D\left(\eta, \frac{Q^D}{\kappa\sigma^2}\right) q_1(\lambda) - \Pi^C\left(\eta, \frac{Q^C}{\kappa\sigma^2}\right) q_2(\rho)}{\eta^2\left(\Pi^D\left(\eta, \frac{Q^D}{\kappa\sigma^2}\right) - \Pi^C\left(\eta, \frac{Q^C}{\kappa\sigma^2}\right)\right)},$$

$$q_1(\lambda) = \left(1 - \frac{1}{2\lambda(1-\lambda)}\right)\log\left(\frac{\lambda}{1-\lambda}\right) - \frac{1-2\lambda}{2\lambda(1-\lambda)}, \quad q_2(\rho) = 2e^{\rho}\sqrt{\pi\rho}(2\rho - 1),$$

where $\rho > 0$ solves (4.27), $\lambda \in (0, 0.5)$ solves (4.26) and $\frac{Q_P}{\kappa\sigma^2} = q(\eta)$.

Since $\frac{\partial f(\eta, \kappa\sigma^2, Q_P)}{\partial Q_P} < 0$ at the point $Q_P = \kappa\sigma^2 q(\eta)$, we have $\Pi^D\left(\eta, \frac{Q^D}{\kappa\sigma^2}\right) < \Pi^C\left(\eta, \frac{Q^C}{\kappa\sigma^2}\right)$. Figure C.6 shows that $q'(\eta)$ is negative.[1]

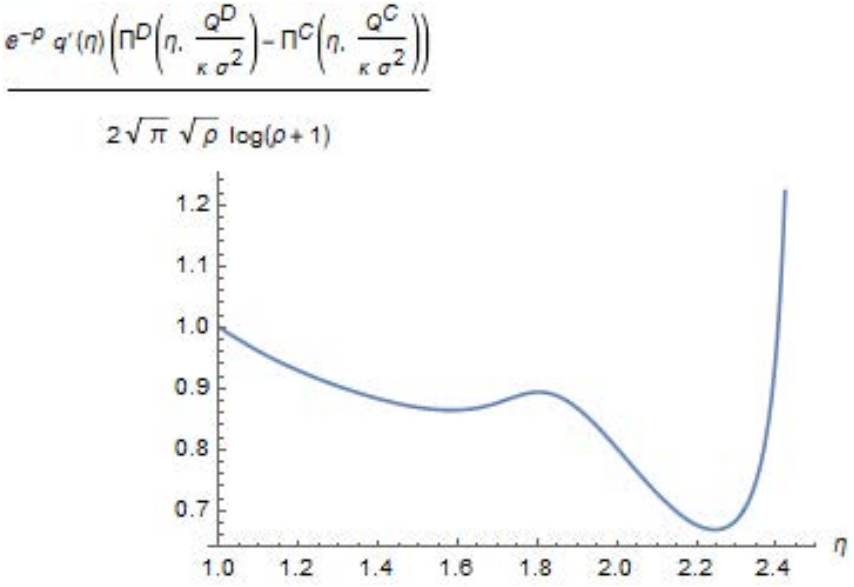


Figure C.6: Function $\frac{q'(\eta)\left(\Pi^D\left(\eta, \frac{Q^D}{\kappa\sigma^2}\right) - \Pi^C\left(\eta, \frac{Q^C}{\kappa\sigma^2}\right)\right)}{2\sqrt{\pi}e^{\rho}\sqrt{\rho}\log(\rho+1)}$, where $\rho > 0$ solves (4.27) and $\frac{Q_P}{\kappa\sigma^2} = q(\eta)$, for $1 < \eta < \frac{\pi^2}{4}$

