Essays on Incentives in Matching Markets

Inauguraldissertation
zur
Erlangung des Doktorgrades
der
Wirtschafts- und Sozialwissenschaftlichen Fakultät
der
Universität zu Köln

2021

vorgelegt
von
Yiqiu Chen, M.Sc.
aus
Mianyang

Referent: Prof. Dr. Alexander Westkamp
Korreferent: Prof. Dr. Christoph Schottmüller

Tag der Promotion: 18. March 2022

This thesis consists of the following works:

Chen, Yiqiu and Möller, Markus (2021):
Regret-Free Truth-Telling in School Choice with Consent,
Working Paper.

Chen, Yiqiu and Westkamp, Alexander (2021):
Optimal Sequential Implementation,
Working Paper.

Chen, Yiqiu (2021):
Partition Obviously Strategy-Proof Rules,
Working Paper.

Acknowledgments

First of all, I would like to express my sincere gratitude to my supervisor Alexander Westkamp for his generous guidance, support and patience during my master and doctoral studies, for the freedom and encouragement he gave such that I can develop my own research ideas and for the valuable discussions on our project that help me to learn how to develop proper research projects. Also, I would like to thank Christoph Schottmüller for co-referring my thesis and providing valuable academic advice.

My sincere thanks also go to Markus Möller. He has been a perfect colleague who made work much fun and who is always open for long and in-depth discussions in both my projects and our joint project. In addition, he has also been a fantastic friend to spend time with. I am also grateful to Marius Gramb who offered me help in many aspects.

For finical support I thank the Cologne Graduate School, which has provided me with scholarships for three years; and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), which has funded me for the last year under Germany's Excellence Strategy – EXC 2126/1–390838866.

I am deeply grateful to my parents for their unconditional love and support in all kinds. Finally but most importantly, I genuinely express my thanks to my wife Sha Li, who is always having me back and gives me precious company and love throughout the past twelve years. I dedicate this thesis to my parents and my wife.

Contents

Economics is about the efficient allocation of scarce resources, and about making resources less scarce.

Alvin E. Roth

1 Introduction

This thesis consists of three theoretical essays that contribute to the research of matching and market design. In particular, these essays focus on incentive analysis in different centralized matching environments, and they aim to provide a better understanding of how people's incentives of behaving truthfully can be influenced by factors including observable information, revelation principles and their abilities to perform contingent reasoning.

Starting with the celebrated work by Gale and Shapley (1962), the literature on matching theory and its applications has developed greatly over the last decades. The way how resources are allocated in practice has been optimized based on theoretical and empirical findings in this field. Examples include

centralized allocation of students to colleges/public schools (Balinski and Sönmez, 1999; Abdulkadiroğlu and Sönmez, 2003), kidneys to patients (Roth et al., 2004), doctors to hospitals (Roth and Peranson, 1999), vaccinations to residents (Pathak et al., 2021) and so on. Participants' preferences are critical to achieve allocations with desirable properties in these markets. However, true preferences are usually private information that is not known by central authorities which are responsible for the allocation procedures. Therefore, an important concern in designing matching rules is to provide incentives to participants such that they reveal their preferences truthfully. This thesis deals with various incentive properties in matching markets and their applications.

In matching markets, the classic approach to incentivize truthful behaviors is through *strategy-proof* matching rules, under which stating their true preferences is a dominant strategy for participants. Nevertheless, researchers have designed a broad range of non-strategy-proof rules in recent years since they aim at achieving promising properties which are at odds with strategy-proofness (Kesten, 2010; Dur et al., 2019; Alva and Manjunath, 2019). As a weakening of strategy-proofness, the first essay (Chapter 2) conducts a regret-based incentive analysis on non-strategy-proof rules. Specifically, while observable information is irrelevant in identifying the dominant strategy under strategy-proof rules, it might be helpful for participants to decide on certain strategies under non-strategy-proof rules where no strategy is dominant. Thus, under certain non-strategy-proof rules, we explore participants' incentives of truth-telling by allowing them to anticipate that they would receive market information which is commonly accessible in public school choice procedures.

On the contrary, economists have also started to seek for solutions with incentive guarantees stronger than those induced by strategy-proof rules. These endeavours are motivated by empirical and experimental findings which suggest that dominated behaviors from participants are routinely observed under strategy-proof rules (see e.g., Chen and Sönmez (2006), Hassidim et al. (2016) and Shorrer and Sóvágó (2018)). A prevailing answer that reacts to these counter-intuitive results is *obviously strategy-proof (OSP)* mechanisms due to Li (2017). In matching markets, OSP mechanisms are essentially se-

quential implementations of strategy-proof rules and they make the incentive given by strategy-proof rules more apparent. Concretely, OSP mechanisms ensure that truth-telling is an *obviously dominant strategy* everywhere, meaning that the worst-case outcome under the truthful report is weakly better than the best-case outcome under any misreport. Li (2017) shows that even participants who are unable to engage in any contingent reasoning would follow the truthful strategy in OSP mechanisms. However, recent findings suggest that popular strategy-proof matching rules, such as Top Trading Cycles (TTC) and Deferred Acceptance (DA), cannot be implemented via OSP mechanisms in general (Li, 2017; Ashlagi and Gonczarowski, 2018). The remaining two essays of this thesis are motivated by these negative results. Specifically, the second essay (Chapter 3) uses obvious dominance as a guideline to design sequential implementations of strategy-proof rules that exist even in the absence of OSP mechanisms. The third essay (Chapter 4) systematically studies the amount of contingent reasoning necessary for a participant to figure out that a rule is strategy-proof.

In particular, Chapter 2 is based on Chen and Möller (2021). It is co-authored by Markus Möller and both authors contributed equally to this project. We consider the many-to-one school choice model with consent (Kesten, 2010) under incomplete information and we interpret the priorities of schools in the form of scores. The incomplete information structure is inspired by common features in public school applications. To be more concrete, students can observe the final allocation and the cutoff at each school that denotes the lowest score with which a student is admitted to that school. Based on the observed information, students can draw inferences about scores of schools and reported preferences of other students. We adopt the regret-based incentive notion by Fernandez (2020) to our setting. Specifically, a strategy is *regret-free* for a student if at any observation, she will not find her strategy to be weakly dominated by another strategy at all inferences of reported preferences and scores, which are consistent with that observation.

The first non-strategy-proof rule we study in this chapter is the *Efficiency Adjusted Deferred Acceptance Rule (EDA)* due to Kesten (2010). A *fair* allocation ensures that for each student, each school

that she prefers to her assignment is assigned to students with higher scores than she has. As it is impossible to achieve a mechanism which always yields allocations that are both efficient and fair (Balinski and Sönmez, 1999), EDA elegantly circumvents this impossibility by asking for students' consents to relax the fairness criterion. The main results of this chapter confirm that reporting their true preferences is regret-free for students under EDA and that no untruthful strategy provides the same guarantee under EDA. We also study the family of *efficient stable dominating rules* (Alva and Manjunath, 2019), which always produce allocations that are efficient and weakly Pareto dominate a fair allocation. We show that no efficient stable dominating rule satisfies our regret-based incentive criterion. Our findings address the unique role of truth-telling under EDA in terms of incentives and provide useful insights for organizations seeking to practically implement EDA.

Chapter 3 is based on Chen and Westkamp (2021). It is joint work with Alexander Westkamp and both authors contributed equally to this work. While Chapter 2 studies an incentive property that is weaker than strategy-proofness, Chapter 3 aims at achieving stronger incentives for truth-telling than those given by strategy-proof rules. We concentrate on a standard priority-based allocation problem without transfers. The main contribution of this chapter is our proposal of *optimal* sequential implementations of matching rules, for which we set two requirements. First, whenever it is obviously dominant for an agent to truthfully reveal certain information about her preferences, an optimal sequential implementation will prioritize picking such decisions over decisions in which truthful revelations are not obviously dominant. Second, an optimal sequential implementation elicits no more than the minimal amount of information necessary to unambiguously determine the outcome under that rule. We find that whenever a strategy-proof rule can be implemented via OSP mechanisms, an optimal sequential implementation of that rule is also an OSP mechanism. Thus, optimal implementations are complementary to the characterization of OSP mechanisms without transfers by Pycia and Troyan (2021).

We develop an optimal sequential implementation of TTC as one main finding of this chapter.

Notably, the existence of our proposed implementation is guaranteed in all markets under consideration. We also introduce a weaker notion of optimality: It loosens the second requirement by imposing no restriction on the amount of information elicited through decisions in which truthful revelations are obviously dominant. We further introduce a weakly optimal sequential implementation for DA that exists generally. Our proposals for TTC and DA are promising solutions in providing incentives for truth-telling for markets where implementations of TTC and DA via OSP mechanisms are unavailable. In this sense, this study complements the line of study initiated by Li (2017), shedding light on incentive design in matching markets. Moreover, this chapter contributes to the design of practical applications of DA and TTC with the goal of maximizing truthful behaviors.

Chapter 4 is based on Chen (2021) and is single authored. It considers the same model as in Chapter 3. However, while Chapter 3 contributes to Li (2017)'s concept by designing sequential games that are closely connected to OSP mechanisms, this chapter relates to Li (2017) by studying the degree of contingent reasoning necessary to understand the strategy-proofness of a rule. I consider participants who are deficient in contingent reasoning and adopt Zhang and Levin (2017)'s method to measure such deficiencies. Concretely, I assume that each participant can partition all states of the world, namely all potential preferences of others, into several events. Accordingly, a rule is called *partition obviously strategy-proof* for that participant if within each event of the just described partition, the worst-case outcome under the truthful report is at least as good as the best-case outcome under any misreport. In this sense, a partition is used to interpret a participant's limited reasoning ability: Given each of her own preferences, she can figure out the set of possible outcomes in each event, but she does not know which outcome results from which preferences of others in that event.

The main result of this chapter states that a participant can understand the incentive given by a strategy-proof rule if and only if given each of her own preferences and in each event of the partition specified by her reasoning ability, her assignment is uniquely determined. In other words, a participant will stick to the truthful strategy under a strategy-proof rule if and only if her reasoning ability

uncovers all uncertainties about her own assignment. Moreover, I find that in sequential implementations of strategy-proof rules, the amount of reasoning ability required to stick to truth-telling is reduced. This finding provides a new angle to understand laboratory results (Klijn et al., 2019; Bó and Hakimov, 2020a; Breitmoser and Schweighofer-Kodritsch, 2021) which observe higher rates of truth-telling in dynamic forms of strategy-proof rules compared to static counterparts.

Overall, the results from the three essays presented in this thesis gain new insights into when and how participants will be incentivized to behave truthfully under various matching rules and mechanisms. First, although not being strategy-proof, EDA still provides participants with reasonably strong incentives to report truthfully considering the information participants usually obtain in practice. Second, a promising complementary solution to OSP mechanisms could be the mechanisms that comply with obvious dominance whenever possible and minimize the amount of information revealed from participants. Third, to understand that a rule is strategy-proof, people must reason to the degree such that there remains no uncertainty about their own assignments. These works add to the literature on incentive studies in matching theory as well as to the literature on their applications.

2

Regret-Free Truth-Telling in School Choice with Consent*

The *Efficiency Adjusted Deferred Acceptance Matching Rule (EDA)* is a promising candidate mechanism for public school assignment. A potential drawback of EDA is that it could encourage students to game the system since it is not strategy-proof. However, to successfully strategize, students typically need information that is unlikely to be available to them in practice. We model school choice under

*This chapter is based on Chen and Möller (2021). We thank especially our advisor, Alexander Westkamp. We are grateful to Christoph Schottmüller, Marcelo Ariel Fernandez, Kevin Breuer and Marius Gramb for helpful comments. All errors remain our own.

incomplete information and show that EDA is regret-free truth-telling, which is a weaker incentive property than strategy-proofness and was introduced by Fernandez (2020). We also show that there is no efficient matching rule that Pareto dominates a stable matching rule and is regret-free truth-telling.

2.1 Introduction

Efficiency and fairness are incompatible in the school choice problem.[1] The *Efficiency Adjusted Deferred Acceptance Rule (EDA)* (Kesten, 2010) elegantly circumvents this incompatibility by allowing students to give their consent to relax the fairness constraint. However, no compromise solution, including EDA, is strategy-proof (Abdulkadiroğlu et al., 2009).[2,3] We study whether EDA satisfies an incentive criterion by Fernandez (2020) which is weaker than strategy-proofness and is based on participants' wish to avoid regret.

We employ the many-to-one school choice model with consent (Kesten, 2010) under incomplete information. Students can reconsider their admission chances for alternative reports, through an observational structure that is based on the cutoff terminology. We express schools' priorities in the form of scores and for each school, the cutoff is the lowest score among all students that have been admitted to that school. Once the final matching has been determined, each student observes which student is assigned to which school and each school's cutoff. Based on her observation, a student can then draw inferences about *plausible scenarios*—pairs of underlying scores of schools and reports of other students that are consistent with the observation. We motivate our model through features common in the context of public school assignment. In practice, matching rules often use scores based on prox-

[1] A student has justified envy at a matching, if there exists a lower prioritized student assigned to a school and the corresponding school is preferred to her assignment (Abdulkadiroğlu and Sönmez, 2003). A matching is *fair* if no justified envy exists and a matching rule is fair if it only produces matchings which are fair. The trade-off between efficiency and fairness follows from Balinski and Sönmez (1999).

[2] Strategy-proofness requires that it is a weakly dominant strategy for students to report their true preferences.

[3] For related results, see also Erdil and Ergin (2008) and Alva and Manjunath (2019).

imity, walk-zone areas, sibling-status and other socioeconomic variables. The composition of scores is usually public information, whereas accurate information on other students' scores and reported preferences will generally be covered by privacy protection. Moreover, students typically receive feedback on the market outcome and cutoffs.

In this model, we adopt the incentive notion by Fernandez (2020). Specifically, a student *regrets* a report through an alternative report, once she finds her submitted report to be dominated by the alternative in any plausible scenario. A rule is *regret-free truth-telling* if no student would regret reporting her preferences truthfully.

The main finding of this chapter is that EDA is *regret-free truth-telling* (Theorem 2.1). Moreover, we show that under EDA, truth-telling is the *unique* option which never leads to regret (Proposition 2.2). Concretely, we show that for any misreport, there exists an observation such that the student regrets the misreport through her true preferences. Our last result concerns matching rules which Pareto dominate a stable matching rule.[4] A stable dominating rule always implements a matching that weakly Pareto dominates a stable matching (Alva and Manjunath, 2019). It is well known that all stable dominating rules, except the well known *Deferred Acceptance Matching Rule (DA)* (Gale and Shapley, 1962), are not strategy-proof (Abdulkadiroğlu et al., 2009).[5] We show that among the *efficient stable dominating rules* no matching rule is regret-free truth-telling (Theorem 2.2). Note that the original formulation of EDA considered in this chapter is not Pareto efficient since EDA respects improvements on efficiency only with students' consents for being exposed to justified envy.

All our results extend to the case where the students only observe their own assignment and the cutoffs. By showing that truth-telling is the unique regret-free strategy, we provide an appropriate statement for the intuition that truth-telling may be a focal strategy under EDA. Thus, our work

[4] A matching rule is *stable* if it produces outcomes which are *fair*, *individually rational* and *non-wasteful*. A matching is non-wasteful if there is no object that is unassigned although there is an agent that prefers it over her assignment. A matching is individually rational if no agent prefers her outside option over her final assignment.

[5] See also Erdil and Ergin (2008), Kesten (2010) and Alva and Manjunath (2019).

contributes to the strand of literature that outlines the many desirable features of EDA for practical implementation.

Related literature To our knowledge, Fernandez (2020) is the first to introduce regret-based incentives in the matching literature.[6] In marriage markets, Fernandez (2020) shows that truth-telling is the unique regret-free strategy under DA for both men and women and that DA is the unique regret-free truth-telling rule among so-called quantile stable rules.[7] Fernandez (2020) sheds light on college admissions problems. He shows that the student-proposing variant of DA is regret-free truth-telling. However, under the college-proposing variant of DA, being truthful does not need to be free of regret for colleges. The key differences of our work to that of Fernandez (2020) is that only the student market side is strategic. Moreover, whereas in Fernandez (2020) participants only observe the realized matching, students in our model additionally observe cutoffs.

This chapter mainly contributes to the literature that deepens the understanding of EDA's incentive properties. Our results complement those of Troyan and Morrill (2020), who show that for cognitively limited participants beneficial misreporting under EDA is not *obvious* in the following sense: a profitable misreport is an *obvious manipulation* if the best-case outcome of the misreport is better than the best-case outcome of telling the truth or, if the worst-case outcome of the misreport is better than the worst-case outcome of telling the truth. The main difference between our work and that of Troyan and Morrill (2020) concerns the source of uncertainty that students face. A profitable misreport is obvious if it is easy to recognize for students whose knowledge on the matching rule is imperfect, given that these students have full access to the scores of other students. That is, non-obvious

[6] Regret-based incentives have a long tradition in economic theory. For instance, in auction theory, regret-based incentives of bidders in first-price auctions have been studied by Filiz-Ozbay and Ozbay (2007) and Engelbrecht-Wiggans (1989). For a more detailed discussion we refer to Fernandez (2020). See Gilovich and Medvec (1995) and Zeelenberg and Pieters (2007) for psychological treatments of regret.

[7] Given any $q \in (0, 1]$, the q-quantile stable rule selects the $\lceil qk \rceil$ best stable school for each student given any report, where k is the number of stable matchings under this report. For more information on quantile stable mechanisms, we refer to Teo and Sethuraman (1998), Klaus and Klijn (2006), or Chen et al. (2014).

manipulability is mainly driven by participants' limited understanding of the matching rule. By contrast, students in our model know how the matching rule works and our results are driven by students' incomplete access to the scores of other students. Notably, the positive result of Troyan and Morrill (2020) covers both EDA and stable dominating rules, where we reach a negative result for efficient stable dominating rules.

Previous results on EDA's incentive properties are inspired by the theoretical benchmark for low information environments from Roth and Rothblum (1999) and Ehlers (2008). Kesten (2010) studies Bayesian incentives of EDA in a setting where it is common knowledge that students' preferences over schools are ordered into shared quality classes and students' beliefs on how other students order schools within each quality class are symmetrically distributed. Kesten (2010) shows that if other students submit their true preferences, then truth-telling stochastically dominates any other strategy. The key difference to our model is that we do not specify any prior probability distribution regarding the beliefs or distribution on other participants' preferences and thus do not impose any symmetry assumptions or correlation of preferences over schools. Thus, in contrast to the approach of Kesten (2010) our information environment follows the 'Wilson doctrine' (Wilson, 1987).

The literature that is concerned with other theoretical properties of EDA is rapidly growing. Tang and Yu (2014), Ehlers and Morrill (2019), Bando (2014) and Dur et al. (2019) recently developed tractable alternatives to Kesten's initial formulation of EDA. Ehlers and Morrill (2019) generalize EDA to a school choice model where school priorities take the form of more flexible choice functions and Kwon and Shorrer (2019) propose a version of EDA for organ exchange.

Our work also relates to the line of literature that uses the cutoff terminology in school choice models. Most prominent in this regard is Azevedo and Leshno (2016) who characterize stable matchings in terms of cutoffs in a continuum school choice model. They show that cutoffs take the form of market-clearing prices that equalize supply and demand and can be used to perform comparative statics with respect to schools' incentives to invest in quality. When used to characterize stable matchings, cutoffs

usually take the form of a guarantee for participants to be admitted at schools. In our framework, final assignments may not correspond to stable matchings. Therefore, the cutoffs do not necessarily provide a student with information about whether she will be admitted at a desired school. Moreover, in our model the cutoffs are incorporated into students' strategic reasoning.

The rest of this chapter is organized as follows. We introduce the basic model and EDA in Section 2.2. We model the informational environment and adopt regret-free truth-telling in Section 2.3. In Section 2.4, we present our main results. Our analysis regarding efficient stable dominating rules is provided in Section 2.5. Finally, Section 2.6 gives a short conclusion. The Appendix contains most of our proofs.

2.2 Model

There is a finite set of students I and a finite set of schools S. Each school $s \in S$ has a fixed capacity q_s and we collect the capacities in $q = (q_s)_{s \in S}$. We add a common outside option $s_\emptyset$ for students which has infinite capacity.

Each school $s \in S$ has a set of scores $g^s = \{g_i^s\}_{i \in I}$, where $g_i^s \in (0,1)$ is i's score at s. We assume that $g_i^s \neq g_j^s$ for any $i, j \in I$ and any $s \in S$, and we say that for each pair of students $i, j \in I$, i has higher priority at s than j if and only if $g_i^s > g_j^s$. That is, for each school s, the school's scores induce a strict priority ranking over I.[8] For each $i \in I$, let $g_i = \{g_i^s\}_{s \in S}$ be the set of scores assigned to student i. Let a score structure $g = (g_i)_{i \in I}$ be a collection of scores for each student and let $g_{-i} = (g_j)_{j \in I \setminus \{i\}}$ be a collection of scores for students in $I \setminus \{i\}$. Moreover, set $\mathcal{G}_I$ as the domain of all possible score structures and $\mathcal{G}_{-i}$ as the domain of all score structures for students other than i.

For each student $i \in I$, let $\succ_i$ be a strict preference relation over $S \cup \{s_\emptyset\}$. The corresponding weak

[8] The incomplete information framework we introduce in Section 2.3 allows students to draw inferences about their admission chances. Our formulation of scores will then ensure that a student typically cannot infer her exact rank on a school's priority list just on the basis of her own score.

preference relation of $\succ_i$ is denoted by $\succeq_i$.[9] Let $\mathcal{P}$ denote the set of all possible strict preference relations over $S \cup \{s_\emptyset\}$. For any $\succ_i \in \mathcal{P}$, a school s is acceptable to i if $s \succ_i s_\emptyset$ and unacceptable if it is not acceptable. A preference profile $\succ = (\succ_i)_{i \in I}$ is a realization of $\mathcal{P}$ for each $i \in I$ and $\succ_{-i} = (\succ_j)_{j \in I \setminus \{i\}}$ is a preference profile for students in $I \setminus \{i\}$. We define $\mathcal{P}_I$ as the domain of all preference profiles and $\mathcal{P}_{-i}$ as the domain of all preference profiles for students in $I \setminus \{i\}$.

A *matching* $\mu : I \to S \cup \{s_\emptyset\}$ is a function such that for each $s \in S$, $|\mu^{-1}(s)| \leq q_s$. Given any μ, we set $\mu_i = \mu(i)$ as the assignment of i and $\mu_s = \mu^{-1}(s)$ as the set of students assigned to s. Denote the set of all possible matchings by $\mathcal{M}$.

In the following, fix any $\succ \in \mathcal{P}_I$. We say a matching μ *weakly Pareto dominates* another matching μ' if for all $i \in I$, $\mu_i \succeq_i \mu'_i$. A matching μ *Pareto dominates* μ' if μ weakly Pareto dominates μ' and for some $j \in I$, $\mu_j \succ_j \mu'_j$. A matching μ is *Pareto efficient* if there does not exist another matching μ' which Pareto dominates μ.

We now introduce two fairness notions, where we start with the well-known notion by Abdulkadiroğlu and Sönmez (2003). Given a matching μ, student i has *justified envy* towards student j at school μ_j under μ if $\mu_j \succ_i \mu_i$ and $g_i^{\mu_j} > g_j^{\mu_j}$. A matching μ is *fair* if no student has justified envy at μ. A matching μ is *individually rational* if for each student the assigned school is acceptable to her. A matching μ is *non-wasteful* if there does not exist a student i and a school s, such that $s \succ_i \mu_i$ and $|\mu_s| < q_s$. A matching μ is *stable* if it is fair, individually rational and non-wasteful.

We also consider a weaker fairness notion that was introduced by Kesten (2010). The notion takes students' willingness to consent for being exposed to justified envy into account. For each student i, the consent is parameterized by a binary variable $c_i \in \{0, 1\}$ where $c_i = 1$ means that i consents to any envy that is justified and otherwise to none. We say a matching μ *violates the priority* of student i given c_i if $c_i = 0$ and if there exists another student $j \in I$ such that i has justified envy towards j at μ. Let $c = (c_i)_{i \in I}$ be a consent profile and let $\mathcal{C}_I$ be the domain of all consent profiles. Denote a

[9] That is, for all $s, s' \in S$, $s \succeq_i s'$ if either $s \succ_i s'$ or $s = s'$.

consent profile of students other than i by $c_{-i} = (c_j)_{j \in I \setminus \{i\}}$ and the respective domain by $\mathcal{C}_{-i}$. Given a matching μ, a profile of preferences $\succ$ and a consent profile c, we say that a matching is *fair with consent* if there exists no student whose priority is violated at μ.

We call a collection $(I, S, q, g, \succ, c)$ a *school choice problem with consent* (or simply a *problem*). Throughout the main body of the chapter, we fix a problem $(I, S, q, g, \succ, c)$. A *report* of student i is pair $(\succ_i', c_i') \in \mathcal{P} \times \{0,1\}$, and a report profile is described by $(\succ', c') \in \mathcal{P}_I \times \mathcal{C}_I$. Analogously, let $(\succ_{-i}', c_{-i}') \in \mathcal{P}_{-i} \times \mathcal{C}_{-i}$ be a report profile of students except i.

A *matching rule* $f : \mathcal{G}_I \times \mathcal{P}_I \times \mathcal{C}_I \to \mathcal{M}$ maps any triple of a score structure, preference profile and consent profile into a matching. Given a report profile $(\succ, c)$ and a score structure g, let the outcome of f be $f(g, \succ, c)$ and for each $i \in I$ let $f_i(g, \succ, c)$ denote student i's respective assignment. If the matching rule does not take consent decisions into consideration, we write $f(g, \succ)$ instead of $f(g, \succ, c)$. A matching rule f is *Pareto efficient* if each outcome of the matching rule is Pareto efficient. Similarly, a matching rule is *stable* if it produces a stable matching for any problem.

We proceed with the description of two incentive notions for students. A matching rule f is *consent-invariant* if $f_i(g, \succ, (c_i, c_{-i})) = f_i(g, \succ, (c_i', c_{-i}))$ for all i and all c_i, c_i'. That is, each student's assignment is independent of her *own* consent decision. Note that the matching rules studied in this chapter are all consent-invariant. A matching rule f is *strategy-proof* if $f_i(g, (\succ_i, \succ_{-i}), c) \succeq_i f_i(g, (\tilde{\succ}_i, \succ_{-i}), c)$ for all i and all $\tilde{\succ}_i \in \mathcal{P}$. That means, for each student, reporting her true preferences is weakly better than reporting untruthfully regardless of other students' reports.

2.2.1 EDA

In this subsection, we present Kesten's *Efficiency Adjusted Deferred Acceptance Rule (EDA)* along with our first result. We use the *Top-Priority (TP) algorithm* (Dur et al., 2019) to calculate the outcomes of EDA and start with some basic terminologies needed for its introduction. For the rest of this section, fix any $(\succ, c)$. For any matching $\mu \in \mathcal{M}$, any student i and any school s, we say that *i demands s at μ*

if $s \succ_i \mu_i$. Moreover, we say that student i is *eligible* for s at μ if i demands s at μ and there exists no j who also demands s with $c_j = 0$ and $g_i^s < g_j^s$. In other words, the set of students eligible for s are those students who, once assigned to s, would not violate the priority of any other student at matching μ. Note that there could be more than one student who is eligible for a school and if two students i, i' are both eligible for s, then $g_i^s > g_{i'}^s$ implies $c_i = 1$.