■

C.12 Proof for Theorem 4.10

The principal's expected utility is equal to

$$U^P\left(\eta, \kappa\sigma^2, Q_P\right) = \begin{cases} \frac{\kappa\sigma^2}{2\lambda(1-\lambda)}\left(\frac{\lambda}{1-\lambda}\right)^{-\frac{1}{\eta}}, & M = 1, \\ \frac{4\kappa\sigma^2 e^{\rho + \frac{\rho}{\eta}}(1+2\rho)}{\sqrt{\eta}}\left(\int_{-\infty}^{\sqrt{\rho/\eta}} e^{-x^2} dx\right)^2, & M = 0, \end{cases}$$

[1] When $\eta \to 1$ from above, $\rho \to +\infty$.

while the agent's actual expected utility is equal to

$$U^A\left(\eta,\kappa\sigma^2,Q_P\right)=\begin{cases}\frac{\kappa\sigma^2}{2\eta}\left(\frac{1+\frac{\lambda}{1-\lambda}}{1+\left(\frac{\lambda}{1-\lambda}\right)^{\frac{1}{\eta}}}\left(\frac{1}{\lambda}-2\right)-\log\left(\frac{\lambda}{1-\lambda}\right)\right), & M=1,\\ \frac{2\kappa\sigma^2}{\eta}\left(2e^{\rho}\sqrt{\rho}\int\limits_{-\infty}^{\sqrt{\rho/\eta}}e^{-x^2}dx-\rho\right), & M=0,\end{cases}$$

where $\lambda\in(0,0.5)$ solves (4.26) and $\rho>0$ solves (4.27).

When $M=1$, we have

$$\frac{\partial U^P\left(\eta,\kappa\sigma^2,Q_P\right)}{\partial\eta}=\frac{\kappa\sigma^2 f_1(\lambda)}{2(1-\lambda)\lambda\eta^2\left(1+\left(\frac{\lambda}{1-\lambda}\right)^{\frac{1}{\eta}}\right)}<0,$$

$$\frac{\partial\left(U^P\left(\eta,\kappa\sigma^2,Q_P\right)+U^A\left(\eta,\kappa\sigma^2,Q_P\right)\right)}{\partial\eta}=\frac{\kappa\sigma^2}{2\eta^2(1+\eta(1-2\lambda))}\times$$
$$\left[\frac{\left(\frac{1}{\lambda(1-\lambda)}-1+\left(\frac{\lambda}{1-\lambda}\right)^{\frac{1}{\eta}}\right)(1-2\lambda)\eta\log\left(\frac{\lambda}{1-\lambda}\right)}{1+\left(\frac{\lambda}{1-\lambda}\right)^{\frac{1}{\eta}}}+\right.$$
$$\left.\frac{f_2(\lambda)\left(\frac{\lambda}{1-\lambda}\right)^{\frac{1}{\eta}}}{\left(1+\left(\frac{\lambda}{1-\lambda}\right)^{\frac{1}{\eta}}\right)^2}+\left(1+\frac{\left(\frac{\lambda}{1-\lambda}\right)^{\frac{1}{\eta}}\frac{1-2\lambda}{\eta(1-\lambda)\lambda}}{\left(1+\left(\frac{\lambda}{1-\lambda}\right)^{\frac{1}{\eta}}\right)^2}\right)\frac{\left(\frac{\lambda}{1-\lambda}\right)^{\frac{1}{\eta}}}{1+\left(\frac{\lambda}{1-\lambda}\right)^{\frac{1}{\eta}}}f_1(\lambda)\right]<0,$$

because $f_1(\lambda)=1-2\lambda+(1-2\lambda(1-\lambda))\log\left(\frac{\lambda}{1-\lambda}\right)<0$ and $f_2(\lambda)=(1-2\lambda)\left(\frac{1}{\lambda(1-\lambda)}-2\right)\left(1+2\log\left(\frac{\lambda}{1-\lambda}\right)\right)+\frac{4\lambda^3}{1-\lambda}\log\left(\frac{\lambda}{1-\lambda}\right)<0$ for all $\lambda\in(0,0.5)$.

When $M=0$, we have

$$\frac{\partial U^P\left(\eta,\kappa\sigma^2,Q_P\right)}{\partial\eta}=-\frac{2\kappa\sigma^2}{\eta^2}e^{\rho}\sqrt{\rho}(2\rho-1)\int\limits_{-\infty}^{\sqrt{\rho/\eta}}e^{-x^2}dx$$

Since $\rho\left(\eta,\kappa\sigma^2,Q_P\right)$ is increasing in η (as follows from (4.27)), $U^P\left(\eta,\kappa\sigma^2,Q_P\right)$ has a unique maximum, and $\rho=0.5$ at this maximum. From (4.27), we have $\hat{\eta}_P\left(\frac{Q_P}{\kappa\sigma^2}\right)=\frac{1}{2y^2}$, where $y>0$ solves

$$4\sqrt{2e\pi}y\left(y+2e^{y^2}\int\limits_{-\infty}^{y}e^{-x^2}dx\right)=\frac{Q_P}{\kappa\sigma^2}. \tag{C.27}$$

As for the social welfare for $M = 0$, we have

$$\frac{\partial\left(U^P\left(\eta,\kappa\sigma^2,Q_P\right)+U^A\left(\eta,\kappa\sigma^2,Q_P\right)\right)}{\partial\eta}=$$
$$\frac{2\kappa\sigma^2 f_2(\eta,\rho)\left(\frac{1}{\sqrt{\eta}(1+2\rho)}+\frac{e^{\rho/\eta}}{\sqrt{\rho}}\left(f_1(\rho)+\int\limits_0^{\sqrt{\rho/\eta}}e^{-x^2}dx\right)\right)}{\eta^2 e^{\frac{\rho}{\eta}-\rho}\left(\frac{1}{\rho}+\left(\frac{1}{2}+\frac{1}{2\eta}+\frac{1}{1+2\rho}\right)Z\left(\frac{\rho}{\eta}\right)\right)}$$

where $Z(z)=\frac{2e^z}{\sqrt{z}}\int\limits_{-\infty}^{\sqrt{z}}e^{-x^2}dx$, $f_1(\rho)=\frac{\sqrt{\pi}}{2}-\frac{e^{-\rho}\sqrt{\rho}}{1+2\rho}>0$ for all $\rho>0$, and

$$f_2(\eta,\rho)=\frac{1}{2}-\rho-\left(\rho+\frac{\sqrt{2}+1}{2}\right)\left(\rho-\frac{\sqrt{2}-1}{2}\right)Z\left(\frac{\rho}{\eta}\right).$$

Then this derivative is positive if and only if $f_2(\eta,\rho)>0$. It is easy to show that if $f_2(\eta,\rho(\eta))=0$ for some η, then $\frac{d}{d\eta}f_2(\eta,\rho(\eta))<0$, where $\rho(\eta)$ is defined from (4.27). Recall that $\rho(\eta)$ is increasing in $\eta>0$ from 0 to $+\infty$, so that $f_2(0,\rho(0))>0$ and $f_2(+\infty,\rho(+\infty))<0$. Thus, there exists a unique $\eta>0$ that solves $f_2(\eta,\rho(\eta))=0$. Denote this η as $\hat{\eta}_{PA}\left(\frac{Q_P}{\kappa\sigma^2}\right)$.

To compare $\hat{\eta}_P\left(\frac{Q_P}{\kappa\sigma^2}\right)$ with $\hat{\eta}_{PA}\left(\frac{Q_P}{\kappa\sigma^2}\right)$, consider function $f_2(\eta,\rho(\eta))$ at point $\eta=\hat{\eta}_P\left(\frac{Q_P}{\kappa\sigma^2}\right)$ (recall that $\rho\left(\hat{\eta}_P\left(\frac{Q_P}{\kappa\sigma^2}\right)\right)=0.5$):

$$f_2\left(\hat{\eta}_P,\rho(\hat{\eta}_P)\right)=-\frac{e^{y^2}}{y}\int\limits_{-\infty}^{y}e^{-x^2}dx<0,$$

where $y>0$ solves (C.27). Thus, $\hat{\eta}_{PA}\left(\frac{Q_P}{\kappa\sigma^2}\right)<\hat{\eta}_P\left(\frac{Q_P}{\kappa\sigma^2}\right)$.