Given a matching $\mu \in \mathcal{M}$, consider the directed graph $G(\mu) = (I, E(\mu))$, where $E(\mu) \subseteq I \times I$ is the set of (directed) edges such that $ij \in E(\mu)$ if and only if i is eligible for μ_j. A set of edges $\{i_1i_2, i_2i_3, ..., i_ni_{n+1}\}$ in $G(\mu)$ is a path if $i_1, i_2, ..., i_{n+1}$ are distinct and it is a cycle if $i_1, i_2, ..., i_n$ are distinct while $i_1 = i_{n+1}$.

A school s has *no demand* at μ if no student demands s at μ. A school s is *underdemanded* at μ if either it has no demand at μ or, there is no path in $G(\mu)$ that ends with some $i \in \mu_s$ which contains students who are part of a cycle in $G(\mu)$. We say that a student is *permanently matched* at μ if she is assigned to an underdemanded school at μ. Furthermore, a student is *temporarily matched* if she is not permanently matched.

Given $\mu \in \mathcal{M}$, we call $G^*(\mu) = (I, E^*(\mu))$ the *Top-priority graph* of μ and its set of edges $E^*(\mu)$ is defined as follows: we have $ij \in E^*(\mu)$ if and only if among the students who are temporarily matched at μ and are eligible for μ_j, student i has the highest score for μ_j. That is, for each $i \in I, E^*(\mu) \subseteq E(\mu)$ contains at most one edge pointing to i. Solving cycle $\gamma = \{i_1i_2, i_2i_3, ...i_ni_1\}$ in $G^*(\mu)$ is defined by the operation $\circ$ and yields matching $\nu = \gamma \circ \mu$, such that $\nu_i = \mu_j$ for each $ij \in \gamma$, and $\nu_{i'} = \mu_{i'}$ for each $i' \notin \{i_1, i_2, ..., i_n\}$.

The TP algorithm iteratively solves cycles based on the top-priority graphs, where one starts with the graph of the *Student Optimal Stable Matching (SOSM)*. The SOSM Pareto dominates all other stable matchings and can be calculated via the popular *Student-Proposing Deferred Acceptance Algorithm (DA)* (Gale and Shapley, 1962) which is presented in Appendix 2.A. The TP algorithm works as follows:

Step 0**:** Calculate the SOSM and denote the matching by μ^0.

Step t**,** $t \geq 1$**:** Given matching μ^{t-1}:

t.1 If there is no cycle in $G^*(\mu^{t-1})$, then stop and let the final outcome be μ^{t-1}.

t.2 Otherwise, select one of the cycles in $G^*(\mu^{t-1})$, say γ^t, and let $\mu^t = \gamma^t \circ \mu^{t-1}$. Move to step $t+1$.

As has been shown in Lemma 6 of Dur et al. (2019), any cycle selection of the algorithm leads to the outcome of EDA and thus the TP algorithm induces EDA.

We now move to our discussion on EDA's incentive properties which is known to be consent-invariant but not strategy-proof (Kesten, 2010). Our first result, Proposition 2.1, states that a certain class of deviations of a student does not affect her own assignment. For any preference relation $\succ_i \in \mathcal{P}$ and school $s \in S$, let the weak lower contour set of $\succ_i$ with respect to s be $L_s^{\succ_i} = \{s' \in S \mid s \succeq_i s'\}$.

Proposition 2.1. *If* $EDA(g, \succ, c) = \mu$ *and* $\tilde{\succ}_i \in \mathcal{P}$ *is such that for all* $s, s' \in L_{\mu_i}^{\succ_i}$, $s \succ_i s'$ *only if* $s \tilde{\succ}_i s'$, *then* $EDA_i(g, (\tilde{\succ}_i, \succ_{-i}), c) = \mu_i$.

Proof. See Appendix 2.B. □

In words, Proposition 2.1 shows that if a student's deviation from her baseline report keeps the same order of the schools in the lower contour set with respect to the baseline assignment, then it yields the same outcome for the deviating student. Note that the set of deviations we consider in Proposition 2.1 is a subset of the monotonic transformations at the student's baseline assignment. Formally, $\succ_i'$ is a *monotonic transformation* of $\succ_i$ at $s \in S \cup \{s_\emptyset\}$ if $s' \succ_i' s$ implies that $s' \succ_i s$. Our main result presented in Theorem 2.1 can be used to illustrate that Proposition 2.1 does not hold for all monotonic transformations at μ_i.

2.3 Regret in school choice

In this section, we introduce the informational environment and regret-based incentives. We first describe the students' information and impose an observational structure. Assume that before submitting the report, each student i knows (I, S, q, g_i) and the matching rule f. After assignments have been determined by f, each student observes the matching and the cutoff at each school, i.e., the lowest score among all applicants matched to the school. More formally, given a report profile $(\hat{\succ}, \hat{c})$, student i observes $\mu = f(g, \hat{\succ}, \hat{c})$ and for each school $s \in S \cup \{s_\emptyset\}$, she observes $\pi_s(\mu, g) = \min_{j \in \mu_s} g_j^s$ when $|\mu_s| = q_s$ and $\pi_s(\mu, g) = 0$ otherwise. Let $\pi(\mu, g) = \{\pi_s(\mu, g)\}_{s \in S \cup \{s_\emptyset\}}$ and let an *observation* of student i be captured by $(\mu, \pi(\mu, g))$.

Next, define any triple $(\succ'_{-i}, c'_{-i}, g'_{-i}) \in \mathcal{P}_{-i} \times \mathcal{C}_{-i} \times \mathcal{G}_{-i}$ as a *scenario* for student i. If i submits $(\hat{\succ}_i, \hat{c}_i)$ and observes $(\mu, \pi(\mu, g))$, then scenario $(\succ'_{-i}, c'_{-i}, g'_{-i})$ is *plausible* if $\pi(\mu, g) = \pi(\mu, (g_i, g'_{-i}))$ and $f((g_i, g'_{-i}), (\hat{\succ}_i, \succ'_{-i}), (\hat{c}_i, c'_{-i})) = \mu$. The set of all plausible scenarios for student i is her *inference set* $\mathcal{I}(\mu, \hat{\succ}_i, \hat{c}_i)$. Moreover, for student $i \in I$ who reports $(\hat{\succ}_i, \hat{c}_i)$ to f, let

$$\mathcal{M}|_{(\hat{\succ}_i, \hat{c}_i)} = \{\mu \in \mathcal{M} \mid \exists (\succ'_{-i}, c'_{-i}) \in \mathcal{P}_{-i} \times \mathcal{C}_{-i} : f(g, (\hat{\succ}_i, \succ'_{-i}), (\hat{c}_i, c'_{-i})) = \mu\}$$

be the set of matchings that could be *observed* by student i. Note that g is fixed in $\mathcal{M}|_{(\hat{\succ}_i, \hat{c}_i)}$, since it is a primitive of the market and independent of the report profile.

Having defined our observational structure, we are ready to introduce the notions of regret and regret-free truth-telling adopted from Fernandez (2020). Recall that all matching rules we study are consent-invariant. To simplify our notation, we define regret with a fixed consent decision for the student under consideration.

Definition 2.1. Fix consent decision $\hat{c}_i$. Student i *regrets* submitting $\hat{\succ}_i$ at $\mu \in \mathcal{M}|_{(\hat{\succ}_i, \hat{c}_i)}$ through $\hat{\succ}'_i$ under f if

1. $\forall (\succ'_{-i}, c'_{-i}, g'_{-i}) \in \mathcal{I}(\mu, \hat{\succ}_i, \hat{c}_i)$: $f_i((g_i, g'_{-i}), (\hat{\succ}'_i, \succ'_{-i}), (\hat{c}_i, c'_{-i})) \succeq_i \mu_i$

2. $\exists(\tilde{\succ}_{-i}, \tilde{c}_{-i}, \tilde{g}_{-i}) \in \mathcal{I}(\mu, \hat{\succ}_i, \hat{c}_i)$: $f_i((g_i, \tilde{g}_{-i}), (\hat{\succ}'_i, \tilde{\succ}_{-i}), (\hat{c}_i, \tilde{c}_{-i})) \succ_i \mu_i$.

In words, a student regrets her report at an observation if there is an alternative report which guarantees her a weakly better assignment in all plausible scenarios and realizes a strict improvement in at least one plausible scenario.

Definition 2.2. Fix consent decision $\hat{c}_i$. A report $\hat{\succ}_i$ is *regret-free* under f if there does not exist a pair $(\mu, \hat{\succ}'_i) \in \mathcal{M}|_{(\hat{\succ}_i, \hat{c}_i)} \times \mathcal{P}$ such that i regrets $\hat{\succ}_i$ at μ through $\hat{\succ}'_i$.

That is, a regret-free report ensures that regardless of the realized observation, the student does not regret her report.

In this chapter, we only consider matching rules that are invariant in the unacceptable set and define reports as truth-telling if the report differs from a student's true preferences only in the order within the unacceptable set. Formally, let $A_i(\succ_i) = \{s \in S | s \succ_i s_\emptyset\}$ collect all acceptable schools and let $U_i(\succ_i) = S \setminus A_i(\succ_i)$ collect all unacceptable schools. Furthermore, let

$$T_i(\succ_i) = \{\succ'_i \in \mathcal{P} \mid A_i(\succ'_i) = A_i(\succ_i) \text{ and } s \succ'_i s' \Leftrightarrow s \succ_i s',\ \forall s, s' \in A_i(\succ_i) \cup \{s_\emptyset\}\}$$

be the set of preferences which differ from $\succ_i$ by only allowing for permutations in $U_i(\succ_i)$. We say that for any i and her true preferences $\succ_i$, a report $\succ'_i \in \mathcal{P}$ is *truth-telling* if $\succ'_i \in T_i(\succ_i)$.

Definition 2.3. A matching rule f is *regret-free truth-telling* if for each problem and for each student, truth-telling is regret-free under f.

Strategy-proofness is stronger than regret-free truth-telling. That is, once truth-telling is weakly dominant under a matching rule, it must also be regret-free. However, the converse is not true. Specifically, strategy-proofness means that truth-telling is the weakly best option under *any* scenario, whereas regret-freeness only requires that, given a students' observation, no alternative report weakly dominates the truth under all *plausible scenarios*.

2.4 Main results

In this section, we present our main result. We show that a student can avoid regret under EDA if she submits her true preferences (Theorem 2.1) and that there is no other reporting behavior that provides the same guarantee (Proposition 2.2). As will be apparent from the corresponding proofs, all our results hold under the assumption that each student can only observe her own assignment and the cutoffs.

Theorem 2.1. *EDA is regret-free truth-telling.*

Proof. See Appendix 2.C. □

The following exposition provides an overview of the main arguments used in the formal proof. Fix any student $i \in I$, suppose that she reports her true preferences $\succ_i$ and she observes $(\mu, \pi(\mu, g))$. Then, any misreport $\tilde{\succ}_i$ can be interpreted as a combination of the following types of permutations, where relative to $\succ_i$:

(A1) for all $s, s' \in S, s \succ_i s'$ and $s' \tilde{\succ}_i s$ only if $s \in S \setminus L_{\mu_i}^{\succ_i}$;

(A2) there exists $s' \in S$ such that $\mu_i \succ_i s'$ and $s' \tilde{\succ}_i \mu_i$, or;

(A3) there exists $s, s' \in L_{\mu_i}^{\succ_i}$ such that $s, s' \in L_{\mu_i}^{\tilde{\succ}_i}$, $s \succ_i s'$ and $s' \tilde{\succ}_i s$.

Type (A1) involves all permutations relative to $\succ_i$ which keep the same ranking of all schools that are truly less preferred to μ_i. Type (A2) considers the misreports which rank some schools that are truly less preferred to μ_i as more preferred and type (A3) considers the misreports which alter the rankings among the schools that are truly less preferred to μ_i.

First note that any permutation $\tilde{\succ}_i$ of type (A1) relates to Proposition 2.1. If $(\tilde{\succ}_{-i}, \tilde{c}_{-i}, \tilde{g}_{-i})$ is plausible, then we have $EDA((g_i, \tilde{g}_{-i}), (\succ_i, \tilde{\succ}_{-i}), (c_i, \tilde{c}_{-i})) = \mu$ and we can apply Proposition 2.1 to obtain $EDA_i((g_i, \tilde{g}_{-i}), (\tilde{\succ}_i, \tilde{\succ}_{-i}), (c_i, \tilde{c}_{-i})) = \mu_i$.

Next, let student i choose a misreport $\tilde{\succ}_i$ that contains permutations of type (A2) and we write $\tilde{S} = \{s' \in S \mid \mu_i \succ_i s' \text{ and } s' \tilde{\succ}_i \mu_i\}$. The key arguments in the proof can roughly be divided into two categories: The submission of $\tilde{\succ}_i$ either would not have effectively influenced the assignment process at all, meaning i's assignment remains μ_i; or there is at least one plausible scenario in which the student is finally assigned to some $s^* \in \tilde{S}$. Here, we discuss the latter and more interesting case. The starting point of our argument is to construct a plausible scenario $(\tilde{\succ}_{-i}, \tilde{c}_{-i}, \tilde{g}_{-i})$ where i is assigned to s^* under $DA((g_i, \tilde{g}_{-i}), (\tilde{\succ}_i, \tilde{\succ}_{-i}))$. Then, we show that either the potential improvements that involve i cannot be realized because the consent of a student is missing; or s^* has no demand under $DA((g_i, \tilde{g}_{-i}), (\tilde{\succ}_i, \tilde{\succ}_{-i}))$. If each student could fully observe the consent decisions of other students, EDA is no longer regret-free truth-telling. Conversely, the uncertainty regarding other students' consent decisions is necessary for our result to hold.[10]

Finally, suppose that the misreport $\tilde{\succ}_i$ contains permutations of type (A3). The key argument for such a misreport is similar to that for type (A2): By submitting $\tilde{\succ}_i$, student i faces the possibility to be assigned to a less preferred school s^* whose order is permuted in $\tilde{\succ}_i$ and which is underdemanded under $DA((g_i, \tilde{g}_{-i}), (\tilde{\succ}_i, \tilde{\succ}_{-i}))$ for a plausible scenario $(\tilde{\succ}_{-i}, \tilde{c}_{-i}, \tilde{g}_{-i})$. However, different from type (A2), here the target school s^* still ranks below μ_i on $\tilde{\succ}_i$. This difference brings an additional challenge to the proof. While for (A2) it is enough to consider a plausible scenario where under truth-telling, i was already assigned to μ_i under DA, for (A3) we need to construct a scenario where under truth-telling, i is involved in at least one solved cycle to improve her from some school $\hat{s} \in L_{\mu_i}^{\succ_i}$ to μ_i. Then, when i submits $\tilde{\succ}_i$, she is assigned to the underdemanded s^* under DA and thus loses the opportunity to be involved in any cycle.

Our final result in this section shows that truth-telling is the unique regret-free choice under EDA.

Proposition 2.2. *For any non-truthful report, there exists an observation at which the student regrets it through truth-telling.*

[10] See Case 3.2.1 in Lemma 2.3 in Appendix 2.C for details.

Proof. See Appendix 2.D. □

At first glance, it might appear that Proposition 2.1 and Proposition 2.2 are in conflict with each other. However, Proposition 2.1 only implies that a certain class of misreports does not change the student's assignment when we fixed an observation that follows from her true preferences. In Proposition 2.2, however, the observation is not fixed. Instead, we show that given any non-truthful report, we can find a corresponding observation, such that truth-telling guarantees weakly better assignments in all plausible scenarios.

As an intuition for Proposition 2.2 note that for every misreport there must exist a pair, say school s and $\tilde{s}$, which compared to the truth, reverse their rankings. Let student i prefer s to $\tilde{s}$ under truth. Now suppose that upon submission of the misreport, she is assigned to $\tilde{s}$ while a seat at s is vacant. Note that the vacant seat at s allows i to infer that the truth would have guaranteed her at worst s. As a result, she will regret not having been truthful. The key step in the proof is to construct an observation of the type just described for any misreport.

2.5 Efficient stable dominating rules

In this section, we extend our analysis to *efficient stable dominating rules*, which are Pareto efficient and only produce outcomes which weakly Pareto dominate a stable matching. In contrast to EDA, consent decisions do not play a role under efficient stable dominating rules and from now on we omit the corresponding notation.

Definition 2.4. A matching rule f is *efficient stable dominating* if for any problem $(I, S, q, g, \succ)$ the matching $f(g, \succ)$ is Pareto efficient and weakly Pareto dominates a stable matching.

Efficient stable dominating rules are a natural refinement of *stable dominating rules*, introduced by Alva and Manjunath (2019). It is well known that efficient rules which Pareto dominate a stable

matching rule are not strategy-proof (Abdulkadiroğlu et al., 2009; Kesten, 2010). As we will show next, among efficient stable dominating rules also the weaker property of regret-free truth-telling cannot be fulfilled.

Theorem 2.2. *No efficient stable dominating rule is regret-free truth-telling.*

The proof below is constructive. We provide a problem with $|S| = 2$ and $|I| = 3$, and show that a student regrets submitting her true preferences under any efficient stable dominating rule. We only need small adaptions in the construction to apply the basic argument to any market with $|S| \geq 2$ and $|I| \geq 3$.

Proof. Consider a problem $(I, S, q, g, \succ)$ with two schools $S = \{s_1, s_2\}$ with capacities $q_{s_1} = q_{s_2} = 1$ and three students $I = \{i_1, i_2, i_3\}$. Suppose that i_1's true preferences $\succ_{i_1}$ are

$$s_2 \succ_{i_1} s_\emptyset \succ_{i_1} s_1.$$

Let $\succ_{-i} \in \mathcal{P}_{-i}$ be such that

$$s_1 \succ_{i_2} s_2 \succ_{i_2} s_\emptyset,$$
$$s_2 \succ_{i_3} s_1 \succ_{i_3} s_\emptyset.$$

and consider the following score structure g with

$$g_{i_1}^{s_1} > g_{i_3}^{s_1} > g_{i_2}^{s_1},$$
$$g_{i_2}^{s_2} > g_{i_1}^{s_2} > g_{i_3}^{s_2}.$$

The unique stable matching with respect to $\succ$ is

$$\nu = \{(\mathbf{i}_1, \mathbf{s}_\emptyset), (i_2, s_2), (i_3, s_1)\}$$

and that matching

$$\mu = \{(\mathbf{i}_1, \mathbf{s}_\emptyset), (i_2, s_1), (i_3, s_2)\},$$

is the unique Pareto efficient matching that Pareto dominates ν. Thus, for an arbitrary efficient stable dominating rule, denoted by f^{ESD}, we must have $f^{ESD}(\succ) = \mu$.

In the following, we construct a misreport $\tilde{\succ}_{i_1}$ through which i_1 regrets $\succ_{i_1}$ at observation $(\mu, \pi(\mu, g))$. Before we can make this misreport explicit, we need to describe i_1's inference set $\mathcal{I}(\mu, \succ_{i_1})$. To start, note that

$$g_{i_1}^{s_1} > \pi_{s_1}(\mu, g) \ , \ g_{i_1}^{s_2} > \pi_{s_2}(\mu, g).$$

We now show that any $\tilde{g}^{s_2}$ must share its ordinal ranking with g^{s_2} for any plausible score structure $\tilde{g}_{-i}$. First, from the observation $(\mu, \pi(\mu, g))$ student i_1 observes that her top choice s_2 is assigned to a lower priority student i_3, i.e. $\tilde{g}_{i_1}^{s_2} > \tilde{g}_{i_3}^{s_2}$. Second, if i_1 would have top priority at s_2 this would imply that i_1 is assigned to s_2 under any stable matching ν' whenever s_2 is submitted as her top choice. Thus, this must also hold true for any Pareto Efficient matching μ' that improves on ν' and hence i_1 can infer that student i_2 must have top priority at s_2. In conclusion, for any plausible $(\tilde{\succ}_{-i_1}, \tilde{g}_{-i_1})$, the corresponding $\tilde{g}^{s_2}$ shares the same ordinal ranking with g^{s_2}.

Next, given $\tilde{g}^{s_2}$, it must hold $\tilde{\succ}_{i_2} = \succ_{i_2}$. First, i_2 must submit s_2 as acceptable since otherwise any stable matching would assign s_2 to i_1. Therefore, i_1 knows $s_2 \tilde{\succ}_{i_2} s_\emptyset$. Second, note that since i_2 has top priority at s_2, f^{ESD} would have assigned s_2 to i_2 if i_2 would have submitted s_2 as her top choice. Thus, i_1 knows $s_1 \tilde{\succ}_{i_2} s_2$. Combining the two relations i_1 can infer that $\tilde{\succ}_{i_2} = \succ_{i_2}$ is the unique candidate contained in any plausible $(\tilde{\succ}_{-i_1}, \tilde{g}_{-i_1})$.

Now, we describe the candidates for $\tilde{g}^{s_1}$. First, by observing $(\mu, \pi(\mu, g))$, student i_1 knows that s_1 is assigned to the lower priority student i_2, i.e., $\tilde{g}^{s_1}_{i_1} > \tilde{g}^{s_1}_{i_2}$. Second, we establish that given the information regarding $\tilde{g}^{s_2}$ and $\tilde{\succ}_{i_2}$, we must have $\tilde{g}^{s_1}_{i_3} > \tilde{g}^{s_1}_{i_2}$. Suppose by contradiction that $\tilde{g}^{s_1}_{i_2} > \tilde{g}^{s_1}_{i_3}$. In this case, in f^{ESD}, i_1 and i_2 must be assigned to their top choices s_2 and s_1, respectively. However, this is incompatible with μ. Thus, there are two remaining ordinal rankings

$$\text{either } \tilde{g}^{s_1}_{i_1} > \tilde{g}^{s_1}_{i_3} > \tilde{g}^{s_1}_{i_2} \text{ or } \tilde{g}^{s_1}_{i_3} > \tilde{g}^{s_1}_{i_1} > \tilde{g}^{s_1}_{i_2}$$

that are compatible with any plausible scenario $(\tilde{\succ}_{-i_1}, \tilde{g}_{-i_1})$.

At last, we show that only $\tilde{\succ}_{i_3} = \succ_{i_3}$ is compatible with i_1's observation. First, since i_3 is assigned to s_2 in μ, student i_1 can conclude that $s_2 \tilde{\succ}_{i_3} s_\emptyset$. If i_3 would have submitted $s_\emptyset \tilde{\succ}_{i_3} s_1$, then any stable matching would have assigned both i_1 and i_2 to their top choices, which is incompatible with the observation. Thus, it must be true that $s_1 \tilde{\succ}_{i_3} s_\emptyset$. Furthermore, suppose by contradiction that $s_1 \tilde{\succ}_{i_3} s_2$. Given that $s_\emptyset \succ_{i_1} s_1$ and $\tilde{g}^{s_1}_{i_3} > \tilde{g}^{s_1}_{i_2}$, student i_3 is assigned to s_1 under f^{ESD}, which is again incompatible with observing μ. Hence, student i_3 can only have submitted $\tilde{\succ}_{i_3} = \succ_{i_3}$.

As a result, we can classify i_1's inference set $\mathcal{I}(\mu, \succ_{i_1})$ into two cases that are distinguished by the remaining candidates of ordinal rankings for scores at s_1.

We now show that i_1 regrets reporting the truth $\succ_{i_1}$ at $(\mu, \pi(\mu, g))$ through

$$\tilde{\succ}_{i_1} : s_2 \tilde{\succ}_{i_1} \mathbf{s}_1 \tilde{\succ}_{i_1} \mathbf{s}_\emptyset.$$

We do so by establishing that among the two possible classes from the inference set, in one class i_1 is strictly better off through the misreport and she is not worse off in the remaining class.

Case 1 Suppose that $(\tilde{\succ}_{-i_1}, \tilde{g}_{-i_1}) \in \mathcal{I}(\mu, \succ_{i_1})$ satisfies $\tilde{g}^{s_1}_{i_1} > \tilde{g}^{s_1}_{i_3} > \tilde{g}^{s_1}_{i_2}$. In this case, we argue that f^{ESD} must assign i_1 to s_2 when i_1 submits $\tilde{\succ}_i$. Hence, student i_1 would strictly improve her assign-

ment from $s_\emptyset$ under truth-telling to her top choice s_2. We first establish that there is a unique stable matching

$$\tilde{\nu} = \{(\mathbf{i}_1, \mathbf{s}_1), (i_2, s_2), (i_3, s_\emptyset)\}.$$

Note that in any stable matching i_1 cannot be assigned to $s_\emptyset$, since i_1 would have justified envy at s_1. This implies that whenever i_1 is not assigned to s_2, she must be assigned to s_1. Furthermore, if i_1 is matched with s_2, then i_2 must be assigned to s_1, which would mean that i_3 has justified envy at s_1. Thus, the unique stable matching corresponds to $\tilde{\nu}$. Hence, any efficient stable dominating rule must select

$$\tilde{\mu} = \{(\mathbf{i}_1, \mathbf{s}_2), (i_2, s_1), (i_3, s_\emptyset)\}$$

since it is the only Pareto efficient matching that dominates $\tilde{\nu}$. Thus, we conclude that conditional on her observation $(\mu, \pi(\mu, g))$, in this scenario, i_1 would have been better off if she had reported $\tilde{\succ}_{i_1}$ to f^{ESD}.

Case 2 It remains to show that given $(\tilde{\succ}_{-i_1}, \tilde{g}_{-i_1}) \in \mathcal{I}(\mu, \succ_{i_1})$ with $\tilde{g}_{i_3}^{s_1} > \tilde{g}_{i_1}^{s_1} > \tilde{g}_{i_2}^{s_1}$, student i_1 is not assigned to a worse option than under truth-telling (namely s_1). Clearly, in this case the unique stable matching is ν, while the unique matching that Pareto dominates ν is μ. Therefore, i_1 will be assigned to $s_\emptyset$ under f^{ESD}, which is the same assignment as under true preferences.

Since the choice of f^{ESD} was arbitrary, we have shown that for any efficient stable dominating rule, student i_1 regrets reporting the truth $\succ_{i_1}$ through misreport $\tilde{\succ}_{i_1}$ at $(\mu, \pi(\mu, g))$. This completes the proof. □

As mentioned before, this example allows us to illustrate one important feature of Theorem 2.1. Concretely, the observation $(\mu, \pi(\mu, g))$ at the beginning of the example is reached through EDA if

one leaves reported preferences unchanged and additionally requires that $c_{i_1} = 1$. Notice that in Case 1 the improvement of i_1's assignment from $s_\emptyset$ to s_2 relies on the consent of student i_3. However, based on i_1's observation, the consent decision of i_3 cannot be inferred by i_1. Specifically, if $c_{i_3} = 0$, then i_1 would be assigned to s_1 in Case 1 which implies that she would not regret that she had told the truth. Thus, EDA being regret-free truth-telling relies partially on the uncertainty regarding other students' consent decisions.