C.13 Dynamic Model, $\delta>0$

The decision maker faces the following optimization problem:[2]

$$\sup_{(\tau,v)}\mathbb{E}\left[e^{-\delta\tau}U(v,p_\tau)-\kappa\int\limits_0^{\tau}e^{-\delta t}dt\right]. \tag{C.28}$$

Before presenting the optimal strategy, we need to make one more assumption:

[2]A more general strategy space includes the opportunity to allocate partial attention to the information flow. If the agent allocates $\Delta\in[0,1]$ amount of attention at time t, then he pays $\kappa\Delta dt$ and observes $X_{t+\Delta dt}-X_t$. It turns out that it is never optimal to allocate partial attention when the discount factor is positive. Suppose $\Delta^*>0$ is optimal. Then $\mathcal{L}(p_t,\Delta^*)=\min\limits_{\Delta\in[0,1]}\mathcal{L}(p_t,\Delta)=0$, where $\mathcal{L}(p,\Delta)=\Delta\left(\kappa-\frac{2}{\sigma^2}p^2(1-p)^2V''(p)\right)+\delta V(p)$, $V(p)$ is the value function. Thus, whenever $\delta V(p)\neq 0$, we must have $\Delta^*=1$.

Assumption C.1. $-\kappa < (Q+R)\delta$.

This assumption says that if $R < -Q$, then δ should be small enough. It guarantees that stopping is optimal if the true state is known. Indeed, if $R < -Q$, then the decision maker gets negative utility when he makes the decision. Basically, the final utility payment acts as a cost. If the discount factor is large, he would want to postpone the payment of this cost forever. If Assumption C.1 does not hold, the optimal strategy does not exist.

Theorem C.2. *The optimal strategy exists and is given by (4.2), where threshold* $\lambda \in (0, 0.5)$ *is uniquely defined by*

$$\frac{2\lambda(1-\lambda)}{\left(1 - \frac{2}{1+\left(\frac{1-\lambda}{\lambda}\right)^{\sqrt{1+2\delta\sigma^2}}}\right)\sqrt{1+2\delta\sigma^2} - 1 + 2\lambda} - (1-\lambda) = \frac{R + \frac{\kappa}{\delta}}{Q}. \tag{C.29}$$

Proof. The Hamilton-Jacobi-Bellman equation (C.1), where

$$\mathcal{L}(p) = \kappa - \frac{2}{\sigma^2}p^2(1-p)^2V''(p) + \delta V(p),$$

gives the sufficient condition for a continuously differentiable function $V\colon [0,1] \to \mathbb{R}$ to be the value function

$$V(p) = \sup_{(\tau,v)} \mathbb{E}\left[e^{-\delta\tau}U(v, p_\tau) - \kappa\int_0^\tau e^{-\delta t}dt \mid p_0 = p\right]. \tag{C.30}$$

Differential equation $\mathcal{L}(p) = 0$ has the following solution:

$$V(p) = -\frac{\kappa}{\delta} + C_1 p^{\frac{1-\sqrt{1+2\delta\sigma^2}}{2}}(1-p)^{\frac{1+\sqrt{1+2\delta\sigma^2}}{2}} + C_2 p^{\frac{1+\sqrt{1+2\delta\sigma^2}}{2}}(1-p)^{\frac{1-\sqrt{1+2\delta\sigma^2}}{2}},$$

where C_1 and C_2 are some constants.