All our results extend to the more restrictive case where instead of observing the full matching μ, each student i observes only her own assignment μ_i and the cutoffs. For Theorem 2.2 this can be explained as follows. In the problem constructed above, there is only one additional consistent matching if i_1 observes only μ_{i_1}. For this matching, which switches the assignments for student i_2 and i_3 compared to μ, a symmetric argument leads to the same conclusion as for μ.

2.6 Conclusion

Telling the truth is a safe choice under EDA if students wish to avoid regret their submitted reports. Strengthening this first result, we have also shown that truth-telling is the unique regret-free option under EDA. Moreover, among the class of efficient stable dominating rules—a class that covers natural alternatives for EDA in practice—no candidate is regret-free truth-telling. Our results open up several avenues for future research. For instance, it would be interesting to study whether EDA is the unique candidate among all non-strategy-proof and constrained Pareto-Efficient rules which is regret-free truth-telling. It is also an open question whether EDA is still regret-free if schools' priorities take the form of more flexible choice functions.[11]

[11] Ehlers and Morrill (2019) introduce a generalized version of EDA that might serve as a starting point for an investigation.

2.A DA

In this section, we first introduce the Student Proposing Deferred Acceptance Algorithm which induces the *Student-Proposing Deferred Acceptance Rule (DA)* due to Gale and Shapley (1962). Thereafter, we present a lemma on *DA* that is necessary to prove Proposition 2.1 and Theorem 2.1. In the following, fix a problem $(I, S, q, g \succ, c)$. The DA algorithm works as follows:

Step 1 Each student $i \in I$ proposes to her most preferred school in $S \cup \{s_\emptyset\}$. Each school $s \in S$ considers all the proposals and tentatively accepts the candidates who apply to s and are among the q_s-highest ranked applicants at that school. The remaining proposals are rejected. If there are fewer than q_s proposals, s accepts all of them. Moreover, all students that propose to the outside option $s_\emptyset$ are accepted.

Step k, $k \geq 2$ Each student who was rejected at step $k-1$ applies to her most preferred school not yet applied to. Each school $s \in S$ considers all the new applicants together with those who are tentatively assigned to it at step $k-1$. Each school s now tentatively accepts the q_s-highest ranked applicants and rejects all others. If there are fewer than q_s proposals, s accepts all of them. Moreover, all students that propose to the outside option $s_\emptyset$ are accepted.

The algorithm terminates with the tentative assignments of the first step in which no student is rejected. For our lemma presented below we define Weak Maskin Monotonicity as in Kojima and Manea (2010). We call $\succ'$ a monotonic transformation of $\succ$ at matching μ, if for each $i' \in I$, $\succ'_{i'}$ is a monotonic transformation of $\succ_{i'}$ at $\mu_{i'}$.

Definition 2.5. A matching rule f is *weakly Maskin monotonic* if, given any $\succ$ and for any $\succ'$ that is a monotonic transformation of $\succ$ at $f(g, \succ, c), f(g, \succ', c)$ weakly Pareto dominates $f(g, \succ, c)$

Kojima and Manea (2010) show that DA is weakly Maskin monotonic. Furthermore, DA is strategy-proof (Dubins and Freedman, 1981; Roth, 1982) and produces the SOSM for a given score structure and preference profile.

Lemma 2.1. *Let $\succ_i' \in \mathcal{P}$ be a monotonic transformation of $\succ_i$ at $DA_i(g, \succ)$, then $DA(g, (\succ_i', \succ_{-i}))$ weakly Pareto dominates $DA(g, \succ)$ and i's outcomes are identical, i.e., $DA_i(g, \succ) = DA_i(g, (\succ_i', \succ_{-i}))$.*

Proof. The first part follows from weak Maskin monotonicity of DA. The second part is proved by means of contradiction. Suppose that $DA_i(g, \succ) \neq DA_i(g, (\succ_i', \succ_{-i}))$, then by weak Maskin monotonicity of DA, $DA_i(g, (\succ_i', \succ_{-i})) \succ_i DA_i(g, \succ)$, which contradicts strategy-proofness of DA. □

2.B Proof of Proposition 2.1

For ease of presentation, we use $EDA(\succ)$ to refer to $EDA(g, (\succ_i, \succ_{-i}), c)$ and $EDA(\tilde{\succ})$ to refer to $EDA(g, (\tilde{\succ}_i, \succ_{-i}), c)$. In a similar way, we use $DA(\succ)$ to refer to $DA(g, (\succ_i, \succ_{-i}))$ and $DA(\tilde{\succ})$ to refer to $DA(g, (\tilde{\succ}_i, \succ_{-i}))$.

We first show that the outcomes of EDA are identical under both profiles given that i consents, i.e., we prove that $EDA(\succ) = EDA(\tilde{\succ})$ when $c_i = 1$. At the end of the proof we extend our arguments to cover the case where $c_i = 0$.

Let $pTP^{\succ} = \{\gamma^t\}_{t=1}^{T}$ be an arbitrary realized process of the TP algorithm with input $(\succ, c, g)$ that are captured by the series of solved top priority cycles $\{\gamma^t\}_{t=1}^{T}$. Specifically, for each $t \leq T$, γ^t is solved at step t of $pTP^{\succ}$ and we set $EDA^t(\succ) = \gamma^t \circ EDA^{t-1}(\succ)$ with $EDA^0(\succ) = DA(\succ)$.

Since the outcome of the TP algorithm is invariant in the choice of the cycle solved in each round, it suffices to construct one TP process with input $((\tilde{\succ}_i, \succ_{-i}), c, g)$, denoted by $pTP^{\tilde{\succ}}$, that leads to the same outcome as $pTP^{\succ}$. As a part of our construction, we make use of the algorithm presented next.

INITIALIZE: Let $t = 1$. Also, let $\nu^0(\tilde{\succ}) = DA(\tilde{\succ})$ and $EDA^0(\succ) = DA(\succ)$.

ROUND $t \leq T$: Let $L^t = \{l \in I \mid \nu_l^{t-1}(\tilde{\succ}) \neq EDA_l^{t-1}(\succ)\}$.

- If each $jk \in \gamma^t$ satisfies that $j, k \in L^t$, let $\nu^t(\tilde{\succ}) = \nu^{t-1}(\tilde{\succ})$. Then, move to Round $t+1$ or terminate the algorithm if $t = T$.
- If there exists $jk \in \gamma^t$ such that $j \notin L^t$ or $k \notin L^t$, let $\nu^t(\tilde{\succ}) = \gamma^t \circ \nu^{t-1}(\tilde{\succ})$. Then, move to Round $t+1$ or terminate the algorithm if $t = T$.

Collect in $\{\tilde{\gamma}^t\}_{t=1}^{\tilde{T}}$ the series of cycles solved in the course of running the algorithm and note that, by construction, we have $\{\tilde{\gamma}^t\}_{t=1}^{\tilde{T}} \subseteq \{\gamma^t\}_{t=1}^{T}$. We now show that the generated cycle selection $\{\tilde{\gamma}^t\}_{t=1}^{\tilde{T}}$ allows to fully describe the desired $pTP^{\tilde{\succ}}$ which terminates at matching $EDA(\succ)$. Our strategy will be as follows. At the first step, we establish that the algorithm is well defined. At the second step, we will argue that $\nu^T(\tilde{\succ}) = EDA^T(\succ)$ and that $G^*(\nu^T(\tilde{\succ}))$ contains no cycles.

STEP 1 We can generate the desired sequence of cycles $\{\tilde{\gamma}^t\}_{t=1}^{\tilde{T}}$ if for each round $t \leq T$, the following three statements are satisfied:

(B1) Either all agents involved in γ^t belong to L^t, or none of them does.

(B2) $\gamma^t \in G^*(\nu^{t-1}(\tilde{\succ}))$ when γ^t contains no agent from L^t.

(B3) $\nu^t(\tilde{\succ})$ weakly Pareto dominates $EDA^t(\succ)$, and $L^{t+1} \subseteq L^t$.

For each t, statement (B1) and statement (B2) ensure that we can find and solve the cycle as described in the algorithm in round t. Then, given that (B1) and (B2) are true, statement (B3), is needed to ensure that (B1) and (B2) will also be true for the next round $t+1$. To prove these three statements, we now argue via induction over t.

For the initial case we build on the following observations. First, it is immediate from Lemma 2.1, that $DA(\tilde{\succ})$ weakly Pareto dominates $DA(\succ)$ and $DA_i(\succ) = DA_i(\tilde{\succ})$. Thus, we can infer that $L^1 = \{l \in I \mid DA_l(\tilde{\succ}) \succ_l DA_l(\succ)\}$ and $i \notin L^1$. Furthermore, by the definition of ν it is true that $DA_l(\tilde{\succ}) = \nu_l^0(\tilde{\succ})$ for any $l \in I$. Note that these conditions resemble those in condition (B3). Moreover, let $S' = \{s \in S \mid s \succ_i \mu_i \text{ and } \mu_i \tilde{\succ}_i s\}$.

We present our arguments in their general form since they are also applicable to the inductive step. That is, for the initial case we do not explicitly insert $t = 1$.

Initial case (Let $t = 1$): *Statement (B1):* Since γ' is a cycle, it suffices to show that $jk \in \gamma'$ and $k \in L^t$ imply $j \in L^t$.

Towards this goal, we first establish that for $jk \in \gamma'$, if $k \in L^t$, then $j \in L^t \cup \{i\}$. More generally, we show that for any $jk \in G^*(EDA^{t-1}(\succ))$, if $k \in L^t$, then we have $j \in L^t \cup \{i\}$. This generality will turn out to be useful proving other statements later on. By contradiction, let $j \notin L^t, j \neq i$ and $k \in L^t$. We aim at a contradiction towards the stability of $DA(\tilde{\succ})$. First, if $k \in L^t$, then there exists $l \in L^t$ such that $\nu_l^{t-1}(\tilde{\succ}) = EDA_k^{t-1}(\succ)$. Since $l \in L^t$, it holds $DA_l(\tilde{\succ}) = \nu_l^{t-1}(\tilde{\succ}) \succ_l EDA_l^{t-1}(\succ)$. Remarkably, for the initial case this argument is immediate since $DA_l(\tilde{\succ}) = \nu_l^0(\tilde{\succ}) \succ_l DA_l(\succ)$. When $t > 1$, the validity of this argument depends on the results we will establish later in the inductive step. Next, the previous observations and $jk \in G^*(EDA^{t-1}(\succ))$ imply that $g_j^{DA_l(\tilde{\succ})} > g_l^{DA_l(\tilde{\succ})}$ and $EDA_k^{t-1}(\succ) \succ_j EDA_j^{t-1}(\succ)$. Furthermore, $j \notin L^t$ implies $EDA_j^{t-1}(\succ) = \nu_j^{t-1}(\tilde{\succ}) \succeq_j DA_j(\tilde{\succ})$ while $j \neq i$ implies $\succ_j = \tilde{\succ}_j$. Combining the relations derived so far, leads to

$$DA_l(\tilde{\succ}) = \nu_l^{t-1}(\tilde{\succ}) = EDA_k^{t-1}(\succ) \tilde{\succ}_j EDA_j^{t-1}(\succ) = \nu_j^{t-1}(\tilde{\succ}) \tilde{\succeq}_j DA_j(\tilde{\succ}).$$

However, this implies that j has justified envy towards l at $DA(\tilde{\succ})$. Hence we arrive at a contradiction to the stability of $DA(\tilde{\succ})$ with respect to $\tilde{\succ}$.

It remains to show that $jk \in \gamma^t$ and $k \in L^t$ imply $j \neq i$. When $EDA_k^{t-1}(\succ) \notin S'$, the arguments above ensure $ik \notin G^*(EDA^{t-1}(\succ))$, and therefore also $ik \notin \gamma^t$. Consider the remaining case where $EDA_k^{t-1}(\succ) \in S'$. Here, if $ik \in \gamma^t$, then it implies that $EDA_k^{t-1}(\succ) = EDA_i^t(\succ) \succ_i \mu_i$. However, this is a contradiction to μ being the final matching. Thus, we must have $j \neq i$.

Now, we can conclude that once there is an edge $jk \in \gamma^t$ with $k \in L^t$, then $j \in L^t$. Therefore, either all agents involved in γ^t belong to L^t, or no such agent does. Statement (B1) is satisfied at round t.

Statement (B2): Given that (B1) is true at round t, we proceed to prove (B2). Suppose that for each $jk \in \gamma^t, j, k \notin L^t$. Thus, we get $EDA_j^{t-1}(\succ) = \nu_j^{t-1}(\tilde{\succ})$ and $EDA_k^{t-1}(\succ) = \nu_k^{t-1}(\tilde{\succ})$. This implies that

$$\nu_k^{t-1}(\tilde{\succ}) \; \tilde{\succ}_j \; \nu_j^{t-1}(\tilde{\succ}).$$

Note that this is also true if $j = i$, since in this case $\nu_k^{t-1}(\tilde{\succ}) \notin S'$. Hence, we obtain that student j must still desire $\nu_k^{t-1}(\tilde{\succ})$ at $\nu^{t-1}(\tilde{\succ})$. Note that the last argument is true for all j such that $jk \in \gamma^t$. Thus, we have that all students involved in γ^t are temporarily matched at $\nu^{t-1}(\tilde{\succ})$. Next, since $\nu^{t-1}(\tilde{\succ})$ weakly Pareto dominates $EDA^{t-1}(\succ)$, there are weakly fewer temporarily matched students who desire $\nu_k^{t-1}(\tilde{\succ})$ at $\nu^{t-1}(\tilde{\succ})$ compared to $EDA^{t-1}(\succ)$. As a result, j still has the highest score among all temporarily matched students pointing to k. Hence $jk \in G^*(\nu^{t-1}(\tilde{\succ}))$. Since this holds for all edges in γ^t, it follows that $\gamma^t \in G^*(\nu^{t-1}(\tilde{\succ}))$.

Statement (B3): We start with showing that the desired weak Pareto dominance relation holds at the end of round t. To begin with, note that $\nu^{t-1}(\tilde{\succ})$ weakly Pareto dominates $EDA^{t-1}(\succ)$ and that if any, only students in γ^t change their assignments in round t of our algorithm (and also in round t of $pTP^{\succ}$). Thus, to conclude that $\nu^t(\tilde{\succ})$ weakly Pareto dominates $EDA^t(\succ)$, it is sufficient to show that for each $jk \in \gamma^t$:

$$\nu_j^t(\tilde{\succ}) \succeq_j EDA_j^t(\succ).$$

Of the two cases we have to consider, we start with the simpler one, in which for any $jk \in \gamma^t$, we have

$j, k \notin L^t$. In this case, γ^t is solved in both $\nu^{t-1}(\tilde{\succ})$ and $EDA^{t-1}(\succ)$. Therefore, $\nu_j^t(\tilde{\succ}) = EDA_j^t(\succ)$ and we obtain the desired result.

In the remaining case, any $jk \in \gamma^t$ satisfies that $j, k \in L^t$. Clearly, we can solve a cycle of this form only if $L^t \neq \emptyset$. Moreover, note that $EDA^t(\succ) = \gamma^t \circ EDA^{t-1}(\succ)$ and $\nu^t(\tilde{\succ}) = \nu^{t-1}(\tilde{\succ})$. We proceed by contradiction and assume that $EDA_j^t(\succ) \succ_j \nu_j^t(\tilde{\succ})$. We derive a contradiction to the stability of $DA(\tilde{\succ})$ with respect to $\tilde{\succ}$. We make the following observations: First, since $k \in L^t$, there must exist $l \in L^t$ such that we have $\nu_l^{t-1}(\tilde{\succ}) = EDA_k^{t-1}(\succ)$. Second, $l \in L^t$ implies that $DA_l(\tilde{\succ}) = \nu_l^{t-1}(\tilde{\succ}) \succ_l EDA_l^{t-1}(\succ)$. Therefore, $jk \in \gamma^t$ also means that $g_j^{DA_l(\tilde{\succ})} > g_l^{DA_l(\tilde{\succ})}$ and $EDA_k^{t-1}(\succ) = EDA_j^t(\succ)$. Third, the algorithm guarantees that $\nu_j^t(\tilde{\succ}) \succeq_j DA_j(\tilde{\succ})$. If we combine all relations above with $\succ_j = \tilde{\succ}_j$, we obtain

$$DA_l(\tilde{\succ}) = \nu_l^{t-1}(\tilde{\succ}) = EDA_k^{t-1}(\succ) = EDA_j^t(\succ) \,\tilde{\succ}_j\, \nu_j^t(\tilde{\succ}) \,\tilde{\succeq}_j\, DA_j(\tilde{\succ})$$

and reach a contradiction, since j has justified envy towards l at $DA(\tilde{\succ})$. Thus, $\nu^t(\tilde{\succ})$ weakly Pareto dominates $EDA^t(\succ)$. Moreover, based on the Pareto dominance result, we can also write L^{t+1} as $L^{t+1} = \{l \in I \mid \nu_l^t(\tilde{\succ}) \succ_l EDA_l^t(\succ)\}$.

To finish the proof for statement (B3) we need to show that $L^{t+1} \subseteq L^t$ for which we again have two cases to consider. If any $jk \in \gamma^t$ satisfies $j, k \notin L^t$, then it is immediate that $L^{t+1} = L^t$. On the contrary, if any $jk \in \gamma^t$ satisfies $j, k \in L^t$, then we make the following two observations. First, for each such j, as $j \in L^t$, we have $\nu_j^{t-1}(\tilde{\succ}) \succ_j EDA_j^{t-1}(\succ)$ and $\nu_j^t(\tilde{\succ}) \succeq_j EDA_j^t(\succ)$. This implies that while j is contained in L^t, she might not be in L^{t+1}. Second, for each $j' \in I$ not involved in γ^t, we have $\nu_{j'}^t(\tilde{\succ}) = \nu_{j'}^{t-1}(\tilde{\succ})$ and $EDA_{j'}^t(\succ) = EDA_{j'}^{t-1}(\succ)$, which implies that $j' \in L^t$ if and only if $j' \in L^{t+1}$. In conclusion, we can infer that $L^{t+1} \subseteq L^t$. Hence statement (B3) is satisfied.

INDUCTIVE STEP: Let $t > 1$ and assume that for all $t' < t$, (B1) - (B3) are satisfied. By assumption of the inductive step and from (B3), we have that $L^{t'+1} \subseteq L^{t'}$ for any $t' < t$, which implies $L^t \subseteq L^{t'}$. Second, through the description of the algorithm, we know that given any $t' < t$, assignments at $\nu^{t'}(\tilde{\succ})$ and $\nu^{t'-1}(\tilde{\succ})$ are identical for each student in $L^{t'}$. Therefore, since $L^t \subseteq L^{t'}$, we can infer that for each $l \in L^t$, $DA_l(\tilde{\succ}) = \nu_l^{t-1}(\tilde{\succ})$. Together with the observations above, the arguments we already presented for (B1) in the initial case also apply to the inductive step. Furthermore, the same holds for (B2). Finally, given we established (B1) and (B2), also (B3) follows again from the same arguments as in the initial case. This completes the induction.

STEP 2: We show that $EDA^T(\succ) = \nu^T(\tilde{\succ})$. Let $t_i \leq T$ be the first step in $pTP^{\succ}$ where i is permanently matched and consider round t_i of our algorithm.

If $EDA^{t_i-1}(\succ) = \nu^{t_i-1}(\tilde{\succ})$, we have that $L^t = \emptyset$ and that γ^t is solved in each round $t > t_i$ of the algorithm. Consequently, it is true that $EDA^T(\succ) = \nu^T(\tilde{\succ})$.

If $EDA^{t_i-1}(\succ) \neq \nu^{t_i-1}(\tilde{\succ})$, then L^{t_i} is non-empty. In this case, we show that there exists $\hat{t} > t_i$ such that $EDA^{\hat{t}}(\succ) = \nu^{\hat{t}}(\tilde{\succ})$. As shown above, this leads to $EDA^T(\succ) = \nu^T(\tilde{\succ})$.

We show that there must be a cycle in $G^*(EDA^{t_i-1}(\succ))$ that solely consists of elements in L^{t_i}. We begin with showing that for any $k \in L^{t_i}$, there exists an edge $jk \in G^*(EDA^{t_i-1}(\succ))$ for some $j \in I$. Since $k \in L^{t_i}$, there exists $l \in L^{t_i}$ such that $EDA_k^{t_i-1}(\succ) = \nu_l^{t_i-1}(\tilde{\succ}) \succ_l EDA_l^{t_i-1}(\succ)$. That is, at $EDA^{t_i-1}(\succ)$, for each student in L^{t_i}, her assignment is desired by at least one student in L^{t_i} whose assignment is further desired by some other student in L^{t_i}. Now, recall that we assume $c_1 = 1$. Since i is permanently matched at step t_i and i consents, then even if i prefers $EDA_k^{t_i-1}(\succ)$ to μ_i, she cannot prevent any agent from being eligible for $EDA_k^{t_i-1}(\succ)$. In other words, at least one edge that is pointing to k, namely lk, is contained in $G(EDA^{t_i-1}(\succ))$. Therefore, we can infer that k is temporarily matched in $EDA^{t_i-1}(\succ)$ and thus there must be $jk \in G^*(EDA^{t_i-1}(\succ))$ for some $j \in I$.

Next, for any such jk, our arguments from (B1) will be sufficient to conclude that $j \in L^{t_i}$. First, we

have already shown $j \in L^{t_i} \cup \{i\}$. Second, we know that $j \neq i$, since i is permanently matched. Thus, we can infer that each student in L^{t_i} is pointed by another student in L^{t_i} in $G^*(EDA^{t_i-1}(\succ))$. Since L^{t_i} is finite, the existence of the desired cycle is guaranteed. Notably, according to (B3) and by iteratively applying the same argument, we can eventually reach a round $\hat{t} > t_i$ where $EDA^{\hat{t}}(\succ) = \nu^{\hat{t}}(\tilde{\succ})$.

We next claim that no cycles can be found in $G^*(\nu^T(\tilde{\succ}))$. Notably, if $G^*(\nu^T(\tilde{\succ}))$ has a cycle, by similar arguments in (B2), we can infer that $G^*(EDA^T(\succ))$ must also have a cycle. However, this contradicts the fact that exactly T cycles are solved in $pTP^{\succ}$.

Based on the statements provided so far, we can construct the desired $pTP^{\tilde{\succ}}$ as $pTP^{\tilde{\succ}} = \{\tilde{\gamma}^t\}_{t=1}^{\tilde{T}}$. This leads to

$$EDA(\succ) = EDA(\tilde{\succ})$$

which completes the proof for $c_i = 1$.

Finally, it remains to prove that our results extend to the case where $c_i = 0$. Note that since EDA is consent-invariant, the following two relations are true: $EDA_i(\succ) = EDA_i(g, (\succ_i, \succ_{-i}), (\tilde{c}_i, c_{-i}))$ and $EDA_i(\tilde{\succ}) = EDA_i(g, (\tilde{\succ}_i, \succ_{-i}), (\tilde{c}_i, c_{-i}))$ for $\tilde{c}_i = 1$. Since we just showed when i consents, submitting $\tilde{\succ}_i$ will not alter the outcome: $EDA(g, (\succ_i, \succ_{-i}), (\tilde{c}_i, c_{-i})) = EDA(g, (\tilde{\succ}_i, \succ_{-i}), (\tilde{c}_i, c_{-i}))$. This allows us to conclude $EDA_i(\succ) = EDA_i(\tilde{\succ})$, which completes the proof.

2.C Proof of Theorem 2.1

Fix an arbitrary problem $(I, S, q, g, \succ, c)$ and consider an arbitrary student $i \in I$. Since EDA only takes acceptable schools into account, for any tuple $(g, \succ_{-i}, c)$ and any $\succ_i' \in T_i$, we can claim that $EDA(g, (\succ_i', \succ_{-i}), c) = EDA(g, (\succ_i, \succ_{-i}), c)$. Hence, if student i does not regret reporting her true preferences $\succ_i$, she does not regret to report any $\succ_i' \in T_i$. Thus, we show that i does not regret to report $\succ_i$.

Lemma 2.2, 2.3 and 2.7 each consider a distinct class of misreports of student i and jointly imply that i cannot regrets submitting her true preferences. In the following exposition, take an arbitrary observation $(\mu, \pi(\mu, g))$ where $\mu \in \mathcal{M}|_{(\succ_i, c_i)}$. We fix i's scores g_i and i's consent decision c_i throughout the proof. From now on, we use $\tilde{g}$ to refer to $(g_i, \tilde{g}_{-i})$ and $\tilde{c}$ to refer to $(c_i, \tilde{c}_{-i})$.

We first show that a misreport is not profitable for i, if it shares the same relative ranking of schools weakly below her own assignment under truth-telling.

Lemma 2.2. *Consider $\tilde{\succ}_i \in \mathcal{P}$ such that for all $s, s' \in L_{\mu_i}^{\succ_i}$, $s \tilde{\succ}_i s'$ if and only if $s \succ_i s'$. For any $(\tilde{\succ}_{-i}, \tilde{c}_{-i}, \tilde{g}_{-i}) \in \mathcal{I}(\mu, \succ_i, c_i)$, it is true that $EDA_i(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i}), \tilde{c}) = \mu_i$.*

Proof. Select any $(\tilde{\succ}_{-i}, \tilde{c}_{-i}, \tilde{g}_{-i}) \in \mathcal{I}(\mu, \succ_i, c_i)$. By definition, $EDA(\tilde{g}, (\succ_i, \tilde{\succ}_{-i}), \tilde{c}) = \mu$ and using Proposition 2.1, we know $EDA_i(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i}), \tilde{c}) = \mu_i$. □

We proceed with misreports in which some schools ranked below μ_i under truth permute their order with μ_i. Our next Lemma shows that the student can either infer that she would have possibly been worse off, or that the misreport would not have affected her assignment in any plausible scenario.

Lemma 2.3. *Consider $\tilde{\succ}_i \in \mathcal{P}$ such that $\mu_i \succ_i s$ and $s \tilde{\succ}_i \mu_i$ for some $s \in S$. Then,*
(1) either there exists $(\tilde{\succ}_{-i}, \tilde{c}_{-i}, \tilde{g}_{-i}) \in \mathcal{I}(\mu, \succ_i, c_i)$ such that $\mu_i \succ_i EDA_i(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i}), \tilde{c})$;
(2) or for any $(\tilde{\succ}_{-i}, \tilde{c}_{-i}, \tilde{g}_{-i}) \in \mathcal{I}(\mu, \succ_i, c_i)$: $EDA_i(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i}), \tilde{c}) = \mu_i$.

Proof. Let $\tilde{S} = \{s' \in S \mid \mu_i \succ_i s' \text{ and } s' \tilde{\succ}_i \mu_i\}$. We start by considering the case where $\tilde{S} = \{s^*\}$ is a singleton. We will explain how to generalize the arguments to cases where $\tilde{S}$ contains more elements at the end of the proof. Given that $\tilde{S}$ is a singleton, we distinguish the following exhaustive cases based on i's observation $(\mu, \pi(\mu, g))$:

Case 1: $\pi_{s^*}(\mu, g) = 0$. If $\pi_{s^*}(\mu, g) = 0$, then s^* has not exhausted its capacity at the observed matching. We use the following argument repeatedly throughout the proof: Note that students

assigned to a school that has not exhausted its capacity under the observed matching cannot be involved in a cycle in any corresponding TP process for any plausible scenario. This implies that at this school also under DA the same set of students must have been assigned there. Furthermore, since DA is non-wasteful, we can conclude that at any plausible scenario the school has no demand under the DA matching. Concretely, since for any $(\tilde{\succ}_{-i}, \tilde{c}_{-i}, \tilde{g}_{-i}) \in \mathcal{I}(\mu, \succ_i, c_i)$, s^* has vacant seat at $EDA(\tilde{g}, (\succ_i, \tilde{\succ}_{-i}), \tilde{c}) = \mu$, then s^* must also have vacant seat at $DA(\tilde{g}, (\succ_i, \tilde{\succ}_{-i}))$ and that for any $i' \in I$, i' weakly prefers $DA_{i'}(\tilde{g}, (\succ_i, \tilde{\succ}_{-i}))$ to s^* given preference profile $\succ$. That is, s^* has no demand.

Next, if i submits $\tilde{\succ}_i$ then we obtain $DA_i(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i})) = s^*$. Now notice that before being matched to the final assignment, the set of applications i sends to reach $DA_i(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i}))$ is a subset of those sent to reach $DA_i(\tilde{g}, (\succ_i, \tilde{\succ}_{-i}))$. Therefore, $DA_{i'}(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i})) \succeq_{i'} DA_{i'}(\tilde{g}, (\succ_i, \tilde{\succ}_{-i}))$ holds for all $i' \neq i$. Accordingly, each agent $i' \in I$ still weakly prefers $DA_{i'}(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i}))$ to s^* given preference profile $\tilde{\succ}$. Hence s^* has again no demand at $DA(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i}))$ and thus no agent is pointing to i in $G^*(DA(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i})))$. As a result, i cannot be involved in any solved cycle during the TP process and thus $EDA_i(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i}), \tilde{c}) = DA_i(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i})) = s^*$. Statement (1) holds.

Case 2: $\pi_{s^*}(\mu, g) \neq 0$, $\pi_{\mu_i}(\mu, g) = 0$ and $g_i^{s^*} < \pi_{s^*}(\mu, g)$. Under this condition, we show that statement (2) is satisfied. Take an arbitrary $(\tilde{\succ}_{-i}, \tilde{c}_{-i}, \tilde{g}_{-i}) \in \mathcal{I}(\mu, \succ_i, c_i)$. To start, note that whenever a student j improves her assignment from one school to another at one step of the TP algorithm, another student with lower priority is assigned to the school that j left at that step. Since $g_i^{s^*} < \pi_{s^*}(\mu, g)$, this implies that student i must have a lower score than any student assigned to s^* at $DA_i(\tilde{g}, (\succ_i, \tilde{\succ}_{-i}))$. Thus, compared to the DA procedure of i submitting $\succ_i$, i's additional application to s^* by submitting $\tilde{\succ}_i$ has no influence on the outcome and we reach $DA(\tilde{g}, (\succ_i, \tilde{\succ}_{-i})) = DA(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i}))$. Moreover, since $\pi_{\mu_i}(\mu, g) = 0$ and as argued in Case 1, μ_i must have vacant seat at $DA(\tilde{g}, (\succ_i, \tilde{\succ}_{-i}))$, thus also at $DA(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i}))$. As a result, μ_i has no

demand at $DA(\tilde{g}, (\tilde{\succ}_i.\tilde{\succ}_{-i}))$ and this implies that $EDA_i(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i}), \tilde{c}) = DA_i(\tilde{g}, (\tilde{\succ}_i.\tilde{\succ}_{-i})) = \mu_i$. Hence, statement (2) holds.

Case 3: $\pi_{s^*}(\mu, g) \neq 0$ and either (C1) $g_i^{s^*} > \pi_{s^*}(\mu, g)$; or (C2) $\pi_{\mu_i}(\mu, g) \neq 0$ and $g_i^{s^*} < \pi_{s^*}(\mu, g)$. Throughout the discussion, we will make it explicit whenever (C1) and (C2) are in need to be distinguished.[12] Furthermore, except for the last subcase (Case 3.2.2.2), statement (1) will apply in Case 3 and our approach for each subcase except this last subcase will be standardized going through the following steps:

Step 1: We construct a candidate scenario $(\tilde{\succ}_{-i}, \tilde{c}_{-i}, \tilde{g}_{-i})$.

Step 2: We show that $(\tilde{\succ}_{-i}, \tilde{c}_{-i}, \tilde{g}_{-i}) \in \mathcal{I}(\mu, \succ_i, c_i)$.

Step 3: We argue that $EDA_i(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i}), \tilde{c}) = s^*$.

Let $j \in I$ be such that $\mu_j = s^*$ and $g_j^{s^*} = \pi_{s^*}(\mu, g)$. Let $\hat{S} = \{s_1, \ldots, s_T\}$ be the set of schools for which i has justified envy at μ and assume without loss of generality $s_1 \succ_i s_2 \succ_i \ldots \succ_i s_T$. Note that our constructions of the candidate scenarios $(\tilde{\succ}_{-i}, \tilde{c}_{-i}, \tilde{g}_{-i})$ below varies for different cardinalities of $\hat{S}$.

In the following, for any $\succ_i' \in \mathcal{P}$ and any $s \in S$, we denote the strict lower contour set of $\succ_i'$ at s by $SL_s^{\succ_i'} = \{s' \in S \mid s \succ_i' s'\}$ and the strict upper contour set of $\succ_i'$ at s by $SU_s^{\succ_i'} = \{s' \in S \mid s' \succ_i' s\}$. Notably, the following observations on $\hat{S}$ will be helpful:

- $\hat{S} = \emptyset$ whenever $c_i = 0$, since EDA does not allow for any priority violations for i.
- If $\hat{S} \neq \emptyset$, non-wastefulness of EDA implies that for each $s' \in \hat{S}$, $\pi_{s'}(\mu, g) \neq 0$.
- Since $\hat{S} \subseteq SU_{\mu_i}^{\succ_i}$ and $s^* \in SL_{\mu_i}^{\succ_i}$, $s^* \notin \hat{S}$.

[12] Note that since we assume that $g_i^s \neq g_j^s$ for any $i, j \in I$ and any $s \in S$, it cannot be true that $\pi_{s^*}(\mu, g) = g_i^{s^*}$, when $i \notin \mu_{s^*}$.

Next, for each $t \in \{1, \ldots, T\}$, let $i_t \in \mu_{s_t}$ be such that $g_{i_t}^{s_t} = \pi_{s_t}(\mu, g)$. Collect all such students in $\hat{I} = \{i_1, \ldots, i_T\}$. For each $i_t \in \hat{I}$, in any TP process corresponding to a plausible scenario, there must exist a solved cycle γ such that $i_t k \in \gamma$ for some $k \in I$ and i_t is assigned to s_t when γ is solved. Moreover, solving γ must be the last step in that TP process in which i_t is improved.

Case 3.1: $|\hat{S}| \neq 1$. For now, assume that (C2) is satisfied. At the end of this subcase, we present a slight modification needed in the construction for (C1).

Step 1: We start with the candidate score structure $\tilde{g}_{-i}$:

- let $\tilde{g}_i^{\mu_i} \geq \pi_{\mu_i}(\mu, g) > \tilde{g}_j^{\mu_i}$ and;
- for any $s' \in S \setminus \{\hat{S} \cup \mu_i\}$ let $\tilde{g}^{s'} = g^{s'}$.

Let $i_0 = i_T$ and $s_{T+1} = s_1$. In case that $\hat{S} \neq \emptyset$:

- for each $s_t \in \hat{S}$, let $\tilde{g}^{s_t}$ be such that $\tilde{g}_{i_{t-1}}^{s_t} > \tilde{g}_i^{s_t} > \tilde{g}_{i_t}^{s_t}$ and for all $l \in \mu_{s_t}$ with $l \neq i_t$, let $\tilde{g}_l^{s_t} > \tilde{g}_{i_{t-1}}^{s_t}$.

Next, select an arbitrary $\tilde{c}_{-i}$ and consider the following preferences $\tilde{\succ}_{-i}$:

$$\mu_i \,\tilde{\succ}_j\, s^* \,\tilde{\succ}_j\, s_\emptyset \,\tilde{\succ}_j\, \ldots,$$

$$s_t \,\tilde{\succ}_{i_t}\, s_{t+1} \,\tilde{\succ}_{i_t}\, s_\emptyset \,\tilde{\succ}_{i_t}\, \ldots \quad \forall t \in \{1, \ldots, T\},$$

$$\mu_k \,\tilde{\succ}_k\, s_\emptyset \,\tilde{\succ}_k\, \ldots \quad \forall k \in I \setminus (\hat{I} \cup \{i, j\}).$$

Step 2: As one can easily see, we have $\pi(\mu, (g_i, \tilde{g}_{-i})) = \pi(\mu, g)$. We now show that the constructed scenario $(\tilde{\succ}_{-i}, \tilde{c}_{-i}, \tilde{g}_{-i})$ yields μ under the TP algorithm. We have two cases to consider: First, if $\hat{S} = \emptyset$, we get $DA(\tilde{g}, (\succ_i, \tilde{\succ}_{-i})) = \mu$ and the TP process terminates with μ since there are no cycles $G^*(\mu)$. Second, suppose that $\hat{S} \neq \emptyset$. We describe how we arrive at the corresponding DA outcome: $DA_k(\tilde{g}, (\succ_i, \tilde{\succ}_{-i})) = \mu_k$ for all $k \in I \setminus \hat{I}$ and $DA_{i_t}(\tilde{g}, (\succ_i, \tilde{\succ}_{-i})) = s_{t+1}$ for all $i_t \in \hat{I}$. Note that

every student $k \in I \setminus \{i,j\}$ considers her assigned school (μ_k) as the top choice in $\tilde{\succ}_{-i}$ and each such student k gets accepted by her top choice at the first step of the corresponding DA process. Moreover, at some step, student i applies to s_1 and gets tentatively accepted. This triggers a series of rejections. Specifically, for each $t \in \{1,\dots,T\}$, i_t gets rejected by s_t and applies to s_{t+1} in the next step, causing i_{t+1} being rejected by s_{t+1} and so forth. This rejection chain ends with i_T applying to s_1 which leads i to be rejected by s_1. Thereafter, i applies to all schools in $SU_{\mu_i}^{\succ_i} \setminus SU_{s_1}^{\succ_i}$ and is rejected until finally being accepted by μ_i. At last, j is rejected by μ_i and applies to s^* to which she is finally assigned in DA.

There is a unique cycle $\gamma = \{i_T i_{T-1}, \dots, i_2 i_1, i_1 i_T\}$ in $G^*(DA(\tilde{g}, (\succ_i, \tilde{\succ}_{-i})))$ which, once solved, produces μ. According to $(\succ_i, \tilde{\succ}_{-i})$, i and j are the only students who do not receive their top choice in μ and therefore the TP algorithm terminates with μ.

Step 3: First, be aware that the outcome $DA(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i}))$ may vary in the position of s^* on $\tilde{\succ}_i$:

- If $s^* \tilde{\succ}_i s_1$, then $DA_i(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i})) = s^*$, $DA_j(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i})) = \mu_i$, $DA_k(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i})) = \mu_k$ for any $k \in I \setminus \{i,j\}$.
- If $s_1 \tilde{\succ}_i s^*$, then $DA_i(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i})) = s^*$, $DA_j(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i})) = \mu_i$, $DA_{i_t}(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i})) = s_{t+1}$ for any $i_t \in \hat{I}$ and $DA_k(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i})) = \mu_k$ for any $k \in I \setminus (\{i,j\} \cup \hat{S})$.

In both instances above s^* has no demand at $DA(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i}))$. As a result, we have that $EDA_i(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i}), \tilde{c}) = s^*$ and thus the argument for (C2) is complete.

Now suppose that (C1) holds. The construction above does not work here generally, since when $\pi_{\mu_i}(\mu, g) = 0$, both i and j get finally assigned to μ_i in $EDA(\tilde{g}, (\succ_i, \tilde{\succ}_{-i}), \tilde{c})$. We make the following adjustments in the construction:

Step 1: Modify the preferences of j to be

$$s^* \tilde{\succ}_j s_\emptyset \tilde{\succ}_j \dots,$$

and keep all other details of our construction the same as in instance (C2) above.

Step 2 and Step 3: The arguments resemble those in instance (C2) above.

Case 3.2: $|\hat{S}| = 1$. The construction in Case 3.1 does not work here. Specifically, we cannot construct a cycle that consists of students in $\hat{I}$ when $|\hat{I}| = |\hat{S}| = 1$. Cycles will therefore contain students not in $\hat{I}$ and moreover, from (C1) to (C2), we need to alter the identity of students involved in the cycle:

Case 3.2.1: $g_i^{s^*} > \pi_{s^*}(\mu, g)$. That is, (C1) holds and we have $g_i^{s^*} > g_j^{s^*}$.

Step 1: Let $\tilde{g}_{-i}$ be such that

- $\tilde{g}_j^{s_1} > \tilde{g}_i^{s_1} > \tilde{g}_{i_1}^{s_1}$;
- $\tilde{g}_i^{s^*} > \tilde{g}_{i_1}^{s^*} > \tilde{g}_j^{s^*}$;
- $\tilde{g}^{s'} = g^{s'}$ for any $s' \in S \setminus \{s^*, s_1\}$.

Now, let $\tilde{c}_{-i}$ be such that $\tilde{c}_{i_1} = 0$[13] and consider the following profile $\tilde{\succ}_{-i}$:

$$s^* \,\tilde{\succ}_j\, s_1 \,\tilde{\succ}_j\, s_\emptyset \ldots,$$

$$s_1 \,\tilde{\succ}_{i_1}\, s^* \,\tilde{\succ}_{i_1}\, s_\emptyset \ \ldots,$$

$$\mu_k \,\tilde{\succ}_k\, s_\emptyset \,\tilde{\succ}_k \ \ldots \ \ \forall k \in I \setminus \{i, j, i_1\}.$$

Step 2: First, it is easily checked $\pi(\mu, (g_i, \tilde{g}_{-i})) = \pi(\mu, g)$. Following a similar application procedure as in Case 3.1, the DA algorithm leads to $DA_j(\tilde{g}, (\succ_i, \tilde{\succ}_{-i})) = s_1$, $DA_{i_1}(\tilde{g}, (\succ_i, \tilde{\succ}_{-i})) = s^*$ and $DA_k(\tilde{g}, (\succ_i, \tilde{\succ}_{-i})) = \mu_k$ for all $k \in I \setminus \{j, i_1\}$. There is a unique cycle $\gamma = \{i_1 j, j i_1\}$ in

[13] It is worth mentioning that this is the only place in the proof of Theorem 2.1, where we need a scenario where a student does not consent.

$G^*(DA(\tilde{g},(\succ_i,\tilde{\succ}_{-i})))$ and once this cycle is solved we obtain μ. In this instance, all students except i receive their top choice in μ. The TP algorithm thus terminates and $EDA(\tilde{g},(\succ_i,\tilde{\succ}_{-i}),\tilde{c})=\mu$.

Step 3: Note that the DA algorithm arrives at $DA_i(\tilde{g},(\tilde{\succ}_i,\tilde{\succ}_{-i}))=s^*$, $DA_j(\tilde{g},(\tilde{\succ}_i,\tilde{\succ}_{-i}))=s_1$, $DA_{i_1}(\tilde{g},(\tilde{\succ}_i,\tilde{\succ}_{-i}))=s_\emptyset$ and $DA_k(\tilde{g},(\tilde{\succ}_i,\tilde{\succ}_{-i}))=\mu_k$ for all $k\in I\backslash\{i,j,i_1\}$. Also, j is not eligible for s^* since $\tilde{c}_{i_1}=0$. Therefore, we cannot add ji to the graph and thus there is no cycle in $G^*(DA(\tilde{g},(\tilde{\succ}_i,\tilde{\succ}_{-i})))$. In conclusion, $EDA_i(\tilde{g},(\tilde{\succ}_i,\tilde{\succ}_{-i}),\tilde{c})=s^*$.

CASE 3.2.2: $\pi_{\mu_i}(\mu,g)\neq 0$ AND $g_i^{s^*}<\pi_{s^*}(\mu,g)$ That is, (C2) holds and we thus have $g_i^{s^*}<g_j^{s^*}$.

Case 3.2.2.1: There exists $s'\in S\setminus\{s_1,\mu_i,s^*\}$ such that $\pi_{s'}(\mu,g)\neq 0$. Pick an arbitrary such s' and denote with j' the student who has the lowest score among all students being assigned to s' under μ.

Step 1: Let $\tilde{g}_{-i}$ be such that

- $\tilde{g}_{j'}^{s_1}>\tilde{g}_i^{s_1}>\tilde{g}_{i_1}^{s_1}$;
- $\tilde{g}_{i_1}^{s'}>\tilde{g}_{j'}^{s'}$;
- $\tilde{g}_i^{\mu_i}>g_j^{\mu_i}$;
- $\tilde{g}^{s''}=g^{s''}$ for any $s''\in S\setminus\{s_1,\mu_i,s'\}$.

Next, fix an arbitrary $\tilde{c}_{-i}$ and consider the following profile $\tilde{\succ}_{-i}$:

$$\mu_i\,\tilde{\succ}_j\,s^*\,\tilde{\succ}_j\,s_\emptyset\,\tilde{\succ}_j\,\ldots,$$

$$s_1\,\tilde{\succ}_{i_1}\,s'\,\tilde{\succ}_{i_1}\,s_\emptyset\,\tilde{\succ}_{i_1}\,\ldots,$$

$$s'\,\tilde{\succ}_{j'}\,s_1\,\tilde{\succ}_{j'}\,s_\emptyset\,\tilde{\succ}_{j'}\,\ldots,$$

$$\mu_k\,\tilde{\succ}_k\,s_\emptyset\,\tilde{\succ}_k\,\ldots\quad\forall k\in I\backslash\{i,i_1,j,j'\}.$$

Step 2 and Step 3: We omit the arguments for Step 2 and Step 3, since they are similar to those in Case 3.1.

Case 3.2.2.2: There does not exist $s' \in S \setminus \{s_1, \mu_i, s^*\}$ *such that* $\pi_{s'}(\mu, g) \neq 0$. Since $\pi_{s^*}(\mu, g) \neq 0$ and $\pi_{\mu_i}(\mu, g) \neq 0$, there are only three schools, namely s_1, μ_i, s^*, which exhaust their capacity under μ. In this last subcase, we show that statement (2) is satisfied.

We first argue that in any plausible scenario, there is only one top priority cycle and it consists of i_1 and one student assigned to s^*. To start, since i has justified envy for s_1 at μ, there exists a cycle containing i_1 that is solved in the EDA process. Second, by non-wastefulness of EDA, we know that if a school is contained in one solved cycle, it exhausts its capacity under the final matching. Recall that only s_1, μ_i, s^* exhaust their capacity at μ. Thus, the candidate student for forming a cycle can only be assigned to s^*. Therefore, we can construct exactly one cycle with i_1 and some $l \in \mu_{s^*}$.

Now select any $(\tilde{\succ}_{-i}, \tilde{c}_{-i}, \tilde{g}_{-i}) \in \mathcal{I}(\mu, \succ_i, c_i)$. Since $g_i^{s^*} < \pi_{s^*}(\mu, g)$ and by our arguments made above, it must be true $\tilde{g}_{i_1}^{s^*} > \tilde{g}_l^{s^*} > g_i^{s^*}$ and $DA_{i_1}(\tilde{g}, (\succ_i, \tilde{\succ}_{-i})) = s^*$. However, this implies that i will be rejected by s^* under DA when she reports $\tilde{\succ}_i$. As a result, we can claim that $DA_i(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i})) = EDA_i(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i}), \tilde{c}) = \mu_i$ and statement (2) thus holds.

This completes the proof for the case in which $\tilde{S}$ is a singleton. To finish the proof, suppose now that $\tilde{S}$ contains multiple elements.

We denote the top ranked school on $\tilde{\succ}_i$ among all schools in $\tilde{S}$ by s_1. Specifically, let $\succ_i^1$ be such that $s_1 \succ_i^1 \mu_i$ and $s \succ_i^1 s'$ if $s \succ_i s'$ for all $S \setminus \{s_1\}$. Since s_1 is the only permuted school on $\succ_i^1$ compared to $\succ_i$, we can apply the arguments above (for singleton $\tilde{S}$) to $\succ_i^1$. Here, we distinguish two cases. In the first case, suppose that the observation $(\mu, \pi(\mu, g))$ is such that statement (1) holds for $\succ_i^1$. That is, we find $(\succ_{-i}^1, c_{-i}^1, g_{-i}^1) \in \mathcal{I}(\mu, \succ_i^1, c_i)$ such that $EDA_i(g^1, (\succ_i^1, \succ_{-i}^1), c^1) = s_1$. Note that all our constructions above satisfy that $DA_i(g^1, (\succ_i^1, \succ_{-i}^1)) = EDA_i(g^1, (\succ_i^1, \succ_{-i}^1), c^1) = s_1$.

Since $SU_{s_1}^{\tilde{\succ}_i} = SU_{s_1}^{\succ_i^1}$, we obtain $DA_i(g^1, (\tilde{\succ}_i, \succ_{-i}^1)) = EDA_i(g^1, (\tilde{\succ}_i, \succ_{-i}^1), c^1) = s_1$. Thus, we can conclude that statement (1) also holds for misreport $\tilde{\succ}_i$ for the first case. In the second case, suppose that the observation $(\mu, \pi(\mu, g))$ falls into the case where statement (2) holds for $\succ_i^1$. Then, we need further consider the second ranked school among $\tilde{S}$ on $\tilde{\succ}$, denoted by s_2.

Specifically, we construct $\succ_i^2$ such that $s_1 \succ_i^2 s_2 \succ_i^2 \mu_i$ and $s \succ_i^2 s'$ if $s \succ_i s'$ for all $s, s' \in S \setminus \{s_1, s_2\}$. Since we assume that $\succ_i^1$ has no influence on the result at all, we can again apply the arguments for the singleton case to $\succ_i^2$. That is, we consider whether statement (1) or statement (2) applies to $\succ_i^2$. If statement (1) holds for $\succ_i^2$, then as explained above we can conclude that statement (1) holds for $\tilde{\succ}_i$. Otherwise, we further consider the third ranked school among $\tilde{S}$ on $\tilde{\succ}$. In the following, we iteratively apply the above arguments by adding a new school from $\tilde{S}$ through each iteration. Once we arrive at a step where statement (1) holds, we stop and conclude that statement (1) holds for $\tilde{\succ}_i$. On the contrary, if for all schools in $\tilde{S}$ the observation (2) holds, then we conclude that statement (2) holds for the misreport $\tilde{\succ}_i$. □

We move to the final class of misreports in which all schools that are truly less preferred to μ_i still rank lower than μ_i. That is, in the rest of the proof, we consider $\tilde{\succ}_i \in \mathcal{P}$ such that $SU_{\mu_i}^{\tilde{\succ}_i} \subseteq SU_{\mu_i}^{\succ_i}$ and for which there exists $s, s' \in SL_{\mu_i}^{\succ_i}$ such that $s \succ_i s'$ and $s' \tilde{\succ}_i s$. Our strategy is to show that if a student could have been improved upon truth through such a misreport $\tilde{\succ}_i$ in a plausible scenario, then the misreport could also have made the misreporting student worse off in another plausible scenario.

Before we formally show the above argument, we provide three auxiliary results. The first result states that a student can improve upon μ_i via reporting $\tilde{\succ}_i$ only if μ_i is not her SOSM assignment under true preferences. Throughout the remaining discussion, we fix some $(\succ'_{-i}, c'_{-i}, g'_{-i}) \in \mathcal{I}(\mu, \succ_i, c_i)$. Also, for any $\succ'_i \in \mathcal{P}$ and any $s \in S$, we denote the weak upper contour set of $\succ'_i$ at s by $U_s^{\succ'_i} = \{s' \in S \mid s' \succeq'_i s\}$.

Lemma 2.4. *If* $EDA_i(g', (\tilde{\succ}_i, \succ'_{-i}), c') \succ_i \mu_i$, *then* $\mu_i \succ_i DA_i(g', (\succ_i, \succ'_{-i}))$.

Proof. EDA guarantees that $\mu_i \succeq_i DA_i(g', (\succ_i, \succ'_{-i}))$. We now prove the contrapositive statement: If $DA_i(g', (\succ_i, \succ'_{-i})) = \mu_i$, then $EDA_i(g', (\tilde{\succ}_i, \succ'_{-i}), c') = \mu_i$. Towards this goal, construct $\hat{\succ}_i \in \mathcal{P}$ such that

(D1) for each $s_1, s_2 \in L^{\succ_i}_{\mu_i}$, $s_1 \hat{\succ}_i s_2$ if and only if $s_1 \succ_i s_2$;

(D2) for each $s_3, s_4 \in U^{\tilde{\succ}_i}_{\mu_i}$, $s_3 \hat{\succ}_i s_4$ if and only if $s_3 \tilde{\succ}_i s_4$ and;

(D3) for all $s \in SU^{\succ_i}_{\mu_i} \setminus SU^{\tilde{\succ}_i}_{\mu_i}$, $s \in SL^{\hat{\succ}_i}_{\mu_i}$.

Since $SU^{\tilde{\succ}_i}_{\mu_i} \subseteq SU^{\succ_i}_{\mu_i}$ and $SU^{\succ_i}_{\mu_i} \cap SL^{\succ_i}_{\mu_i} = \emptyset$, one obtains $L^{\succ_i}_{\mu_i} \cap U^{\tilde{\succ}_i}_{\mu_i} = \{\mu_i\}$. Therefore, (D1) and (D2) consider distinct sets of schools, and more concretely, (D1) - (D3) defines the full order of $\hat{\succ}_i$. With (D1) we can immediately apply Proposition 2.1 such that we reach $EDA_i(g', (\hat{\succ}_i, \succ'_{-i}), c') = EDA_i(g', (\succ_i, \succ'_{-i}), c') = \mu_i$. Clearly, (D1) means that $\hat{\succ}_i$ is a monotonic transformation of $\succ_i$ at μ_i. Thus, according to Lemma 2.1 we have $DA_i(g', (\hat{\succ}_i, \succ'_{-i})) = DA_i(g', (\succ_i, \succ'_{-i}))$. Thus, we obtain $EDA_i(g', (\hat{\succ}_i, \succ'_{-i}), c') = DA_i(g', (\hat{\succ}_i, \succ'_{-i})) = \mu_i$. Moreover, (D2) and (D3) jointly ensure that $DA(g'_{-i}, (\tilde{\succ}_i, \succ'_{-i})) = DA(g'_{-i}, (\hat{\succ}_i, \succ'_{-i}))$.