Consider the following class of functions defined on $p \in [0,1]$ parameterized with $\lambda \in (0, 0.5]$:

$$V_\lambda(p) = \begin{cases} pQ + R, & p \geq 1-\lambda, \\ (1-p)Q + R, & p \leq \lambda, \\ \dfrac{\left(\left(\frac{1-p}{p}\right)^{\frac{\sqrt{1+2\delta\sigma^2}}{2}} + \left(\frac{p}{1-p}\right)^{\frac{\sqrt{1+2\delta\sigma^2}}{2}}\right)\sqrt{\frac{(1-p)p}{(1-\lambda)\lambda}}\left((1-\lambda)Q + R + \frac{c}{\delta}\right)}{\left(\frac{1-\lambda}{\lambda}\right)^{\frac{\sqrt{1+2\delta\sigma^2}}{2}} + \left(\frac{\lambda}{1-\lambda}\right)^{\frac{\sqrt{1+2\delta\sigma^2}}{2}}} - \frac{c}{\delta}, & \text{otherwise.} \end{cases}$$

Note that these functions are continuous, symmetric around $p = 0.5$, that is $V_\lambda(p) = V_\lambda(1-p)$, and satisfy $\mathcal{L}(p,1) = 0$ for $\lambda < p < 1-\lambda$. Moreover, function $V_\lambda(p)$ is continuously differentiable if and only if

$$\lim_{p \to \lambda + 0} V'_\lambda(p) = -Q,$$

which is equivalent to (C.29). Note that the left hand side of (C.29) is an increasing function of $\lambda \in (0, 0.5]$ from -1 to $+\infty$. Thus,

- if $\frac{R+\frac{\kappa}{\delta}}{Q} > -1$, then the solution to (C.29) always exists and unique,
- if $\frac{R+\frac{\kappa}{\delta}}{Q} \leq -1$, then there is no $\lambda \in (0, 0.5]$ such that $V_\lambda(p)$ is continuously differentiable.

Note that (C.29) implies that $(1-\lambda)Q + R + \frac{\kappa}{\delta} > 0$.

Finally, note that

1. $V_\lambda(p) \geq \max\limits_{v \in \{G,I\}} U(v,p)$ for $\lambda < p < 1-\lambda$ since $V_\lambda(p)$ is convex for $\lambda < p < 1-\lambda$ as long as $(1-\lambda)Q + R + \frac{\kappa}{\delta} > 0$,
2. $\mathcal{L}(p) \geq 0$ for $p \geq 1-\lambda$ and $p \leq \lambda$ as long as $(1-\lambda)Q + R + \frac{\kappa}{\delta} \geq 0$.

Thus, $V_\lambda(p)$ is the value function and therefore the strategy (4.2) is the unique optimal one. ■

Note that when Assumption C.1 just holds, that is when $(Q+R)\delta + \kappa$ is very small, the optimal λ is very close to 0, which corresponds to long learning.

Theorem C.3. *$\mathcal{X}(\sigma^2)$ defined by (4.10) and (C.29) is increasing in σ^2.*

Proof. First, we calculate $\lambda'(\sigma^2)$ from (C.29) holding κ, Q, R, δ fixed:

$$\lambda'\left(\sigma^2\right) = \frac{(1-\lambda)\lambda\left(\left(\frac{1-\lambda}{\lambda}\right)^{2\sqrt{1+2\delta\sigma^2}} + 2\sqrt{1+2\delta\sigma^2}\left(\frac{1-\lambda}{\lambda}\right)^{\sqrt{1+2\delta\sigma^2}}\log\left(\frac{1-\lambda}{\lambda}\right) - 1\right)}{\left(1+\left(\frac{1-\lambda}{\lambda}\right)^{\sqrt{1+2\delta\sigma^2}}\right)^2 \sigma^2\sqrt{1+2\delta\sigma^2}}. \tag{C.31}$$