Now, note that $DA_i(g', (\hat{\succ}_i, \succ'_{-i})) = EDA_i(g', (\hat{\succ}_i, \succ'_{-i}), c')$, which implies that i cannot be improved to any school more preferred than μ_i on $\hat{\succ}_i$ by EDA. Since by (D2) and (D3) we know that $\tilde{\succ}_i$ and $\hat{\succ}_i$ share the same ranking for schools more preferred than μ_i. Thus, it follows that $DA_i(g', (\tilde{\succ}_i, \succ'_{-i})) = EDA_i(g', (\tilde{\succ}_i, \succ'_{-i}), c')$. Thus, we obtain $EDA_i(g', (\tilde{\succ}_i, \succ'_{-i}), c') = \mu_i$. This completes the proof. □

Next, we show that if a student could improve upon μ_i via a misreport $\tilde{\succ}_i$, then at least one school satisfies the following three conditions: First, the student prefers her assignment to this school. Second, the relative ranking of this school is lowered under the misreport compared to truth-telling. Third, the student's score at this school is higher than this school's cutoff.

Let $S' = \{s \in SL_{\mu_i}^{\succ_i} | \exists \tilde{s} \in SL_{\mu_i}^{\succ_i} : s \succ_i \tilde{s} \text{ and } \tilde{s} \tilde{\succ}_i s\}$. Recall that we now consider misreport $\tilde{\succ}_i$ of the last class where $SU_{\mu_i}^{\tilde{\succ}_i} \subseteq SU_{\mu_i}^{\succ_i}$. According to Proposition 2.1, we know that S' must be non-empty since $EDA_i(g', (\tilde{\succ}_i, \succ'_{-i}), c') \neq EDA_i(g', (\succ_i, \succ'_{-i}), c')$.

Lemma 2.5. *If $EDA_i(g', (\tilde{\succ}_i, \succ'_{-i}), c') \succ_i \mu_i$, then there exists $s' \in S'$ such that $g_i^{s'} > \pi_{s'}(\mu, g) > 0$.*

Proof. We prove by means of contradiction. That is, given $EDA_i(g', (\tilde{\succ}_i, \succ'_{-i}), c') \succ_i \mu_i$, we suppose that for each $\hat{s} \in S'$ either $\pi_{\hat{s}}(\mu, g) > g_i^{\hat{s}}$ or $\pi_{\hat{s}}(\mu, g) = 0$. We aim at a contradiction by showing that we arrive at $EDA_i(g', (\tilde{\succ}_i, \succ'_{-i}), c') = \mu_i$.

Select any TP process with input $(g', (\succ_i, \succ'_{-i}), c')$ and denote it by $pTP^{\succ}$. Let $EDA^t(\succ)$ be the outcome of the t_{th} step in $pTP^{\succ}$. Since $EDA_i(g', (\tilde{\succ}_i, \succ'_{-i}), c') \succ_i \mu_i$, by Lemma 2.4 we have that $\mu_i \succ_i DA_i(g', (\succ_i, \succ'_{-i}))$. Then, we collect the set of schools to which i is (temporarily) assigned during $pTP^{\succ}$ in $S_i = \{\hat{s} \in S \,|\, \exists t \in \mathbb{N} : EDA_i^t(\succ) = \hat{s}\}$. As mentioned before, during the process of the TP algorithm, scores of assigned students are weakly decreasing at each school from step to step. Thus, for any $s'' \in S_i$ we have $g_i^{s''} \geq \pi_{s''}(\mu, g)$. Also, schools in S_i must have positive cutoffs. Therefore, by assumption of S', we have $S' \cap S_i = \emptyset$. Hence, we can use the following two features:

1. for any $s' \in S_i$, $SU_{s'}^{\tilde{\succ}_i} \subseteq SU_{s'}^{\succ_i}$; and
2. for any $s', s'' \in S_i$, $s' \tilde{\succ} s''$ if and only if $s' \succ_i s''$.

In the following, we first assume that $c_i = 1$. Under this assumption, we claim that with the above two features of $\succ_i$ and $\tilde{\succ}_i$, we can implement the algorithm in proof of Proposition 2.1 with profiles $(g', (\tilde{\succ}_i, \succ'_{-i}), c')$ to construct a process $pTP^{\tilde{\succ}}$ that yields the same outcome as $pTP^{\succ}$ does. Concretely, compared to the misreports studied in Proposition 2.1, the misreport $\tilde{\succ}_i$ considered here allows for additional permutations which move some $s \in L_{\mu_i}^{\succ_i}$ from above some $s'' \in S_i$ to below. Note that the first feature above ensures that all cycles solved in $pTP^{\succ}$ which do not involve i are no

different from those already covered by the algorithm in Proposition 2.1. Concretely, although agent i demands some additional schools in S', since i consents and i has a lower score than the cutoff at each of these schools, the additional demand of student i does not change the formation of cycles at each step of the algorithm. Next, note that for cycles solved in $pTP^{\succ}$ which contain i, the second feature above guarantees that such a cycle can still be solved at the corresponding step. Therefore, we arrive at $EDA_i(g', (\tilde{\succ}_i, \succ'_{-i}), c') = EDA_i(g', (\succ_i, \succ'_{-i}), c') = \mu_i$, which contradicts our initial assumption.

Next, assume $c_i = 0$. Recall that in Proposition 2.1, we extend the conclusions to the case where $c_i = 0$. Here, we can also conclude $EDA_i(g', (\tilde{\succ}_i, \succ'_{-i}), c') = EDA_i(g', (\succ_i, \succ'_{-i}), c')$ with the same line of reasoning. Again, we reach the desired contradiction. □

From now on, assume that $\mu_i \succ_i DA_i(g'_{-i}, (\succ_i, \succ'_{-i}))$. The reason for this assumption is that, as shown in Lemma 2.5, misreporting $\tilde{\succ}_i$ could potentially be profitable only if this assumption is satisfied. If reporting $\tilde{\succ}_i$ is not profitable at all, then the agent will never regret telling the truth through such a misreport. Notably, this assumption also implies that we have $\pi_{\mu_i}(\mu, g) \neq 0$ in the rest of the proofs. Moreover, Lemma 2.5 shows that there exists a maximal and non-empty set $S_1 \subseteq S'$ such that $s' \in S_1$ if and only if $g_i^{s'} > \pi_{s'}(\mu, g) > 0$. For the rest of the proof, let $s^* \in S_1$ be such that $s^* \succeq_i s'$ for any $s' \in S_1$. Furthermore, we collect in $S_2 = \{r' \in L_{\mu_i}^{\succ_i} \mid s^* \succ_i r', r' \tilde{\succ}_i s^*\}$ and denote with $r^* \in S_2$ the school such that $r^* \tilde{\succeq}_i r'$ for any $r' \in S_2$. For our construction for the last class of misreports, we rely on the following property of r^*.

Lemma 2.6. *If $EDA_i(g', (\tilde{\succ}_i, \succ'_{-i}), c') \succ_i \mu_i$, then $\pi_{r^*}(\mu, g) \neq 0$.*

Proof. We aim to show the contrapositive statement. That is, given $\pi_{r^*}(\mu, g) = 0$, we prove that $\mu_i \succeq_i EDA_i(g', (\tilde{\succ}_i, , \succ'_{-i}), c')$. Let $DA(g', (\succ_i, \succ'_{-i})) = \nu_i$. Since we assume $\pi_{\mu_i}(\mu, g) \neq 0$, it follows immediately $\pi_{\nu_i}(\mu, g) \neq 0$. That is, $\nu_i \neq r^*$. In the following, we consider two cases that are distinguished by the relative ranking of r^* and ν_i on $\tilde{\succ}_i$.

In the first case, suppose $\nu_i \tilde{\succ}_i r^*$. We claim $EDA_i(g', (\tilde{\succ}_i, \succ'_{-i}), c') = \mu_i$ here and will use similar arguments as in the proof of Lemma 2.5. Recall that we have $SU^{\tilde{\succ}_i}_{\mu_i} \subseteq SU^{\succ_i}_{\mu_i}$. Together with the assumption $\nu_i \tilde{\succ}_i r^*$ and the fact $\mu_i \succeq_i \nu_i$, we can infer $\mu_i \tilde{\succ}_i r^*$. Now, select an arbitrary TP process with input $(g', (\succ_i, \succ'_{-i}), c')$, denoted by $pTP^{\succ'}$ and let $EDA^t(\succ')$ be the outcome of the t_{th} step in $pTP^{\succ'}$. Also, let $S'_i = \{s' \in S \mid \exists t \in \mathbb{N} : EDA^t_i(\succ') = s'\}$ be the set of schools to which i is (temporarily) assigned during $pTP^{\succ'}$. As argued before, for each $s' \in S'_i$, it is true $g^{s'}_i > \pi_{s'}(\mu, g) > 0$. Note that $\nu_i \in S'_i$ and by assumption $\nu_i \tilde{\succ}_i r^*$, the selection of r^* ensures that:

1. for any $s' \in S'_i$, $SU^{\tilde{\succ}_i}_{s'} \subseteq SU^{\succ_i}_{s'}$; and

2. for any $s', s'' \in S'_i$, $s' \tilde{\succ} s''$ if and only if $s' \succ_i s''$.

Notably, as argued in Lemma 2.5, this leads to $EDA_i(g', (\tilde{\succ}_i, \succ'_{-i}), c') = \mu_i$.

In the second case, suppose $r^* \tilde{\succ}_i \nu_i$. We show $\mu_i \succ_i EDA_i(g', (\tilde{\succ}_i, \succ'_{-i}), c')$ here. Towards this goal, we first argue $SU^{\tilde{\succ}_i}_{r^*} \subseteq SU^{\succ_i}_{\nu_i}$. By contradiction, suppose that there exists $s' \in S$ such that $r' \in SU^{\tilde{\succ}_i}_{r^*}$ and $r' \notin SU^{\succ_i}_{\nu_i}$. Then, we know that (1) $\nu_i \succ_i r'$, (2) $r' \tilde{\succ}_i \nu_i$ and (3) $r' \tilde{\succ}_i r^*$. Since $g^{\nu_i}_i > \pi_{\nu_i}(\mu, g) > 0$, by (1) and (2) we can infer $\nu_i \in S_1$. Thus, the selection of s^* ensures that $s^* \succeq_i \nu_i$, which combined with (1) shows $s^* \succ_i r'$. Moreover, from (3) and the construction of S_2 we have $r' \tilde{\succ}_i r^* \tilde{\succ}_i s^*$. Note that $s^* \succeq_i \nu_i$ and $r' \tilde{\succ}_i r^* \tilde{\succ}_i s^*$, we reach a contradiction to how r^* is selected. Thus, we have $SU^{\tilde{\succ}_i}_{r^*} \subseteq SU^{\succ_i}_{\nu_i}$. Next, since by assumption r^* has vacant seat at $EDA(g', (\succ_i, \succ'_{-i}), c')$, it also has vacant seat at $DA(g', (\succ_i, \succ'_{-i}))$. With the two findings above, we can implement the arguments in Case 2 of Lemma 2.3 and conclude that $DA_i(g', (\tilde{\succ}_i, \succ'_{-i})) = r^*$ is underdemanded. Thus, student i cannot improve her assignment above r^* and we reach $EDA_i(g', (\tilde{\succ}_i, \succ'_{-i}), c') = r^*$. Since $\mu_i \succ_i r^*$, this completes the proof. □

We now show the formal arguments for the last class of misreports. Concretely, we show that when i reports $\tilde{\succ}_i$, she could have been worse off by being assigned to r^*.

Lemma 2.7. *If* $EDA_i(g', (\tilde{\succ}_i, \succ'_{-i}), c') \succ_i \mu_i$, *there exists* $(\tilde{\succ}_{-i}, \tilde{c}_{-i}, \tilde{g}_{-i}) \in \mathcal{I}(\mu, \succ_i, c_i)$ *such that* $\mu_i \succ_i EDA_i(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i}), \tilde{c}) = r^*$.

Proof. Note that by Lemma 2.6, we only need to construct such a scenario for case $\pi_{r^*}(\mu, g) > 0$. Similar as in the proof of Lemma 2.3, we go through a series of steps to show the desired result:

Step 1: We construct a candidate scenario $(\tilde{\succ}_{-i}, \tilde{c}_{-i}, \tilde{g}_{-i})$.

Step 2: We show that $(\tilde{\succ}_{-i}, \tilde{c}_{-i}, \tilde{g}_{-i}) \in \mathcal{I}(\mu, \succ_i, c_i)$.

Step 3: We argue that $EDA_i(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i}), \tilde{c}) = r^*$.

Recall that $s^* \in S_1$ is the school that ranks highest on $\succ_i$ among all schools in S_1. Let $j \in I$ be such that $\mu_j = r^*$, and let $l \in I$ be such that $\mu_l = s^*$ and $g_l^{s^*} = \pi_{s^*}(\mu, g)$. Moreover, consider the set $\bar{S} = \{s' \in SU_{s^*}^{\succ_i} \mid g_i^{s'} > \pi_{s'}(\mu, g)\}$ and denote $\bar{S} = \{s_1, s_2, \ldots, s_T\}$. Without loss of generality, let $s_1 \succ_i s_2 \succ_i \ldots \succ_i s_T$. Since $r^* \in SL_{s^*}^{\succ_i}$, we know that $r^* \notin \bar{S}$. For each $t \in \{1, \ldots, T\}$, denote the student with the lowest score assigned to s_t in μ by i_t and collect all such students in $\bar{I} = \{i_1, \ldots, i_T\}$. Similar to Lemma 2.3, we make a case distinction based on different observations. However, since we already know that $\pi_{\mu_i}(\mu, g) \neq 0$ and $\pi_{r^*}(\mu, g) \neq 0$, it suffices to consider different cardinalities of $\bar{S}$.

Case 1: $|\bar{S}| \neq 1$. *Step 1:* We start with the candidate score structure. Let $\tilde{g}_{-i}$ be such that

- $\tilde{g}_l^{\mu_i} > \tilde{g}_j^{\mu_i} > \tilde{g}_i^{\mu_i}$; and for any $k \in \mu_{\mu_i} \setminus \{i\}, g_k^{\mu_i} > g_l^{\mu_i}$;
- $g_i^{s^*} > g_l^{s^*}$; and for any $k \in \mu_{s^*} \setminus \{l\}, \tilde{g}_k^{s^*} > \tilde{g}_i^{s^*}$;
- $\tilde{g}^{s'} = g^{s'}$ for any $s' \in S \setminus \{s_1, \ldots, s_T, \mu_i, s^*\}$.

Let $i_0 = i_T$ and $s_{T+1} = s_1$. In case that $\bar{S} \neq \emptyset$, for any $s_t \in \bar{S}$:

- $\tilde{g}_{i_{t-1}}^{s_t} > \tilde{g}_i^{s_t} > \tilde{g}_{i_t}^{s_t}$; and for any $k \in \mu_{s_t} \setminus \{i_t\}, \tilde{g}_k^{s_t} > \tilde{g}_{i_{t-1}}^{s_t}$.

Next, we specify $\tilde{c}_{-i}$ such that for all $i' \in I \setminus \{i\}$ it holds that $\tilde{c}_{i'} = 1$ and consider preference profile $\tilde{\succ}_{-i} \in \mathcal{P}_{-i}$:

$$s_t \tilde{\succ}_{i_t} s_{t+1} \tilde{\succ}_{i_t} s_\emptyset \tilde{\succ}_{i_t} \ \ldots \ \ \forall t \in \{1, \ldots, T\},$$

$$s^* \tilde{\succ}_l \mu_i \tilde{\succ}_l s_\emptyset \tilde{\succ}_l \ \ldots \ ,$$

$$\mu_i \tilde{\succ}_j r^* \tilde{\succ}_j s_\emptyset \tilde{\succ}_j \ \ldots \ ,$$

$$\mu_k \tilde{\succ}_k s_\emptyset \tilde{\succ}_k \ \ldots \ \ \forall k \in I \setminus (\bar{I} \cup \{i, j, l\}).$$

Step 2: It is easily checked that $\pi(\mu, (g_i, \tilde{g}_{-i})) = \pi(\mu, g)$. Next, we show that DA leads to $DA_i(\tilde{g}, (\succ_i, \tilde{\succ}_{-i})) = s^*, DA_j(\tilde{g}, (\succ_i, \tilde{\succ}_{-i})) = r^*, DA_l(\tilde{g}, (\succ_i, \tilde{\succ}_{-i})) = \mu_i, DA_{i_t}(\tilde{g}, (\succ_i, \tilde{\succ}_{-i})) = s_{t+1}$ for each $t \in \{1, \ldots, T\}$ and $DA_k(\tilde{g}, (\succ_i, \tilde{\succ}_{-i})) = \mu_k$ for $k \in I \setminus (\bar{I} \cup \{i, j, l\})$.

Consider the corresponding application process of DA under the constructed scenario. For each student $k \in I \setminus (\bar{I} \cup \{i, j, l\})$, either μ_k is in $U_{s^*}^{\succ_i}$ and k is among the top q_{μ_k} scored students at μ_k; or μ_k is in $SL_{s^*}^{\succ_i}$ and at most q_{μ_k} students apply to μ_k according to $(\succ_i, \tilde{\succ}_{-i})$. Therefore, at the first step of the DA process, each such k applies to μ_k and is finally assigned to μ_k. Furthermore, the following students will be tentatively accepted at the first step:

- student j applies to μ_i,
- student l applies to s^*,
- for all $t \in \{1, \ldots, T\}$, student i_t applies to s_t.

At the first step of the application process, also i applies to her top choice. If i's top choice is not s_1, let $t_1 \in \mathbb{N}$ be the step in the application process, in which i applies to s_1. In all the previous steps $t < t_1$, student i is rejected at each school she proposes to. However, at step t_1 student i is tentatively accepted at s_1 and student i_1 is rejected. In fact, being initial for student i_1 being rejected at s_1, student i induces

a sequence of rejections. This sequence ends in student i being rejected at s_1 and for all $t \in \{2, ..., T\}$ student i_t is rejected from school s_t in favor of student i_{t-1} at step $t_1 + t$. Finally at step $t_1 + T$, student i_T applies to s_1 such that student i gets rejected. In the following steps, only student i makes new applications until she gets accepted. Precisely, student i proposes to each remaining school in $SU_{s^*}^{\succ_i}$ that she has not yet proposed to and gets immediately rejected at each of these schools. Finally, student i applies to s^* and gets accepted in favor of student l. Student l being rejected at s^* applies now to μ_i such that student j gets rejected. Next, j applies to r^* and gets accepted. Notice that at this step no student is rejected, the application process ends and the algorithm terminates.

Starting with the final outcome $DA_i(\tilde{g}, (\succ_i, \tilde{\succ}_{-i}))$ of the just described process, we now show that the cycle selection under a TP process ends in the observed matching μ. Since j is permanently matched in $DA(\tilde{g}, (\succ_i, \tilde{\succ}_{-i}))$ and $\tilde{c}_j = 1$, we know that $G^*(DA(\tilde{g}, (\succ_i, \tilde{\succ}_{-i})))$ contains the cycle $\gamma^1 = \{il, li\}$ and solving it yields $EDA^1(\tilde{g}, (\succ_i, \tilde{\succ}_{-i}), \tilde{c}) = \gamma^1 \circ DA(\tilde{g}, (\succ_i, \tilde{\succ}_{-i}))$, where compared to in $DA(\tilde{g}, (\succ_i, \tilde{\succ}_{-i}))$, only i and l switch their assignments.

Next, since $c_i = 1$ and i is permanently matched to μ_i in $EDA^1(\tilde{g}, (\succ_i, \tilde{\succ}_{-i}), \tilde{c})$, whenever $\bar{S}$ is non-empty, $G^*(EDA^1(\tilde{g}, (\succ_i, \tilde{\succ}_{-i}), \tilde{c}))$ contains a unique cycle

$$\gamma^2 = \{i_T i_{T-1}, i_{T-1} i_{T-2}, ..., i_{t+1} i_t, ... i_2 i_1, i_1 i_T\}$$

which once solved yields matching μ. Since all students except i and j get their top-choice, and both i, j are permanently matched, there is no cycle in $G^*(\mu)$. Therefore, $EDA(\tilde{g}, (\succ_i, \tilde{\succ}_{-i}), \tilde{c}) = \mu$.

Step 3: Reviewing the application process above, we get $DA_i(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i})) = r^*$. Moreover, note that apart from the students who are matched with school r^* at $DA(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i}))$, student j is the only one who ranks r^* above $s_\emptyset$ in $\tilde{\succ}_{-i}$. However, notice that $DA_j(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i})) = \mu_i \tilde{\succ}_j r^*$ and thus school r^* is underdemanded in $DA(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i}))$. As a result, i is permanently matched with r^* at $DA(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i}))$, which implies $EDA_i(\tilde{g}, (\tilde{\succ}_i, \tilde{\succ}_{-i}), \tilde{c}) = r^*$. This completes the proof for Case 1.

Case 2: $|\bar{S}| = 1$. *Step 1:* Let $\tilde{g}_{-i}$ be such that

- $\tilde{g}_l^{s_1} > \tilde{g}_i^{s_1} > \tilde{g}_{i_1}^{s_1}$; and for all $k \in \mu_{s_1} \setminus \{i_1\}, \tilde{g}_k^{s_1} > \tilde{g}_l^{s_1}$;
- $\tilde{g}_{i_1}^{\mu_i} > \tilde{g}_j^{\mu_i} > \tilde{g}_i^{\mu_i}$; and for all $k \in \mu_{\mu_i} \setminus \{i\}, \tilde{g}_k^{\mu_i} > \tilde{g}_{i_1}^{\mu_i}$;
- $\tilde{g}_i^{s^*} > \tilde{g}_{i_1}^{s^*} > \tilde{g}_l^{s^*}$; and for all $k \in \mu_s \setminus \{l\}, \tilde{g}_k^{s^*} > \tilde{g}_i^{s^*}$;
- $\tilde{g}^{s'} = g^{s'}$ for any $s' \in S \setminus \{s_1, \mu_i, s^*\}$.

Also, let $\tilde{c}_{-i}$ be such that for all $i' \in I \setminus \{i\}$ it holds that $\tilde{c}_{i'} = 1$ and consider preference profile $\tilde{\succ}_{-i} \in \mathcal{P}_{-i}$:

$$s_1 \,\tilde{\succ}_{i_1}\, s \,\tilde{\succ}_{i_1}\, \mu_i \,\tilde{\succ}_{i_1}\, s_\emptyset \,\tilde{\succ}_{i_1}\, \dots,$$

$$s^* \,\tilde{\succ}_l\, s_1 \,\tilde{\succ}_l\, s_\emptyset \,\tilde{\succ}_l\, \dots,$$

$$\mu_i \,\tilde{\succ}_j\, r^* \,\tilde{\succ}_j\, s_\emptyset \,\tilde{\succ}_j\, \dots,$$

$$\mu_k \,\tilde{\succ}_k\, s_\emptyset \,\tilde{\succ}_k\, \dots \quad \forall k \in I \setminus \{i, j, l, i_1\}.$$

Step 2 and Step 3: We can resemble the arguments in Step 2 and Step 3 for Case 1 to conclude that i is worse off by being finally assigned to r^* in this constructed scenario. □

Since the conclusion holds for any observation, any student and any problem, we conclude that EDA is regret-free truth-telling.

2.D Proof of Proposition 2.2

With a similar technique as in the proof of Proposition 1 in Fernandez (2020), we now show that any non-truthful report is regretted through the truth at some observation. Throughout the discussion, fix an arbitrary problem $(I, S, q, g, \succ, c)$ and fix an arbitrary $i \in I$. We divide the set of possible

misreports into three exhaustive cases. In each case, we consider an arbitrary misreport $\succ_i'$. We then construct an observation following $\succ_i'$ such that the truth $\succ_i$ would have granted i a weakly better assignment in any plausible scenario. Moreover, there exists at least one plausible scenario in which the improvement is strict.

Case 1 Suppose that for $\succ_i'$ there exists $s \in S$ such that $s_\emptyset \succ_i s$ and $s \succ_i' s_\emptyset$. Let i submit $\succ_i'$ and consider the pair $(\mu, \pi(\mu, g))$ such that $\mu_i = s$ and $g_i^{s'} < \pi_{s'}(\mu, g)$ for all $s' \in SU_s^{\succ_i'}$. At first, we show that $\mu \in \mathcal{M}|_{(\succ_i', c_i)}$ by constructing $(\tilde{\succ}_{-i}, \tilde{c}_{-i}, \tilde{g}_{-i})$ that leads to $(\mu, \pi(\mu, g))$: That is, we show that $(\mu, \pi(\mu, g))$ is an observation under EDA. Let $\tilde{g}_{-i}$ be such that, for each $s' \in SU_{\mu_i}^{\succ_i'}$, each student in $\mu_{s'}$ is among the top $q_{s'}$'s scored students at school s'. Let i rank highest on $\tilde{g}^s$ and suppose that the remaining scores are arbitrary. Let $\tilde{\succ}_{-i}$ be such that for each $j \in I \setminus \{i\}$, $\tilde{\succ}_j$ only ranks μ_j as acceptable and suppose that $\tilde{c} = c$. Apparently, we have $\pi(\mu, (g_i, \tilde{g}_{-i})) = \pi(\mu, g)$ and $EDA(\tilde{g}, (\succ_i', \tilde{\succ}_{-i}), \tilde{c}) = \mu$. Thus, $\mu \in \mathcal{M}|_{(\succ_i', c_i)}$.

It remains to be shown that student i regrets $\succ_i'$ through the truth $\succ_i$. Note that since EDA is individually rational, it holds that $EDA_i(\tilde{g}, (\succ_i, \tilde{\succ}_{-i}), \tilde{c}) \succeq_i s_\emptyset$ for any $(\tilde{\succ}_{-i}, \tilde{c}_{-i}, \tilde{g}_{-i}) \in \mathcal{I}(\mu, \succ_i, c_i)$. Since $s_\emptyset \succ_i s$, student i thus regrets $\succ_i'$ through the truth at $(\mu, \pi(\mu, g))$.

Case 2 Let for $\succ_i'$ exist $s \in S$ such that $s_\emptyset \succ_i' s$ and $s \succ_i s_\emptyset$. Suppose i submits $\succ_i'$ and consider $(\mu, \pi(\mu, g))$ such that $\mu_i = s_\emptyset$, $\pi_s(\mu, g) = 0$ and $g_i^{s'} < \pi_{s'}(\mu, g)$ for all $s' \in SU_{s_\emptyset}^{\succ_i'}$. Notably, by doing the same construction $(\tilde{\succ}_{-i}, \tilde{c}_{-i}, \tilde{g}_{-i})$ as in Case 1, we can infer $\mu \in \mathcal{M}|_{(\succ_i', c_i)}$.

It remains to be shown that student i regrets $\succ_i'$ through the truth $\succ_i$. To see this, note that since EDA is non-wasteful, it holds that $EDA_i(\tilde{g}, (\succ_i, \tilde{\succ}_{-i}), \tilde{c}) = s$ for any $(\tilde{\succ}_{-i}, \tilde{c}_{-i}, \tilde{g}_{-i}) \in \mathcal{I}(\mu, \succ_i, c_i)$. Since $s \succ_i s_\emptyset$, student i thus regrets $\succ_i'$ through the truth at $(\mu, \pi(\mu, g))$.