Substituting (C.31) into (C.4), we get $\mathcal{X}'(\sigma^2) = g\left(\lambda(\sigma^2), \sqrt{1+2\delta\sigma^2}\right)$, where for any $\lambda \in (0, 0.5)$ and $\alpha > 1$ function $g(\lambda, \alpha)$ is defined as

$$g(\lambda,\alpha) = \frac{1 + \alpha\left(1+\left(\frac{1-\lambda}{\lambda}\right)^{2\alpha}\right)\log\left(\frac{1-\lambda}{\lambda}\right) - \left(\frac{1-\lambda}{\lambda}\right)^{2\alpha}}{2\alpha\left(1+\left(\frac{1-\lambda}{\lambda}\right)^{\alpha}\right)^2}.$$

For any fixed $\alpha > 1$, function $g(\lambda, \alpha)$ is decreasing in $\lambda \in (0, 0.5)$ from $+\infty$ to 0. Thus, it is always positive and therefore $\mathcal{X}'(\sigma^2) > 0$. ■

C.14 Dynamic Asymmetric Model

Consider a general form of the utility function:

$$U(v, p) = pu(v, I) + (-p)u(v, G).$$

Assumption C.2.

$$u(A, I) > \max\{u(A, G), u(C, I)\}, \qquad u(C, G) > \max\{u(C, I), u(A, G)\}.$$

Denote

$$p^* = \frac{u(C, G) - u(A, G)}{u(C, G) - u(A, G) + u(A, I) - u(C, I)}.$$

Then, the judge wants to acquit when $p > p^*$ and he wants to convict when $p < p^*$.

Without loss of generality, assume that the judge is weakly biased towards convicting

Assumption C.3. $p^* \geq 0.5$.

Theorem C.4. *The optimal strategy exists and is given by*

$$\tau = \inf\{t \geq 0 \colon p_t \notin (\lambda, \mu)\}, \quad v = \begin{cases} A, & p_\tau \geq \mu, \\ C, & p_\tau \leq \lambda, \end{cases} \tag{C.32}$$

where p_t is the belief that the true state is I at time t. Thresholds $0 < \lambda < p^* < \mu < 1$ are uniquely defined by

$$f(\mu) - f(\lambda) = \frac{G}{2}, \quad \mu = \frac{1}{2}\left(1 + \sqrt{1 - \frac{4(1-\lambda)\lambda}{1 + (2p^* - 1)G(1-\lambda)\lambda}}\right), \tag{C.33}$$

where $G = \frac{2(u(A,I) - u(A,G) + u(C,G) - u(C,I))}{\kappa\sigma^2}$, $f(x) = \log\left(\frac{x}{1-x}\right) - \frac{1-2x}{2(1-x)x}$.

The proof is similar to Theorem 4.1.

In asymmetric case, the choice of the welfare function is not so obvious. In symmetric case, it is natural to take the probability of the correct decision as the welfare criterion. When there is bias in prior belief and / or in preferences $u(v, z)$, there are many different options one can take as the welfare criterion. We are not going to consider them all and just focus on how the strategy changes with the overconfidence level.

Theorem C.5 confirms the conclusion of Theorem 4.2 for the upper threshold:

Theorem C.5. $\frac{\sigma^2}{2}\log\left(\frac{\mu(\sigma^2)}{1-\mu(\sigma^2)}\right)$ *is increasing in* σ^2.

However, this conclusion is no longer true for the lower threshold:

Theorem C.6. *When* $p^* > \frac{1}{2}$*, there exists* $\Sigma^2 > 0$ *such that* $\frac{\sigma^2}{2}\log\left(\frac{\lambda(\sigma^2)}{1-\lambda(\sigma^2)}\right)$ *is decreasing for* $\sigma^2 < \Sigma^2$ *and it is increasing for* $\sigma^2 > \Sigma^2$.

Theorem C.6 states that there is a unique level of overconfidence $\eta = \frac{\sigma^2}{\Sigma^2}$ that minimizes the lower threshold for X_t. Intuitively, there is a trade-off between the preference bias and the overall precision of the decision. When there is a lot of noise in information, the bias is more prominent since the decision is not precise anyway. As information becomes more precise, the trade-off optimal resolution moves towards the decision precision, which means that the thresholds become more symmetric.