Case 3 In this last case consider $\succ_i'$ which only contains permutations in the acceptable and unacceptable set, i.e., $A_i(\succ_i') = A_i(\succ_i)$ and $U_i(\succ_i') = U_i(\succ_i)$.

The following labeling for any $\succ_i'' \in \mathcal{P}$ in the acceptable set $A_i(\succ_i'')$ ensures that a school's index corresponds to its position in $\succ_i''$. Precisely, we denote s_1'' as the $\succ_i''$-maximal element on $A_{i,1}(\succ_i'') = A_i(\succ_i'')$ and s_2'' as the $\succ_i''$-maximal element on $A_{i,2}(\succ_i'') = A_{i,1}(\succ_i'') \setminus \{s_1''\}$, and so forth.

Now suppose that $|A_i(\succ_i)| = N \in \mathbb{N}$ is the number of acceptable schools under true preferences of student i and consider a permutation $\succ_i'$ as described above. Since $\succ_i'$ is a permutation, there exists $n^* = \arg\min_n \{n \leq N \,|\, s_n' \neq s_n\}$.

Next, let student i observe $(\mu, \pi(\mu, g))$ such that $\mu_i = s_{n^*}'$, $\pi_{s_{n^*}}(\mu, g) = 0$ and $g_i^{s'} < \pi_{s'}(\mu, g)$ for all $s' \in UC_{s_{n^*}'}^{\succ_i'}$. Again, by doing the same construction $(\tilde{\succ}_{-i}, \tilde{c}_{-i}, \tilde{g}_{-i})$ as in Case 1, we can infer $\mu \in \mathcal{M}|_{(\succ_i', c_i)}$.

It remains to be shown that student i regrets $\succ_i'$ through the truth $\succ_i$. Since s_{n^*} has capacity left, this allows us to conclude that if i would have reported $\succ_i$ then, for any $(\tilde{\succ}_{-i}, \tilde{c}_{-i}, \tilde{g}_{-i}) \in \mathcal{I}(\mu, \succ_i, c_i)$, student i would have been matched to s_{n^*}. Since $s_{n^*} \succ_i s_{n^*}'$, we conclude that i regrets $\succ_i'$ through $\succ_i$ at $(\mu, \pi(\mu, g))$. This completes the proof. □

3

Optimal Sequential Implementation*

We introduce an optimality notion for sequential implementations of matching rules in priority-based matching problems. An optimal sequential implementation (1) complies with obvious dominance (Li, 2017) whenever possible and (2) does not elicit more information about preferences from agents than necessary to determine the outcome. We show that optimal sequential implementations of a strategy-proof rule are *obviously strategy-proof (OSP)* when that rule is OSP-implementable. As a promising solution in providing incentives for truth-telling, we derive an optimal implementation

*This chapter is based on Chen and Westkamp (2021). For their insightful discussions and suggestions, we are grateful to Christoph Schottmüller, Markus Möller and Marius Gramb. This work has also benefited from the comments made by the seminar participants at the University of Cologne and the University of Bonn.

of Top Trading Cycles (TTC) that exists even when OSP-implementations of TTC are unavailable. We further introduce a weaker notion of optimality that imposes no restriction on the amount of information elicited through decisions in which truthful revelations are obviously dominant. Finally, we introduce a weakly optimal implementation of Deferred Acceptance which exists generally.

3.1 Introduction

Many public school choice procedures and entry-level labor markets use a central clearinghouse to match agents to resources. To achieve desirable outcomes in the presence of strategic agents, clearinghouses are often based on *strategy-proof* matching rules, in which truthful behaviors are weakly dominant for agents. Unfortunately, although advised to report their true preferences, agents are routinely observed to misreport both in experiments (Chen and Sönmez, 2002, 2006; Pais and Pintér, 2008) and in the field (Hassidim et al., 2016; Shorrer and Sóvágó, 2018; Rees-Jones, 2018) in *static* implementations of strategy-proof matching rules.[1]

To make the incentives provided by strategy-proof rules more apparent, Li (2017) proposes implementations via sequential revelation games that are *obviously strategy-proof (OSP)*. Loosely speaking, a game is OSP if acting truthfully is an *obviously dominant strategy* for each agent and at each stage she plays in that game, and a strategy is obviously dominant if based on the information elicited so far, the worst-case outcome under that strategy is weakly better than the best-case outcome under any deviation. Li (2017) shows that the strategic simplicity of OSP-implementations can be recognized by agents even without contingent reasoning, and such implementations lead to more truthful preference revelations in the lab.

Despite their appealing incentive properties, OSP-implementations of popular strategy-proof rules,

[1] An implementation of a rule implements that rule under truthful behaviors: For any preference profile, when agents act truthfully, it always results in the unique outcome consistent with the allocation that rule assigns to that profile.

such as Top Trading Cycles (TTC) and Deferred Acceptance (DA), only exist when priorities satisfy stringent acyclicity conditions that rarely hold in practice (Troyan, 2019; Thomas, 2021). Notably, while a number of papers study how to achieve OSP in various frameworks, few of them ask what is the "next best thing" when it is impossible to implement a strategy-proof rule in obviously dominant strategies. This chapter aims at an answer to this question.

To this end, we employ a standard priority-based allocation problem without monetary transfers and use obvious dominance as a guiding principle to develop an optimality notion of implementations that exists even when OSP-implementations are absent. In particular, a sequential implementation of a matching rule is *optimal* if it satisfies the following two conditions: First, whenever it is obviously dominant for an agent to truthfully reveal certain information about her preferences, such decision is prioritized in picking over decisions in which truthful revelations are not obviously dominant. Second, it elicits no more than the minimal amount of information necessary to unambiguously determine the outcome under that rule. Intuitively, one can break down the revelation of preferences into a set of small decisions. Each time an agent reveals some information about her preferences, she essentially makes a subset of those small decisions.[2] From the perspective of obvious strategy-proofness, it is natural to expect that an agent may make mistakes in each such small decision once it is not obviously dominant. Therefore, our notion of optimality reaches the extreme in the sense of avoiding mistakes.

We show that whenever a rule can be implemented in obviously dominant strategies, any optimal implementation of that rule is also OSP (Proposition 3.2). Moreover, in this case, our concept refers to the set of OSP-implementations in which revelations are efficient. On the contrary, when OSP-implementations are impossible, our concept provides a solution that incentivizes agents to behave truthfully from the perspective of minimizing mistakes from cognitively limited agents. Consider that

[2] For instance, for any agent, each such small decision can be a revelation of how she ranks two given objects. Accordingly, if an agent reveals that an object is her top choice, she essentially makes a set of small decisions each of which reveals that she prefers that object to some other object.

an agent reports an object to be her top choice through a preference revelation. Then, that preference revelation is obviously dominant for that agent if and only if she will secure that object through the underlying revelation (Proposition 3.1).[3]

The first main result of this chapter shows how to construct an optimal implementation of TTC. Notably, our proposed implementation exists in all problems under consideration (Theorem 3.1). It turns out that our two conditions of optimality might sometimes be incompatible in DA implementations, and this is essentially caused by the tentative nature of acceptances in DA. Reacting to such an incompatibility, we develop a weaker optimality notion which loosens the second condition by only minimizing the amount of information elicited through non-obviously dominant revelations. It is worth noting that Proposition 3.2 still holds with this weaker notion. At last, we classify the objects which an agent can secure during DA implementations (Proposition 3.3) and introduce a weakly optimal implementation for DA that exists generally (Conjecture 3.1). Our proposals for TTC and DA contribute to the design of practical implementations of these rules that aims at maximizing the rate of truthful preference revelations.

Related Literature This chapter belongs to the growing line of research that relates to Li (2017)'s obvious strategy-proofness. Pycia and Troyan (2021) introduce a family of simplicity standards and characterize simple mechanisms in broad domains under their *richness* assumption. In terms of obvious strategy-proofness, they introduce *millipede games* that characterize general OSP mechanisms without monetary transfers. Troyan (2019) focuses on TTC and characterizes all priorities that ensure TTC to be implementable in obviously dominant strategies. Ashlagi and Gonczarowski (2018) show that stable mechanisms (including DA) are only OSP-implementable in very restrictive environments, and Thomas (2021) follows their study characterizing all such environments that make DA OSP-implementable. Bade and Gonczarowski (2016) study OSP-implementations of Pareto-efficient

[3] Specifically, this result relates closely to Pycia and Troyan (2021)'s characterization about OSP mechanisms without transfers. A detailed discussion will be presented in the main text.

social choice rules in various domains. The present chapter complements prior works on OSP mechanisms since we provide solutions to problems where OSP mechanisms are absent.[4]

Our work also relates to the broader literature that considers sequential implementations of matching rules. From the theoretical perspective, the closest paper to ours in this class is Bó and Hakimov (2020b), who propose the family of *pick-an-object (PAO)* mechanisms, in which agents are asked to choose their favorites from individualized menus of available objects. They show that both DA and TTC can be implemented via PAO mechanisms in robust truthful Bayesian equilibria. Notably, while our concept of optimal implementations has the same equilibrium property as PAO mechanisms, our focus is different since we aim at finding the most obvious implementations. Bó and Hakimov (2020a) and Klijn et al. (2019) conduct experimental research, suggesting that compared to static implementations of strategy-proof rules, sequential counterparts result in higher rates of truth-telling. Moreover, there are studies that evaluate various practical applications of sequential matching rules (Gong and Liang, 2016; Chen and Kesten, 2017; Dur et al., 2018; Bo and Hakimov, 2019; Haeringer and Iehlé, 2019; Grenet et al., 2019).

Our paper also contributes to the literature exploring information efficiency in matching. Immorlica et al. (2020) introduce *regret-free stability* which requires both a stable outcome and efficient information revelations. They show that no mechanism is regret-free stable and introduce a mechanism that yields approximately regret-free stable outcomes. Similar to their work, we also set a model where information communication is not costly. However, while they concentrate on a specific stochastic model of preferences, we consider a setting that is more general. Other papers study information efficiency in matching problems where information communication is costly (Nisan and Segal, 2006; Gonczarowski et al., 2019; Ashlagi et al., 2020).

Besides those already discussed, there is an increasing number of studies proposing innovative con-

[4]See also Zhang and Levin (2017), Mackenzie (2020) and Troyan and Morrill (2020) that introduce new concepts that closely relate to obvious strategy-proofness.

cepts that apply to sequential mechanisms. Börgers and Li (2019) introduce a notion of simplicity on mechanisms which guarantees that agents can recognize the optimal strategies with their first-order beliefs on others' preferences. The key difference to our work is that our setting does not take beliefs into consideration. Furthermore, Akbarpour and Li (2020) classify credible mechanisms to which authorities have commitment power. Hakimov and Raghavan (2020) investigate the transparency of a mechanism which requires any deviation from that mechanism to be detected by agents. Both credibility and transparency indicate the benefits of sequential mechanisms in views different from our incentive considerations.

The remainder of this chapter is organized as follows. Section 3.2 presents the basic model. Section 3.3 formally defines an optimal sequential implementation. While Section 3.4 provides the application to TTC, Section 3.5 defines the weaker version of optimality and presents our results on DA. Section 3.6 concludes.

3.2 Model

We consider a standard priority-based allocation problem without monetary transfers. Formally, a priority-based allocation problem without transfers, *problem* from now on, is described by a quadruple $(I, O, \rhd_O, \succeq)$, where

- I is a finite set of agents,
- O is a finite set of indivisible objects,
- $\rhd_O = (\rhd_o)_{o \in O}$ is a *priority structure*, where for each $o \in O$, $\rhd_o$ is a strict priority ordering of $I \cup \{o\}$, and
- $\succeq = (\succeq_i)_{i \in I}$ is a *preference profile*, where for each $i \in I$, $\succeq_i$ is a strict preference ranking of $O \cup \{i\}$.

For the remainder of this chapter, we fix the set of agents I, the set of objects O and the priority structure $\rhd_O$. Hence, problems are parameterized by agents' preferences. For agent $i \in I$, let $\mathcal{P}_i$ denote the set of possible preferences, and let $\mathcal{P} = (\mathcal{P}_i)_{i\in I}$ denote the set of all possible preference profiles. Given some preference profile $\succeq \in \mathcal{P}$ and some $I' \subseteq I$, let $\succeq_{I'} = (\succeq_i)_{i\in I'}$ denote the preferences of agents in I', and let $-I'$ denote the set $I \setminus I'$. Moreover, given some preference relation $\succeq_i \in \mathcal{P}_i$, some subset $O' \subseteq O \cup \{i\}$, and some integer k, let $top_k(\succeq_i |_{O'}) \in O \cup \{i\}$ denote the k_{th} most preferred option among O' according to $\succeq_i$.

A *matching* is a function $\mu : I \to O \cup I$ such that $|\mu^{-1}(o)| \leq 1$ for each $o \in O$ and $\mu(i) \in O \cup \{i\}$ for each $i \in I$. Denote the set of all matchings by $\mathcal{M}$. A *rule* maps preference profiles into matchings, i.e., a matching rule is a mapping $f : \mathcal{P} \to \mathcal{M}$. Given some $\succeq \in \mathcal{P}$ and some $i \in I$, $f(\succeq)$ is the matching chosen by f for $\succeq$ and $f_i(\succeq)$ is i's match ($\in O \cup \{i\}$).

A matching rule f is said to be *strategy-proof* if $f_i(\succeq_i, \succeq_{-i}) \succeq_i f_i(\succeq'_i, \succeq_{-i})$ for all $i \in I$, all $\succeq_i, \succeq'_i \in \mathcal{P}_i$ and all $\succeq_{-i} \in \mathcal{P}_{-i}$. In words, a matching rule f is strategy-proof when submitting the true preferences is a weakly dominant strategy for all agents.

3.2.1 Preliminaries

In this subsection, we introduce key concepts for the development of optimal sequential implementations. First, we formally introduce "decisions" for the agents and distinguish between those that are obvious and those that are not. Fix any $\tilde{\mathcal{P}} \subseteq \mathcal{P}$.

Definition 3.1. 1. A *decision by $i \in I$ at $\tilde{\mathcal{P}}$* is a non-trivial partition $\tilde{\mathbb{P}}_i$ of $\tilde{\mathcal{P}}_i$.[5]

2. Decision $\tilde{\mathbb{P}}_i$ is *obvious for i* at $\tilde{\mathcal{P}}$ if, for any pair $\succeq_i, \succeq'_i \in \tilde{\mathcal{P}}_i$ such that $\succeq_i$ and $\succeq'_i$ belong to

[5] That is, $\tilde{\mathbb{P}}_i = (\tilde{P}_i^1, \ldots, \tilde{P}_i^M)$ is a disjoint partition of $\tilde{\mathcal{P}}_i$ such that $\tilde{P}_i^m \neq \emptyset$ for all $m \leq M$ and $M \geq 2$.

different elements of $\tilde{\mathbb{P}}_i$, we have that

$$\min_{\succeq_{-i} \in \tilde{\mathcal{P}}_{-i}} f_i(\succeq_i, \succeq_{-i}) \succeq_i \max_{\succeq_{-i} \in \tilde{\mathcal{P}}_{-i}} f_i(\succeq_i', \succeq_{-i}), \tag{3.1}$$

where min and max refer to the worst and best possible outcome with respect to $\succeq_i$, respectively.

Asking agent $i \in I$ to take a decision at $\tilde{\mathcal{P}}$ in the sense of Definition 3.1 means asking her to reveal additional information about her preferences. The notion of an obvious decision $\tilde{\mathbb{P}}_i$ is inspired by the notion of obvious strategy-proofness due to Li (2017). The requirement here is that if i's true preferences are given by $\succeq_i \in \tilde{\mathcal{P}}_i$, then the worst thing that could happen to i when she truthfully chooses the element of $\tilde{\mathbb{P}}_i$ that contains $\succeq_i$ is no worse than the best thing that could happen to her if she chooses any other element of $\tilde{\mathbb{P}}_i$.

Finally, we introduce some further notation that will be used in our construction of optimal sequential implementations in the next section and that will also enable us to provide a general characterization of obvious decisions. Given some agent $i \in I$ and a matching rule f, let

$$O_i(\tilde{\mathcal{P}}, f) = \{o \in O \cup \{i\} : f_i(\succeq) = o \text{ for some } \succeq \in \tilde{\mathcal{P}}\}.$$

Let

$$S_i(\tilde{\mathcal{P}}, f) = \{o \in O : f_i(\succeq) = o \text{ for all } \succeq \in \tilde{\mathcal{P}} \text{ such that } top(\succeq_i |_{O_i(\tilde{\mathcal{P}}, f)}) = o\}^6$$

be the (possibly) empty set of objects that i can secure by ranking them first among objects in $O_i(\tilde{\mathcal{P}}, f)$. We call each object in $S_i(\tilde{\mathcal{P}}, f)$ a *secure object* for i at $\tilde{\mathcal{P}}$. Clearly, we have that $S_i(\tilde{\mathcal{P}}, f) \subseteq O_i(\tilde{\mathcal{P}}, f)$. For any $o \in O_i(\tilde{\mathcal{P}}, f)$, let $\tilde{\mathcal{P}}_i^o$ be the set of all possible preference relations for i in $\tilde{\mathcal{P}}_i$ that have o as their top

[6] In the following discussion, we omit the subscript in $top_1(\succeq_i |_{O_i(\tilde{\mathcal{P}}, f)}) = o$ when we refer to the object which ranks first among $O_i(\tilde{\mathcal{P}}, f)$ on $\succeq_i$.

choice among $O_i(\tilde{\mathcal{P}}, f)$, i.e.

$$\tilde{\mathcal{P}}_i^o = \{\succeq_i \in \tilde{\mathcal{P}}_i : top(\succeq_i |_{O_i(\tilde{\mathcal{P}},f)}) = o\}.$$

For any $O' \subseteq O_i(\tilde{\mathcal{P}}, f)$, let $\tilde{\mathcal{P}}_i^{O'} = \cup_{o \in O'} \tilde{\mathcal{P}}_i^o$. We are now ready to state and prove a characterization of obvious decisions in our setting. Specifically, this result is closely related to the characterization of OSP mechanisms without transfers in Pycia and Troyan (2021).

Proposition 3.1. *Let f be a strategy-proof matching rule, $i \in I$ be arbitrary, and $\tilde{\mathcal{P}}_i$ be such that, for all $o \in O_i(\tilde{\mathcal{P}}, f)$ where $|O_i(\tilde{\mathcal{P}}, f)| \geq 2$, if there exists $\succeq_i \in \tilde{\mathcal{P}}_i$ for which $top(\succeq_i |_{O_i(\tilde{\mathcal{P}},f)}) = o$, then there exists $\succeq_i' \in \tilde{\mathcal{P}}_i$ for which $top_2(\succeq_i' |_{O_i(\tilde{\mathcal{P}},f)}) = o$. The decision $\mathbb{P}_i = (\tilde{\mathcal{P}}_i^t)_{t \subseteq O_i(\tilde{\mathcal{P}},f)}$ is obvious for i at $\tilde{\mathcal{P}}$ if and only if there exists a t such that $\tilde{\mathcal{P}}_i^{O_i(\tilde{\mathcal{P}},f) \setminus S_i(\tilde{\mathcal{P}},f)} \subseteq \tilde{\mathcal{P}}_i^t$.*

Proof. Let $S_i(\tilde{\mathcal{P}}, f) = \{o_1, \ldots, o_M\}$ for some $M \geq 0$ (where, obviously, $S_i(\tilde{\mathcal{P}}, f) = \emptyset$ if $M = 0$).

We argue first that $\mathbb{P}_i^* = (\tilde{\mathcal{P}}_i^{o_1}, \ldots, \tilde{\mathcal{P}}_i^{o_M}, \tilde{\mathcal{P}}_i^{O_i(\tilde{\mathcal{P}},f) \setminus S_i(\tilde{\mathcal{P}},f)})$ is an obvious decision for i. Let $\succeq_i \in \tilde{\mathcal{P}}_i$ be arbitrary. We distinguish two cases:

- **Case 1: $\mathbf{top}(\succeq_\mathbf{i} |_{\mathbf{O_i}(\tilde{\mathcal{P}},\mathbf{f})}) = \mathbf{o_m}$ for some $\mathbf{m} \geq 1$.**

  By definition of $S_i(\tilde{\mathcal{P}}, f)$, we have that $f_i(\succeq_i, \succeq_{-i}) = o_m$ for all $\succeq_{-i} \in \tilde{\mathcal{P}}_{-i}$. Hence, Eq. 3.1 is satisfied.

- **Case 2: $\mathbf{top}(\succeq_\mathbf{i} |_{\mathbf{O_i}(\tilde{\mathcal{P}},\mathbf{f})}) \in \mathbf{O_i}(\tilde{\mathcal{P}}, \mathbf{f}) \setminus \mathbf{S_i}(\tilde{\mathcal{P}}, \mathbf{f})$.**

  Given the structure of the decision $\mathbb{P}_i^*$, any possible deviation for i is a preference relation $\succeq_i'$ such that $top(\succeq_i' |_{O_i(\tilde{\mathcal{P}},f)}) = o_m$ for some $m \geq 1$. By definition of $S_i(\tilde{\mathcal{P}}, f)$, we have that $f_i(\succeq_i', \succeq_{-i}) = o_m$ for all $\succeq_{-i} \in \tilde{\mathcal{P}}_{-i}$.

  Hence, if Proposition 3.1 were not true, there would exist $\succeq_{-i}' \in \tilde{\mathcal{P}}_{-i}$ which satisfies that $o_m \succ_i f_i(\succeq_i, \succeq_{-i}')$. Now fix some $\succeq_i' \in \tilde{\mathcal{P}}_i^{o_m}$. By construction, we have that $o_m = f_i(\succeq_i', \succeq_{-i}')$. But then, f cannot be strategy-proof. This contradiction completes the proof in Case 2.

Next, we argue that if $\tilde{\mathbb{P}}_i = (\tilde{\mathcal{P}}_i^t)_t$ is obvious, then there exists a t such that $\tilde{\mathcal{P}}_i^{O_i(\tilde{\mathcal{P}},f)\setminus S_i(\tilde{\mathcal{P}},f)} \subseteq \tilde{\mathcal{P}}_i^t$. Let $\succeq_i, \succeq_i' \in \tilde{\mathcal{P}}_i^{O_i(\tilde{\mathcal{P}},f)\setminus S_i(\tilde{\mathcal{P}},f)}$ be arbitrary. We argue that these two preference relations have to belong to the same element of $\tilde{\mathbb{P}}_i$ in order for the decision to be obvious for i. We again distinguish two cases.

- **Case 1: $\mathbf{top}(\succeq_\mathbf{i}|_{\mathbf{O_i}(\tilde{\mathcal{P}},\mathbf{f})}) = \mathbf{top}(\succeq_\mathbf{i}'|_{\mathbf{O_i}(\tilde{\mathcal{P}},\mathbf{f})})$.**

  Let $o = top(\succeq_i|_{O_i(\tilde{\mathcal{P}},f)})$. Since $o \in O_i(\tilde{\mathcal{P}},f) \setminus S_i(\tilde{\mathcal{P}},f)$, there exists $\succeq_{-i}^1 \in \tilde{\mathcal{P}}_{-i}$ such that $f_i(\succeq_i, \succeq_{-i}^1) \neq o$.

  We claim that there is some $\succeq_{-i}^2 \in \tilde{\mathcal{P}}_{-i}$ such that $f_i(\succeq_i', \succeq_{-i}^2) = o$: By the assumption that $o \in O_i(\tilde{\mathcal{P}},f)$, there exist $\succeq^2 \in \tilde{\mathcal{P}}$ such that $f_i(\succeq^2) = o$; by strategy-proofness of f, we must have $o = f_i(\succeq_i', \succeq_{-i}^2)$.

  Hence, $\tilde{\mathbb{P}}_i$ cannot be obvious if $\succeq_i$ and $\succeq_i'$ belong to different elements of the partition.

- **Case 2: $\mathbf{top}(\succeq_\mathbf{i}|_{\mathbf{O_i}(\tilde{\mathcal{P}},\mathbf{f})}) \neq \mathbf{top}(\succeq_\mathbf{i}'|_{\mathbf{O_i}(\tilde{\mathcal{P}},\mathbf{f})})$.**

  Let $o = top(\succeq_i|_{O_i(\tilde{\mathcal{P}},f)})$ and $p = top(\succeq_i'|_{O_i(\tilde{\mathcal{P}},f)})$. By our assumption about $\tilde{\mathcal{P}}_i$, there exists a preference relation $\succeq_i''$ such that $o \succ_i'' p \succ_i'' q$ for all $q \in O_i(\tilde{\mathcal{P}},f)$. By our arguments in Case 1, there must exist a t such that $\{\succeq_i, \succeq_i''\} \subseteq \tilde{\mathcal{P}}_i^t$.

  Assume first that there exists some $\succeq_{-i}^1 \in \tilde{\mathcal{P}}_{-i}$ such that $f_i(\succeq_i'', \succeq_{-i}^1) = p$ is true. Since $p \in O_i(\tilde{\mathcal{P}},f) \setminus S_i(\tilde{\mathcal{P}},f)$, there exists $\succeq_{-i}^2 \in \tilde{\mathcal{P}}_{-i}$ such that $f_i(\succeq_i', \succeq_{-i}^2) \neq p$. Hence, in order for $\tilde{\mathbb{P}}_i$ to be obvious for i, we must have $\succeq_i' \in \tilde{\mathcal{P}}_i^t$.

  Next, assume $f_i(\succeq_i'', \succeq_{-i}) \neq p$ for all $\succeq_{-i} \in \tilde{\mathcal{P}}_{-i}$. Since $o \in O_i(\tilde{\mathcal{P}},f) \setminus S_i(\tilde{\mathcal{P}},f)$, there exists $\succeq_{-i}^1 \in \tilde{\mathcal{P}}_{-i}$ such that $f_i(\succeq_i'', \succeq_{-i}^1) \neq o$. In the case which we consider here, we must have that $f_i(\succeq_i'', \succeq_{-i}^1) \neq p$ as well. Since $p \in O_i(\tilde{\mathcal{P}},f)$, there is $\succeq_{-i}^2 \in \tilde{\mathcal{P}}_{-i}$ such that $f_i(\succeq_i', \succeq_{-i}^2) = p$. Hence, in order for $\tilde{\mathbb{P}}_i$ to be obvious for i, we must again have $\succeq_i' \in \tilde{\mathcal{P}}_i^t$.

□

Notably, the above proof allows us to conclude the following result that turns out to be useful in the following sections.

Corollary 3.1. *Let f be a strategy-proof matching rule and let $i \in I$, $o \in O$ and $\tilde{\mathcal{P}}$ be arbitrary. If $o \in S_i(\tilde{\mathcal{P}}, f)$, then for any $\hat{\mathcal{P}} \subseteq \tilde{\mathcal{P}}$ such that $top(\succeq_i |_{O_i(\hat{\mathcal{P}}, f)}) = o$ for some $\succeq_i \in \hat{\mathcal{P}}_i$, it holds that $o \in S_i(\hat{\mathcal{P}}, f)$.*

Pycia and Troyan (2021) characterize that any OSP mechanism without transfers can be interpreted as a "millipede game" where each obvious decision comprises several "clinching" options and a "passing" option. Concretely, a "clinching" option ensures the player to get certain outcomes, and the "passing" option allows the player to wait for better outcomes while keeping all outcomes in the "clinching" option still open for that player. The decision $\mathbb{P}_i^* = (\tilde{\mathcal{P}}_i^{o_1}, \ldots, \tilde{\mathcal{P}}_i^{o_M}, \tilde{\mathcal{P}}_i^{O_i(\tilde{\mathcal{P}}, f) \setminus S_i(\tilde{\mathcal{P}}, f)})$ discussed in the above proof provides the same guarantees. That is, each of $\{\tilde{\mathcal{P}}_i^{o_1}, \ldots, \tilde{\mathcal{P}}_i^{o_M}\}$ can be regarded as a "clinching" option and $\tilde{\mathcal{P}}_i^{O_i(\tilde{\mathcal{P}}, f) \setminus S_i(\tilde{\mathcal{P}}, f)}$ can be recognized as the "passing" option. As shown in the proof, if i chooses any of $\{\tilde{\mathcal{P}}_i^{o_1}, \ldots, \tilde{\mathcal{P}}_i^{o_M}\}$, she is assigned to her top choice for sure. Moreover, according to Corollary 3.1, if i chooses $\tilde{\mathcal{P}}_i^{O_i(\tilde{\mathcal{P}}, f) \setminus S_i(\tilde{\mathcal{P}}, f)}$, she can get back to objects in $S_i(\mathcal{P}, f)$ anytime when the more preferred objects in $O_i(\tilde{\mathcal{P}}, f) \setminus S_i(\tilde{\mathcal{P}}, f)$ are no longer available.

An immediate implication of Proposition 3.1 is that for each agent, the existence of secure objects is necessary and sufficient for the existence of obvious decisions.

Corollary 3.2. *Let f be a strategy-proof matching rule, let $i \in I$ be arbitrary, and $\tilde{\mathcal{P}}_i$ be such that, for all $o \in O_i(\tilde{\mathcal{P}}, f)$ where $|O_i(\tilde{\mathcal{P}}, f)| \geq 2$, if there exists $\succeq_i \in \tilde{\mathcal{P}}_i$ for which $top(\succeq_i |_{O_i(\tilde{\mathcal{P}}, f)}) = o$, then there exists $\succeq_i' \in \tilde{\mathcal{P}}_i$ for which $top_2(\succeq_i' |_{O_i(\tilde{\mathcal{P}}, f)}) = o$. There is an obvious decision for i at $\tilde{\mathcal{P}}$ if and only if $S_i(\tilde{\mathcal{P}}, f) \neq \emptyset$.*

3.3 Optimal sequential implementation

In this section, we develop a notion of optimality for extensive-form implementations of a given rule. We start by formally introducing the types of sequential revelation games that we consider in this chapter. The definition is identical to that in Ashlagi and Gonczarowski (2018), which in turn provided a more concise reformulation of the general extensive-form revelation games considered by Li (2017) that applies to matching markets.

Definition 3.2. An *extensive-form revelation game with perfect information* Γ is a quadruple (R, τ, π, φ), where

1. R is a rooted game tree
   (a) r is the root node of R
   (b) $N(R)$ is the set of non-terminal nodes of R, where $r \in N(R)$
   (c) $L(R)$ is the set of terminal nodes of R
   (d) $E(R) = \{E(n)\}_{n \in N(R)}$ is the set of edges of R, where $E(n)$ is the set of edges originating from non-terminal node n; given any edge $e \in E(R)$, we denote the origin node of e by $n(e)$
2. $\tau : L(R) \to \mathcal{M}$ maps terminal nodes to matchings
3. $\pi : N(R) \to I$ maps non-terminal nodes to agents
4. $\varphi = (\varphi_n)_{n \in N(R)}$, where for each $n \in N(R)$, $\varphi_n : E(n) \to 2^{\mathcal{P}_{\pi(n)}} \setminus \{\emptyset\}$ maps edges to sets of preference relations for agent $\pi(n)$ such that:
   (a) for any two distinct $e, e' \in E(n)$, $\varphi_n(e) \cap \varphi_n(e') = \emptyset$, and

(b) if e^* is the first edge along the path from n back to the root r such that $\pi(n(e^*)) = \pi(n)$, then $\cup_{e\in E(n)}\varphi(e) = \varphi(e^*)$; if no such edge exists, then $\cup_{e\in E(n)}\varphi(e) = \mathcal{P}_{\pi(n)}$.

For the following discussion, fix an extensive-form revelation game $\Gamma = (R, \tau, \pi, \varphi)$. For each $n \in N(R) \cup L(R)$, let $\mathcal{P}^n(\Gamma) \subseteq \mathcal{P}$ be the set of remaining preference profiles upon reaching n, i.e. for each $i \in I, \mathcal{P}_i^n(\Gamma) = \varphi(e_i)$ where e_i is the most recent edge along the path back from n to r such that $\pi(n(e_i)) = i$ (and $\mathcal{P}_i^n(\Gamma) = \mathcal{P}_i$ if no such edge exists). For each $n \in N(R)$, let $\mathbb{P}^n(\Gamma) = (\varphi_n(e))_{e\in E(n)}$ be the decision (in the sense of Definition 3.1) facing agent $\pi(n)$, and let Γ_n be the sub-game of Γ such that n is the root node of Γ_n. Note that Definition 3.2 implies $\mathbb{P}^\Gamma \equiv (\mathcal{P}^l(\Gamma))_{l\in L(R)}$ is a partition of $\mathcal{P}$. The next definition introduces a key property of Γ and $\mathbb{P}^\Gamma$.

Definition 3.3. An extensive-form revelation game Γ is *an implementation of* f if, for all $l \in L(R)$ and all pairs $\succeq, \succeq' \in \mathcal{P}^l(\Gamma), f(\succeq) = f(\succeq')$.

Note that if Γ is an implementation of f, then if agents always act truthfully, Γ is guaranteed to always elicit enough information from agents in order to implement f. We now make this precise. Fix an agent i, and denote the nodes where i plays in Γ by $N_i(\Gamma) = \{n \in N(R) : \pi(n) = i\}$. A *strategy* for i in Γ is a function $s_i : N_i(\Gamma) \times \mathcal{P}_i \to E(R)$ such that $s_i(n, \succeq_i) \in E(n)$ for all $n \in N_i(\Gamma)$ and $\succeq_i \in \mathcal{P}_i$. We say that strategy s_i is *truthful*, if it always chooses edges that correspond to the agent's true preferences, that is, for any $\succeq_i \in \mathcal{P}_i$ and any $n \in N_i(\Gamma)$ such that $\succeq_i \in \cup_{e\in E(n)}\varphi_n(e), s_i(n, \succeq_i) = e^{\succeq_i}$ where $e^{\succeq_i} \in E(n)$ is the unique edge such that $\succeq_i \in \varphi_n(e^{\succeq_i})$. Given some truthful strategy-profile s and a preference profile $\succeq \in \mathcal{P}$, let $s(\succeq) \equiv s(\cdot, \succeq)$ and let $l(s(\succeq))$ be the terminal node of Γ that is reached when agents play according to $s(\succeq)$. If Γ is an implementation of f, then $f(\succeq') = f(\succeq)$ for all $\succeq' \in \mathcal{P}^{l(s(\succeq))}(\Gamma)$.

The idea behind Definition 3.3 is that $\mathbb{P}^\Gamma$ always collects enough information from the agents in order to unambiguously determine the outcome under f. Hence, if agents always act truthfully in the

sense that they choose the actions that correspond to their true preferences, then for each $\succeq \in \mathcal{P}$, the outcome chosen via Γ will coincide with $f(\succeq)$.

Definition 3.4 (Li (2017)). A rule f is *OSP-implementable* if there exists an implementation Γ of f such that for each non-terminal node $n \in N(R)$ in Γ, $\mathbb{P}^n(\Gamma)$ is obvious for $\pi(n)$ at $\mathcal{P}^n(\Gamma)$.[7]

Note that all the concepts that were introduced so far apply equally well to any sub-domain $\mathcal{P}^n(\Gamma)$, where $n \in N(R)$, that can be generated via Γ. In particular, we can define an extensive-form revelation game Γ' on any $\tilde{\mathcal{P}} \subset \mathcal{P}$. We say that Γ' is an implementation of f on $\tilde{\mathcal{P}} \subset \mathcal{P}$ if $\mathbb{P}^{\Gamma'}$ is a partition of $\tilde{\mathcal{P}}$ and for each $\succeq \in \tilde{\mathcal{P}}$, the outcome chosen via Γ' under agents' truthful strategies is $f(\succeq)$. We are now ready to define our notion of an optimal implementation of f.

Definition 3.5. The extensive-form revelation game $\Gamma^* = (R^*, \tau^*, \pi^*, \varphi^*)$ is an *optimal implementation of f* if

1. Γ^* is an implementation of f,
2. for all $n \in N(R^*)$,
   (a) if $\mathbb{P}^n(\Gamma^*)$ is not obvious, then there is no obvious decision at $\mathcal{P}^n(\Gamma^*)$, and
   (b) there is no implementation Γ' of f on $\mathcal{P}^n(\Gamma)$ such that $\mathbb{P}^{\Gamma'}$ is coarser than $\mathbb{P}^{\Gamma^*_n}$.

As already mentioned in introduction, popular strategy-proof rules are not OSP-implementable in general. Nevertheless, it is often possible to elicit at least partial information about an agent's preferences via obvious decisions. For example, if, as is true for most mechanisms that are studied in the literature, an agent is guaranteed to be matched to an object o when she ranks it first and she has the highest priority for it, then a decision which asks the agent to reveal whether her top choice is o or some

[7] Note that if all decisions an agent made in Γ are obvious, then the truthful strategy is obviously dominant for that agent in Γ. Therefore, our definition of OSP-implementability is equivalent to that in Li (2017).

other object is obvious. More generally, we can break the revelation of a preference $\succeq_i$ into a finite sequence of decisions each of which asks the agent to reveal partial information about $\succeq_i$. From the perspective of obvious strategy-proofness, it is natural to expect that agents may make mistakes each time they are asked to reveal more about their preferences. Definition 3.5 therefore requires optimal implementations to minimize the amount of information that they collect from agents. In Appendix 3.B, we present two examples that describe the types of information (elicited by some non-obvious decisions) which are usually avoided by optimal implementations.

In order to motivate Definition 3.5, we first relate it to OSP-implementability via the following proposition.

Proposition 3.2. *If f is OSP-implementable, then, for any optimal implementation Γ^* of f, all decisions are obvious.*

Proof. Since f is OSP-implementable, we can find an implementation $\Gamma = (R, \tau, \pi, \varphi)$ of f where all decisions are obvious. Now, select any optimal implementation $\Gamma^* = (R^*, \tau^*, \pi^*, \varphi^*)$ of f and select any non-terminal node $n^* \in N(R^*)$. To reach the desired result, we show that we can always find an obvious decision for some agent at $\mathcal{P}^{n^*}(\Gamma^*) \subseteq \mathcal{P}$.

Since $\mathcal{P}^r(\Gamma) = \mathcal{P}$, this ensures us to find (at least) a node $n' \in N(R) \cup L(R)$ in Γ such that $\mathcal{P}^{n^*}(\Gamma^*) \subseteq \mathcal{P}^{n'}(\Gamma)$. Specifically, let n be the last such node in Γ. That is, either $n \in L(R)$, or for each immediate successor $\tilde{n}$ of n in R, $\mathcal{P}^{n^*}(\Gamma^*) \subseteq \mathcal{P}^{\tilde{n}}(\Gamma)$ does not hold. Note that we might find multiple such nodes in Γ, and it is sufficient to focus on an arbitrary one. In the remaining proof, we distinguish three cases based on the properties of n.

- **Case 1:** $n \in L(R)$**.**

  Since Γ implements f, then $f(\succeq) = f(\succeq')$ for any $\succeq, \succeq' \in \mathcal{P}^n(\Gamma)$. However, note that since $\mathcal{P}^{n^*}(\Gamma^*) \subseteq \mathcal{P}^n(\Gamma)$ and n^* is a non-terminal node in Γ^*, by eliminating all the decisions made in Γ^*_n while keeping all other decisions the same as in Γ^*, we get an implementation of f that

induces a coarser partition than that induced by Γ^*. We can then infer that Γ^* fails to satisfy condition 2.(b) of Definition 3.5, which contradicts to Γ^* being an optimal implementation of f. Thus, we have $n \notin L(R)$.

- **Case 2:** $n \in N(R)$ **and** $\mathcal{P}^{n^*}_{\pi(n)}(\Gamma^*) = \mathcal{P}^{n}_{\pi(n)}(\Gamma)$**.**

  Let $\pi(n) = i$. Recall that Γ provides i with an obvious decision at n. According to Corollary 3.2, it must be true that $S_i(\mathcal{P}^n(\Gamma), f) \neq \emptyset$. Next, as state in Corollary 3.1, all secure objects for an agent will remain secure for her until she is assigned. Therefore, $\mathcal{P}^{n^*}_i(\Gamma^*) = \mathcal{P}^n_i(\Gamma)$ and $\mathcal{P}^{n^*}_{-i}(\Gamma^*) \subseteq \mathcal{P}^n_{-i}(\Gamma)$ jointly indicate $S_i(\mathcal{P}^n(\Gamma), f) \subseteq S_i(\mathcal{P}^{n^*}(\Gamma^*), f)$. Since this shows that $S_i(\mathcal{P}^{n^*}(\Gamma^*), f)$ is non-empty, we conclude by Corollary 3.2 that there is an obvious decision for i at $\mathcal{P}^{n^*}(\Gamma^*)$.

- **Case 3:** $n \in N(R)$ **and** $\mathcal{P}^{n^*}_{\pi(n)}(\Gamma^*) \subset \mathcal{P}^{n}_{\pi(n)}(\Gamma)$**.**

  Again let $\pi(n) = i$. First, by the selection of n, we can find at least two immediate successors n_1, n_2 of n in Γ such that $\mathcal{P}^{n^*}_i(\Gamma^*) \cap \mathcal{P}^{\tilde{n}}_i(\Gamma) \neq \emptyset$ for each $\tilde{n} \in \{n_1, n_2\}$. Next, since Γ provides i with an obvious decision at n, by Proposition 3.1, there exists exactly one immediate successor n' of n in Γ such that $\succeq'_i \in \mathcal{P}^{n'}_i(\Gamma)$ if $top(\succeq'_i |_{O_i(\mathcal{P}^n_i(\Gamma), f)}) \in O_i(\mathcal{P}^n_i(\Gamma), f) \setminus S_i(\mathcal{P}^n_i(\Gamma), f)$. Let $\tilde{n} \in \{n_1, n_2\} \setminus \{n'\}$ be arbitrary and fix any $\succeq \in \mathcal{P}^{n^*}_i(\Gamma^*) \cap \mathcal{P}^{\tilde{n}}_i(\Gamma)$, then it must hold that $top(\succeq |_{O_i(\mathcal{P}^n(\Gamma), f)}) \in S_i(\mathcal{P}^n(\Gamma), f)$. Let $top(\succeq |_{O_i(\mathcal{P}^n(\Gamma), f)}) = o$.

  At last, we show that o is secure for i at $\mathcal{P}^{n^*}(\Gamma^*)$. Since o is secure for i at $\mathcal{P}^n(\Gamma)$, it holds that $f_i(\succeq) = o$ and thus $o \in O_i(\mathcal{P}^{n^*}(\Gamma^*), f)$. Moreover, since $\mathcal{P}^{n^*}_{-i}(\Gamma^*) \subseteq \mathcal{P}^n_{-i}(\Gamma)$, it follows that $O_i(\mathcal{P}^{n^*}(\Gamma^*), f) \subseteq O_i(\mathcal{P}^n(\Gamma), f)$. We can then infer that $top(\succeq |_{O_i(\mathcal{P}^{n^*}(\Gamma^*), f)}) = o$. Therefore, we know by Corollary 3.1 that $o \in S_i(\mathcal{P}^{n^*}(\Gamma^*), f)$ and thus $S_i(\mathcal{P}^{n^*}(\Gamma^*), f)$ is non-empty. As a result, we can again conclude by Corollary 3.2 that there is an obvious decision for i at $\mathcal{P}^{n^*}(\Gamma^*)$.

□

For our purpose, it will prove useful to focus on a specific class of possible implementations of a given matching rule that we define next.

Definition 3.6. Let $\Gamma = (R, \tau, \pi, \varphi)$ be an implementation of f. Then, Γ *only asks about top choices* if, for all $n \in N(R)$, if $top(\succeq_{\pi(n)} |_{O_{\pi(n)}(\mathcal{P}^n(\Gamma),f)}) = top(\succeq'_{\pi(n)} |_{O_{\pi(n)}(\mathcal{P}^n(\Gamma),f)})$ for two preferences $\succeq_{\pi(n)}, \succeq'_{\pi(n)} \in \mathcal{P}^n_{\pi(n)}(\Gamma)$, then $\succeq_{\pi(n)}$ and $\succeq'_{\pi(n)}$ belong to the same cell of the partition induced by $\mathbb{P}^n(\Gamma)$.

In following sections, we introduce implementations of different strategy-proof rules that only ask about top choices. We focus on two well-studied strategy-proof rules: the top trading cycle (TTC) and the agent-proposing deferred acceptance (DA).[8] Aside from their popularity, the key driver of our choice of TTC and DA is that both rules are found to be not OSP-implementable in general (Li, 2017; Ashlagi and Gonczarowski, 2018; Troyan, 2019). In this sense, our upcoming proposals are promising solutions to markets where OSP-implementation of TTC or DA is absent.

3.4 Optimal implementation of TTC

We now introduce a sequential revelation game with perfect information, denoted by Γ^T, that implements TTC under truthful behavior. At each point in the game there is a set of remaining agents I, a set of remaining objects O and a directed path G on node set I. We say that a remaining agent $i \in I$ *owns* a remaining object $o \in O$ at I if i has the highest priority on $\rhd_o$ among all agents in I. At the start of Γ^T, let I and O be the same as in the original problem and let G be empty. The triple (I, O, G) is updated as follows.

[8] Since the formal definitions of TTC and agent-proposing DA are familiar to most readers, we relegate them to Appendix 3.A

Stage 1: For each agent $i \in I$, let

$$I_i = \{i\} \cup \{k \in I : \text{there exists a path from } k \text{ to } i \text{ in } G\}$$

and let

$$S_i = \{i\} \cup \{o \in O : \text{there exists } k \in I_i \text{ who owns } o \text{ at } I\}.$$

Ask each agent $i \in I$ to reveal whether her top choice out of the set $O \cup \{i\}$ is in $O \setminus S_i$ or which one of the objects in S_i.[9]

Once one such i reveals her top choice to be some $o^* \in S_i$, let i point to the agent who owns o^* at I in G and move to Stage 2. If each such i reveals her top choice to be in $O \setminus S_i$, move to Stage 3.

Stage 2: There is a directed cycle C in G. Ask each agent involved in C to further specify her top choice among the objects owned by the agent she points to in C. Assign all these agents to their reported top choices, remove them from I and O respectively, update G accordingly and move back to Stage 1.

Stage 3: If G is non-empty, select the agent i in G who does not have an outgoing edge; otherwise, let i be the smallest indexed agent among those who own most objects at I.[10] Then, ask i to reveal the agent $k \in I$ who owns her top choice at I, add the edge ik to G and move back to Stage 1.

The game terminates when O or I becomes empty. After termination, Γ^T yields the matching that comprises the assignments made in all nodes where Stage 2 is reached.

According to Proposition 3.1, all decisions in Stage 1 are obvious since for any agent $i \in I$, the set S_i contains all objects that she could secure by reporting as favorite. The same also applies to Stage

[9] We assume that repeated decisions are avoided. That is, i is asked to make such a revelation in Stage 1 if and only if compared to last time the game reaches i in Stage 1, the set S_i or O has changed.

[10] That is, we always select the smallest indexed agent i from those agents k for whom $|S_k|$ is maximal among all agents in I.

2, since in Stage 2 agents get whatever they reveal as top choices. The decisions in Stage 3 will not in general be obvious to agents. Exceptions include when there are only two agents in I, denoted by i_1, i_2, who do not have an outgoing edge in G. Note that G is always a path upon reaching Stage 3 (and after Stage 3). Thus, in this case, i_1's top choice is in S_{i_2} and i_2's top choice is in S_{i_1}. As a result, both agents receive whatever they report as top choices – the decisions are obvious. A detailed discussion about another kind of exceptions can be found right below in Example 3.1.

The following is our main result on TTC.

Theorem 3.1. *The extensive-form revelation game Γ^T is an optimal implementation of TTC.*

Proof. See Appendix 3.C. □

The following corollary follows immediately from Proposition 3.2 and Theorem 3.1.

Corollary 3.3. *Γ^T is OSP whenever TTC is OSP-implementable.*

Notably, in problems where Γ^T is not OSP, we might find some nodes in Γ^T such that there exist preference profiles which satisfy the following two statements: First, these preference profiles are compatible with the information collected at that node. Second, Γ^T does not implement the TTC outcomes of these profiles in obviously dominant strategies while one can implement those TTC outcomes in obviously dominant strategies. The following example presents such a scenario.

Example 3.1. There are four agents $I = \{i_1, i_2, i_3, i_4\}$ and four objects $O = \{o_1, o_2, o_3, o_4\}$. Priority structure $\rhd_O$ is given by the following table.

$\rhd_{o_1}$	$\rhd_{o_2}$	$\rhd_{o_3}$	$\rhd_{o_4}$
i_1	i_1	i_2	i_3
i_2	i_4	i_4	i_4
i_3	i_2	i_1	i_1
i_4	i_3	i_3	i_2

Note that since there are three agents who rank top at some priorities, the *weak acyclicity* condition[11] of Troyan (2019) is violated in $\rhd_O$ and TTC is thus not OSP-implementable. Specifically, suppose that agents' true preferences are given by the following table.

$\succeq_{i_1}$	$\succeq_{i_2}$	$\succeq_{i_3}$	$\succeq_{i_4}$
o_3	o_4	o_3	o_4
o_4	o_3	o_1	o_3
o_1	o_1	o_4	o_1
o_2	o_2	o_2	o_2

Assume that agents are truthful until the first time we reach Stage 3 in Γ^T. Then, we claim that if we deviate from Γ^T by first asking i_2 about her top choice in Stage 3, we can find an implementation Γ' in which it is obviously dominant for all agents with the above profile to tell the truth. Concretely, the preference for i_2 ensures that truth-telling is obviously dominant for her in Stage 3:

- With respect to all preference profiles that are still possible at the first time we reach Stage 3, the worst case for reporting truthfully for i_2 is being assigned to o_3.
- The best case for misrepresenting top choice is to be assigned o_3: If i_2 does not get her top choice o_4, then i_3 must have traded o_4 to i_1; best remaining object in this case is o_3.

Assume that i_2 follows the obviously dominant strategy and we have inferred i_2's top choice with her obviously dominant revelation. Then, let Γ' proceed to Stage 1 with i_3. Note that since i_3 will be assigned to her top choice o_3 by reporting truthfully, truth-telling is then obviously dominant for her. After i_2 and i_3 is assigned to o_4 and o_3 respectively, let Γ' turn to i_1 as it is then obviously dominant for

[11] A *strong cycle* in a priority structure $\rhd_O$ is described by three agents $i, j, k \in I$ and three objects $a, b, c \in C$ such that $i \rhd_a j, k, j \rhd_b i, k$ and $k \rhd_c i, j$. If there are no strong cycles, the priority structure is said to be *weakly acyclic*. Troyan (2019) shows that TTC is OSP-implementable in a market if and only if the underlying priority structure is weakly acyclic.

i_1 to report her top choice between o_1 and o_2. In conclusion, for the above profile, Γ' can implement the TTC outcome so that truth-telling is obviously dominant for all agents.

Since acting truthfully is not obviously dominant for i_1 in Γ^T (as she is the first one to be approached in Stage 3), the example shows that, for a given preference profile, Γ^T may fail to implement the TTC outcome in obviously dominant strategies even though that outcome is implementable in obviously dominant strategies by another implementation Γ' of TTC.

As a remark, if we change the preferences of above agents i_1 and i_2 to $o_3 \succeq_{i_1} o_1 \succeq_{i_1} o_2 \succeq_{i_1} o_4$ and $o_1 \succeq_{i_2} o_4 \succeq_{i_2} o_2 \succeq_{i_2} o_3$, we construct a profile for which the TTC outcome is implemented in obviously dominant strategies in Γ^T but not in the deviation Γ'. In fact, for any implementation of TTC that outperforms Γ^T in terms of implementing the TTC outcome of certain preference profiles in obviously dominant strategies, we can find some other preference profiles for which Γ^T outperforms that implementation.

3.5 Weakly optimal implementation and DA

In this section, we focus on DA, which we denote by f^{DA} hereafter. In the following discussion, for any $i \in I$ and $\tilde{\mathcal{P}} \in \mathcal{P}$, we simplify $O_i(\tilde{\mathcal{P}}, f^{DA})$ as $O_i(\tilde{\mathcal{P}})$ if there is no risk of confusion.

3.5.1 Incompatibility result for DA

We start this section by presenting a challenge in designing an implementation of DA that is optimal in the sense of Definition 3.5. Concretely, in some cases, we will face the incompatibility between selecting an existing obvious decision and collecting the least necessary information from agents, as illustrated in the next example.

Example 3.2. Consider a problem with six agents $I = \{i_1, \dots, i_6\}$ and five objects $O = \{o_1, \dots, o_5\}$. Let $\rhd_O$ be

$\rhd_{o_1}$	$\rhd_{o_2}$	$\rhd_{o_3}$	$\rhd_{o_4}$	$\rhd_{o_5}$
i_1	i_2	i_3	i_4	i_5
i_6	i_6	i_4	i_5	i_3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	i_6	i_6	i_6
$\vdots$	$\vdots$	i_1	i_1	i_1

The above table indicates that i_6 and i_1 rank lowest at $\rhd_{o_3}$, $\rhd_{o_4}$ and $\rhd_{o_5}$. Let the true preference profile $\succeq$ be given in the following table.

$\succeq_{i_1}$	$\succeq_{i_2}$	$\succeq_{i_3}$	$\succeq_{i_4}$	$\succeq_{i_5}$	$\succeq_{i_6}$
o_2	o_1	o_1	o_1	o_1	o_1
o_1	o_2	o_4	o_5	o_3	o_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Let Γ be an implementation of DA that always asks obvious decisions whenever they exist. In the following, we will describe the path in Γ that corresponds to the truthful behaviors of agents with the above profile. Specifically, we show that at a certain node in the underlying path, i_1 reveals more information than necessary to compute the outcome as Γ prioritizes picking an obvious decision (of i_1) over non-obvious decisions (of other agents).

At the initial steps, Γ asks $i_1, \ldots, i_5$ whether their top choices (among $\{o_1, \ldots, o_5\}$) are the ones for which they have top priority. Note that each agent can secure the objects for which she ranks highest by reporting them as her top choice. Thus, all these decisions are obvious according to Proposition 3.1. For the given profile $\succeq$, we know that all agents will answer "No". Then, there are no obvious decisions.

Next, Γ turns to i_6 and asks her whether her top choice (among $\{o_1, \ldots, o_5\}$) is o_1, for which she ranks second. Note that i_6 answers "Yes" in the underlying path and thus she is temporarily assigned

to o_1. Since o_1 is then no longer available to $i_2, \ldots, i_5$, there are again obvious decisions for $i_2, \ldots, i_5$, namely whether their top choices (among $\{o_2, \ldots, o_5\}$) are objects for which they have top priority. At this point, it is obviously dominant for i_2 to report that o_2 is her top choice. After i_2 reports so, she is assigned to o_2 with certainty. Then, Γ turns to i_3, i_4 and i_5 with obvious decisions: Whether their top choices (among $\{o_3, o_4, o_5\}$) are those at which they rank top.

In the following, consider the node in Γ where all i_3, i_4 and i_5 answer "No" to the just mentioned obvious decisions. Specifically, denote this node by n. Note that it remains exactly one obvious decision at n, which is to ask i_1 to reveal whether o_1 is her top choice among $\{o_1, o_3, o_4, o_5\}$. By the assumption that Γ picks an obvious decision whenever one exists, Γ can only ask i_1 to make this obvious decision at n. We next argue that some of the information revealed by i_1 at n is redundant for computing the outcome.

In fact, n is a node in Γ where i_1's obvious decision will not be informative to i_3, i_4 and i_5 at all. In particular, if i_1 reveals that her top choice is not o_1, it will not contribute to ease the decisions of i_3, i_4, i_5 since they all have higher priority than i_1 at $\rhd_{o_3}, \rhd_{o_4}$ and $\rhd_{o_5}$. On the contrary, if i_1 reveals that her top choice is o_1, it will cause i_6 to be rejected by o_1. However, since all i_3, i_4, i_5 already know that o_1 is not available, this information will not be helpful either. Moreover, note that since i_6 ranks lower than i_3, i_4, i_5 at $\rhd_{o_3}, \rhd_{o_4}$ and $\rhd_{o_5}$, then i_6's any further decision also makes no difference for i_3, i_4, i_5.

Finally, we construct an implementation Γ' of DA that induces a coarser partition than that induced by Γ. Concretely, Γ' is identical to Γ except in the sub-game Γ_n: When n is reached, Γ' directly picks the decision that Γ picks at an immediate successor of n. Apparently, Γ' is not optimal. However, at any successor of n in Γ' where i_1 plays, she reveals weakly less about her preferences than she reveals at n in Γ. To be more specific, recall that i_1 reveals whether o_1 is her top choice among $\{o_1, o_3, o_4, o_5\}$ at n in Γ. In Γ', instead, i_1 only reveals whether o_1 is her top choice among a subset of $\{o_1, o_3, o_4, o_5\}$ since the decisions of i_3, i_4, i_5 might cause some of $\{o_3, o_4, o_5\}$ to be no longer available to i_1. As a

conclusion, Γ' induces a coarser partition than Γ does in terms of i_1's decisions (in the sub-game Γ_n).

Example 3.2 describes a specific scenario where we fail to induce the coarsest partition among DA implementations if we always provide obvious decisions whenever they exist. Moreover, it remains unclear how to avoid encountering such scenarios when we design an optimal implementation of DA that generally exists. Towards a solution concept which accounts for such a challenge, it seems reasonable to have tolerance for redundant decisions as long as they are obvious for agents. Thus, we next provide a weakening of optimality in Definition 3.5 that allows agents to reveal unnecessary information via obvious decisions.

Fix a matching rule f and two implementations Γ and $\hat{\Gamma}$ of f. If $\mathbb{P}^{\hat{\Gamma}}$ is coarser than $\mathbb{P}^{\Gamma}$, then at least two elements $\mathcal{P}^1, \mathcal{P}^2 \in \mathbb{P}^{\Gamma}$ are contained in the same cell of $\mathbb{P}^{\hat{\Gamma}}$. Accordingly, there must be (at least) one agent $i \in I$ who plays at some $n \in N(R)$ such that $\mathcal{P}_i^1, \mathcal{P}_i^2$ are contained in different elements of the decision $\mathbb{P}^n(\Gamma)$ for which f is constant. We now formally define how the decision $\mathbb{P}^n(\Gamma)$ causes $\mathbb{P}^{\Gamma}$ to be more revealing than $\mathbb{P}^{\hat{\Gamma}}$.

Definition 3.7. Consider two implementations Γ and $\hat{\Gamma}$ of f such that $\mathbb{P}^{\hat{\Gamma}}$ is coarser than $\mathbb{P}^{\Gamma}$. Fix any $n \in N(R)$ and let $\pi(n) = i$. We say $\mathbb{P}^{\hat{\Gamma}}$ is coarser than $\mathbb{P}^{\Gamma}$ *caused by* $\mathbb{P}^n(\Gamma)$ if there exists $\hat{\mathcal{P}}^s \in \mathbb{P}^{\hat{\Gamma}}$ and $\succeq, \succeq' \in \hat{\mathcal{P}}^s$ such that

1. $\succeq, \succeq' \in \mathcal{P}^n(\Gamma)$, and
2. $\succeq_i \in \mathcal{P}_i^t$ and $\succeq_i' \in \mathcal{P}_i^{t'}$ for distinct $\mathcal{P}_i^t, \mathcal{P}_i^{t'} \in \mathbb{P}^n(\Gamma)$.

We are now ready to define a weakly optimal implementation of f.

Definition 3.8. The extensive-form game $\Gamma^* = (R^*, \tau^*, \pi^*, \varphi^*)$ is a *weakly optimal implementation of* f if

1. Γ^* is an implementation of f,

2. for all $n \in N(R^*)$,

   (a) if $\mathbb{P}^n(\Gamma^*)$ is not obvious, then there is no obvious decision at $\mathcal{P}^n(\Gamma^*)$, and

   (b) if $\mathbb{P}^n(\Gamma^*)$ is not obvious, then there is no implementation $\hat{\Gamma}$ of f on $\mathcal{P}^n(\Gamma^*)$ such that $\mathbb{P}^{\hat{\Gamma}}$ is coarser than $\mathbb{P}^{\Gamma^*_n}$ caused by $\mathbb{P}^n(\Gamma^*)$.

In words, an implementation is weakly optimal if it always picks obvious decisions when they exist and if it guarantees agents the minimal amount of *non-obvious* decisions. Compared to the optimality in Definition 3.5, we loosen condition 2.(b) in Definition 3.8 by only checking for non-obvious decisions. In weakly optimal implementation, agents might reveal information that turns out to be irrelevant for the final matchings. However, it is ensured that all such irrelevant information is elicited through obvious decisions. It is worthy of mentioning that Proposition 3.2 also holds with our weaker notion of optimality.

3.5.2 Characterization of secure objects in DA implementations

In this subsection, we characterize the set of secure objects at any node in an implementation of DA. According to Corollary 3.2, this characterization is helpful for figuring out obvious decisions under DA, which in turn contributes to the design of a weakly optimal implementation of DA in the next subsection.

In general, the identification of secure objects in DA is more challenging than for TTC. Recall that in implementations of TTC, we can identify a secure object of an agent by simply checking whether that agent is assigned to that object if she reports it as favorite. In implementations of DA, however, this strategy is not working since the assignments made during the process of DA are temporary. That is, even if an agent is assigned to an object at some intermediate stage of DA, she might be rejected later.

As a starting point, note that there are secure objects that can be easily identified in DA implementations: For agents who own some objects, the objects they own are secure for them. That is, revealing whether their favorite objects are among the ones they own, is an obvious decision. Thus, when any of the answers to these obvious decisions is "Yes", we can update the market immediately. When all answers are "No", however, it is not clear how to uncover secure objects then. Towards a solution, let us first consider the following example that describes such a scenario.

Example 3.3. There are four agents $I = \{i_1, i_2, i_3, i_4\}$ and four objects $O = \{o_1, o_2, o_3, o_4\}$. The priority structure $\rhd_O$ is given in the following table.

$\rhd_{o_1}$	$\rhd_{o_2}$	$\rhd_{o_3}$	$\rhd_{o_4}$
i_1	i_1	i_1	i_2
i_2	i_3	i_3	i_1
i_3	i_2	i_2	i_3
i_4	i_4	i_4	i_4

At the initial step, there are obvious decisions for i_1 and i_2: Whether among $\{o_1, o_2, o_3, o_4\}$, her top choice is in $\{o_1, o_2, o_3\}$ and $\{o_4\}$, respectively. Suppose that both agents' answers are "No". Clearly, in this case, i_1 implicitly reveals that her top choice is o_4 and thus the following two statements are true. First, i_1 is temporarily assigned to o_4 and no revelation from i_1 is needed until she is rejected by o_4. Second, regardless of i_1's final assignment, the stability of DA ensures that i_3 and i_4 (who rank lower than i_1 on $\rhd_{o_4}$) cannot be matched with o_4 under DA. That is, o_4 is no longer available to i_3 and i_4.

For the remainder of this example, we study whether there are secure objects for i_2, i_3 and i_4 after both i_1 and i_2 answer "No" to the above obvious decisions. First, it is immediate that i_4 has no secure objects since she ranks lowest at all objects.

Next, we claim that o_1 is then a secure object for i_2. Concretely, if i_2 reveals that her top choice is o_1, then the temporarily assigned pairs are $\{(i_1, o_4), (i_2, o_1)\}$. Since only i_1 could let i_2 be rejected by

o_1 and only i_2 could let i_1 be rejected by o_4, and since both of them are temporarily assigned to their top choices, they will not be rejected anymore. Thus, o_1 is secure for i_2.

Finally, we claim that i_3 has no secure object at this point. Since o_4 is no longer available to i_3, the available objects for i_3 are $\{o_1, o_2, o_3\}$. We show that o_2 (for which i_3 ranks second) is not secure for i_3. For instance, if i_3's true preference is $\succeq_{i_3}$: $o_2, o_3 \ldots$, then the worst possible outcome of reporting o_2 as favorite is o_3 (with the preferences $\succeq_{i_1}$: $o_4, o_2, \ldots$ and $\succeq_{i_2}$: $o_2, o_4, \ldots$), and the best possible outcome of reporting o_3 as favorite is o_2 (with the preferences $\succeq'_{i_1}$: $o_4, o_3, \ldots$ and $\succeq_{i_2}$: $o_3, o_4, \ldots$). Since the arguments for the other two objects are similar, we omit the details here.

Intuitively, compared to other agent-object pairs, the pair (i_2, o_1) is distinct since i_1, who could potentially let i_2 be rejected by o_1, is temporarily assigned to o_4 at which only i_2 ranks higher than i_1. In fact, o_1 could be secure for i_2 even when i_1 is not temporarily assigned. To illustrate this, we slightly adjust $\rhd_{o_3}$ in the above example such that we cannot infer the exact top choice of i_1 from the "No" answers.

Example 3.3 (Continued)**.** The adjusted priorities are:

$\rhd_{o_1}$	$\rhd_{o_2}$	$\rhd_{o_3}$	$\rhd_{o_4}$
i_1	i_1	$\mathbf{i}_2$	i_2
i_2	i_3	$\mathbf{i}_1$	i_1
i_3	i_2	$\mathbf{i}_3$	i_3
i_4	i_4	i_4	i_4

Again, consider the point where both i_1 and i_2 answer "No" at the beginning. We claim that o_1 is still a secure object for i_2 then. Note that when i_1 answers "No", her top choice can only be o_3 or o_4. If i_2 reports o_1 as favorite, the potential set of temporarily assigned pairs is either $\{(i_1, o_3), (i_2, o_1)\}$ or $\{(i_1, o_4), (i_2, o_1)\}$. With the same reasoning as above, we know that i_2 is ultimately assigned to o_1 under DA in both cases.

Generally speaking, at the underlying points in the above examples, once i_2 is temporarily assigned to o_1, all agents who could potentially cause i_2 to be rejected by o_1 are (or will be) assigned to objects by which they will not be rejected. In other words, after o_1 is temporarily allocated to i_2, no continuation of the game will have some i with $i \rhd_{o_1} i_2$ applying to o_1. Notably, as will become clear soon, this feature is vital for o_1 being secure for i_2.

In the reminder of this subsection, we first formalize the generalization of the above described feature, and then use it to characterize secure objects under DA. To this end, fix any matching $\mu \in \mathcal{M}$, we denote $\mu = \{(i_1, o_1), \ldots\}$ such that it contains only agent-object pairs and $(i, o) \in \mu$ if and only if $\mu(i) = o$. For any two matchings $\mu, \mu' \in \mathcal{M}$, we write $\mu' \subseteq \mu$ if for each $(i, o) \in \mu'$, it holds $\mu(i) = o$.

Next, fix an implementation Γ of DA and a non-terminal node $n \in N(R)$. We use a matching μ^n to describe all temporarily assigned pairs at n. Formally, $(i, o) \in \mu^n$ if and only if $top(\succeq_i |_{O_i(\mathcal{P}^n(\Gamma))}) = o$ for all $\succeq_i \in \mathcal{P}_i^n(\Gamma)$. Abusing notation, we say $i \in \mu^n$ when i is contained in some $(i, o) \in \mu^n$. The next definition regarding μ^n generalizes the feature we derive from Example 3.3.

Definition 3.9. Fix any implementation Γ of DA and $n \in N(R)$. A matching $\mu' \subseteq \mu^n$ is *anchored* in $(I, O, \rhd_O)$ if for all $(i, o) \in \mu'$ and all $j \in I$ such that $j \rhd_o i$, it holds $j \in \mu'$.

That is, a matching μ' is anchored during a DA implementation when for any $\mu \in \mathcal{M}$ such that $\mu' \subseteq \mu$, no agent $i \notin \mu'$ can have justified envy towards any $i' \in \mu'$ at μ. For instance, in Example 3.3 when i_2 reports o_1 as her top choice, the matching $\{(i_1, o_4), (i_2, o_1)\}$ is anchored.[12] Importantly, note that since any $i \in \mu^n$ is assigned to her top choice among her available objects at $\mathcal{P}^n(\Gamma)$, she will not have justified envy when μ^n is part of the final matching. Therefore, once an anchored matching $\mu' \subseteq \mu^n$ is formed during DA implementations, it is immune to justified envy from both inside and outside μ'. In conclusion, μ' will be fixed at any continuation of the underlying revelation game. We formally present this result.

[12] Note that an anchored matching can also be singleton: If an agent i is temporarily assigned to one of her owned objects o, then no one ranks higher than i on $\rhd_o$ and $\mu' = \{(i, o)\}$ is anchored.

Corollary 3.4. *Fix any implementation* Γ *of DA and* $n \in N(R)$. *If* $\mu' \subseteq \mu^n$ *is anchored in* $(I, O, \rhd_O)$, *then for all* $\succeq \in \mathcal{P}^n(\Gamma)$ *and* $(i, o) \in \mu'$, *it holds* $f_i^{DA}(\succeq) = o$.

According to Corollary 3.4, we can infer that to predict whether an object o is secure for an agent i, we just need to check whether (i, o) is part of an anchored matching in any consistent future. Notably, it is not enough to only check with the set of temporarily assigned pairs at the underlying point. To see this, consider in Example 3.3 (Continued), when both i_1 and i_2 answer "No" and i_2 reveals that o_1 is her top choice, we have $\mu^n = \{(i_2, o_1)\}$. However, since we further know that i_1 prefers either o_3 or o_4 most, there are two candidates for μ^n, namely $\{(i_1, o_3), (i_2, o_1)\}$ and $\{(i_1, o_4), (i_2, o_1)\}$, and both candidates are anchored. As a result, even if we cannot find an anchored matching in μ^n, we still know that o_1 is secure for i_2.

This motivates us to define a set of potential matchings at n, which is the final notation necessary for our characterization of secure objects under DA. Let $I(n)$ be the set of agents who have played at least once when n is reached in Γ and let $K = \{i \in I(n) : i \notin \mu^n\}$ be the set of agents from $I(n)$ who have not revealed their top choices until n. That is, for each $k \in K$, there are $\succeq_k, \succeq'_k \in \mathcal{P}_k^n(\Gamma)$ such that $top(\succeq_k |_{O_k(\mathcal{P}^n(\Gamma))}) \neq top(\succeq'_k |_{O_k(\mathcal{P}^n(\Gamma))})$. Moreover, for any $i \notin \mu^n$, *updating* a pair (i, o) to μ^n is defined by the operation $\odot$ that yields a new matching $\mu = (i, o) \odot \mu^n$ such that

$$\mu = \begin{cases} \mu^n \cup \{(i, o)\} \setminus \{(i', o)\}, & \text{if } \exists i' \neq i \text{ such that } (i', o) \in \mu^n; \\ \mu^n \cup \{(i, o)\}, & \text{otherwise.} \end{cases}$$

Fix any $i \notin \mu^n$, let $\mathcal{U}_i^n$ be the set of all possible matchings that i could infer at n. In particular, let $\{k_1, \ldots, k_T\}$ denote all agents in $K \setminus \{i\}$, then $\mu \in \mathcal{U}_i^n$ if and only if

$$\mu = (k_1, o_1) \odot \ldots \odot (k_T, o_T) \odot \mu^n$$

where for each $t \leq T$, there exists $\succeq_{k_t} \in \mathcal{P}^n_{k_t}(\Gamma)$ such that $top(\succeq_{k_t}|_{O_{k_t}(\mathcal{P}^n(\Gamma))}) = o_t$.

We are now ready to characterize the secure objects in DA implementations.

Proposition 3.3. *Let Γ be an implementation of DA that only asks about top choices, $n \in N(R)$ be arbitrary and $\pi(n) = i$. An object $o \in O_i(\mathcal{P}^n(\Gamma))$ is secure for i at $\mathcal{P}^n(\Gamma)$ if and only if for any $\mu \in \mathcal{U}^n_i$, there exists $\mu' \subseteq \mu$ such that $\mu^* = (i,o) \odot \mu'$ is anchored in $(I, O, \rhd_O)$.*

Proof. We first prove the "if" part. To do so, construct the following implementation $\hat{\Gamma}$ of DA that deviates from Γ from node n onwards. At n, $\hat{\Gamma}$ asks $k_1 \in K \setminus \{i\}$ to reveal her top choice among $O_{k_1}(\mathcal{P}^n(\Gamma))$. After k_1 reveals, $\hat{\Gamma}$ turns to $k_2 \in K \setminus \{i, k_1\}$ and asks her to reveal her top choice among $O_{k_2}(\mathcal{P}^n(\Gamma))$, and so on. After all agents in $K \setminus \{i\}$ have revealed, $\hat{\Gamma}$ turns to i and provides i with the same decision as Γ provides at n. Let $\hat{N}$ collect all nodes in $\hat{\Gamma}$ where i faces such a decision. Then, it follows $\mathcal{P}^n(\Gamma) = \{\mathcal{P}^{\hat{n}}(\hat{\Gamma})\}_{\hat{n} \in \hat{N}}$.

Select any $\hat{n} \in \hat{N}$. By definition of $\mathcal{U}^n_i$, we can find $\mu \in \mathcal{U}^n_i$ such that $\mu^{\hat{n}} = \mu$. That is, there exists $\mu' \subseteq \mu^{\hat{n}}$ such that $\mu^* = (i,o) \odot \mu'$ is anchored in $(I, O, \rhd_O)$. Fix any $\succeq^*_i \in \mathcal{P}^{\hat{n}}_i(\hat{\Gamma})$ such that $top(\succeq^*_i|_{O_i(\mathcal{P}^n(\Gamma))}) = o$. According to Corollary 3.4, it holds that $f^{DA}_i(\succeq^*_i, \hat{\succeq}_{-i}) = o$ for all $\hat{\succeq}_{-i} \in \mathcal{P}^{\hat{n}}_{-i}(\hat{\Gamma})$. Note that since $\hat{n}$ is arbitrarily taken, the above result holds for all nodes in $\hat{N}$ and thus $f^{DA}_i(\succeq^*_i, \succeq_{-i}) = o$ for all $\succeq_{-i} \in \mathcal{P}^n_{-i}(\Gamma)$.

We proceed with the "only if" part. Suppose that $o \in O_i(\mathcal{P}^n(\Gamma))$ is secure for i at $\mathcal{P}^n(\Gamma)$, then we have $f^{DA}_i(\succeq^*_i, \succeq_{-i}) = o$ for all $\succeq_{-i} \in \mathcal{P}^n_{-i}(\Gamma)$. Select any $\mu \in \mathcal{U}^n_i$. In the remaining proof, we extract $\mu' \in \mu$ such that $\mu^* = (i,o) \odot \mu'$ is anchored in $(I, O, \rhd_O)$.

We first claim that for any j such that $j \rhd_o i$, it holds $j \in \mu$. Suppose that there exists $j \in I$ such that $i \rhd_o j$ and $j \notin \mu$, and we aim at a contradiction to o being secure for i. As argued before, we can find $\hat{n} \in \hat{N}$ in the constructed $\hat{\Gamma}$ such that $\mu^{\hat{n}} = \mu$. Select any $\succeq_j \in \mathcal{P}^{\hat{n}}_j(\hat{\Gamma})$, and let $top(\succeq_j|_{O_j(\mathcal{P}^n(\Gamma))}) = o_j$. According to $\hat{\Gamma}$, j has revealed that her top choice among $O_j(\mathcal{P}^n(\Gamma))$ is o_j before $\hat{n}$. Then, $j \notin \mu^{\hat{n}}$ implies that j is rejected by o_j and $o_j \notin O_j(\mathcal{P}^{\hat{n}}_j(\hat{\Gamma}))$. Construct $\succeq'_j$ such that o ranks right after o_j on

$\succeq'_j$ and $o' \succeq'_j o''$ if and only if $o' \succeq_j o''$ for all $o', o'' \in O \setminus \{o\}$. Since Γ only asks about top choices, it follows $\succeq'_j \in \mathcal{P}_j^{\hat{n}}(\hat{\Gamma})$. However, this implies that when j reports according to $\succeq'_j$ at the continuation of $\hat{n}$ in $\hat{\Gamma}$, i will be rejected by o given $j \rhd_o i$. This contradicts to the assumption that o is secure for i.

Thus, we collect in μ' all pairs (j, o_j) from μ such that $j \rhd_o i$. If $(i, o) \odot \mu'$ is anchored in $(I, O, \rhd_O)$, we are done. If not, this indicates that for some $j \in \mu'$, there exists k such that $k \notin \mu'$ and $k \rhd_{o_j} j$. We next claim that for such k, we can find a pair $(k, o_k) \in \mu$. By contradiction, if $k \notin \mu$, then k is rejected by o_k before $\hat{n}$ and $o_k \notin O_k(\mathcal{P}_k^{\hat{n}}(\hat{\Gamma}))$. Notably, with the same arguments as above, we can construct a consistent profile such that after k is rejected by o_k, she applies to o_j which causes j to be rejected by o_j and then to apply to o. In such a scenario, i is finally rejected by o, which again contradicts to o being secure for i. As a result, $k \in \mu$ and we update $\mu' = (k, o_k) \odot \mu'$.

We inductively apply the same reasoning and update corresponding pairs from $\mu \setminus \mu'$ to μ'. Since μ is finite, we will finally reach a matching $\mu' \subseteq \mu$ such that $(i, o) \odot \mu'$ is anchored in $(I, O, \rhd_O)$. This completes the proof. □

Looking carefully at the proof of Proposition 3.3, we know that to predict whether an object o is secure for an agent i, we just need to ensure that no consistent future will have some agent j with $j \rhd_o i$ applying to o. Moreover, it is guaranteed to exist no such future when all such j will be assigned to whatever she could still report as her top choice. In this sense, Proposition 3.3 provides a method to identify secure objects without computing the results for all remaining preference profiles.

3.5.3 Weakly optimal implementation of DA

In this section, we introduce a sequential revelation game with perfect information, denoted by Γ^D, that implements DA under truthful behavior. Towards this goal, we first define some additional terminology. Given some $\mathcal{P}^n \subseteq \mathcal{P}$,

1. let $(I, O, \rhd_O)$ be the reduced problem consisting only of agents and objects for whom the DA-matching is *not* identical across all $\succeq \in \mathcal{P}^n$;

2. let $J \subseteq I$ be the set of active agents j who have revealed their top choices, that is, for whom $top(\succeq_j |_{O_j(\mathcal{P}^n)}) = top(\succeq'_j |_{O_j(\mathcal{P}^n)})$ for all $\succeq_j, \succeq_{j'} \in \mathcal{P}^n_j$;

3. let $\{I_o\}_{o \in O}$ be such that each I_o is the set of agents i who can still reveal (or have already revealed) that o is her top choice among $O_i(\mathcal{P}^n)$. That is, let

$$I_o = \{i \in I : o \in O_i(\mathcal{P}^n) \text{ and } o = top(\succeq_i |_{O_i(\mathcal{P}^n)}) \text{ for some } \succeq_i \in \mathcal{P}^n_i\}.$$

   In the following, denote by i_o the agent who ranks top on $\rhd_o$ among all agents in I_o;[13]

4. let $\{K_o\}_{o \in O}$ be such that each K_o consists of all agents who rank higher than i_o on $\rhd_o$, that is, $K_o = \{k \in I : k \rhd_o i_o\}$ contains all agents who already state that o are not their top choices among their available objects; and

5. let $O^u \subseteq O$ be the set of all *uninformative objects* in the problem $(I, O, \rhd_O)$ that are yielded via the following algorithm:

   Step 0**:** Let O^u be empty.

   Step k, $k \geq 1$**:** Find any set $O' \subseteq O \setminus O^u$ such that $O' \neq O$ and

$$(\cup_{o' \in O'} I_{o'}) \cap (\cup_{o \in O \setminus (O^u \cup O')} K_o) = \emptyset,$$

   add all objects in O' to O^u and move to the next step. If no such set exists, terminate and output O^u.

[13] Note that if $i_o \in J$, then $I_o = \{i_o\}$. Moreover, this implies that all agents who are still eligible for o have revealed that o is not their top choice. In this case, no further decision regarding o can be made at $\mathcal{P}^n$.

In short, for any $o \in O$, we have $o \in O^{\mu}$ if and only if there is $\bar{O} \subset O$ such that $o \in \bar{O}$ and $(\cup_{\bar{o}\in\bar{O}} I_{\bar{o}}) \cap (\cup_{o'\in O\setminus\bar{O}} K_{o'}) = \emptyset$. Note that it is possible $O^{\mu} = \emptyset$ or $O^{\mu} = O$.

Let $O^* \equiv O \setminus O^{\mu}$ be the set of *informative objects* in the problem $(I, O, \rhd_O)$. Notably, for each $o^* \in O^*$ and each $o \in O \setminus \{o^*\}$, o^* being assigned to i_{o^*} could trigger rejections or invoke new decisions that cause i_o to be changed (to one agent in K_o).

Accordingly, for any subset $\hat{O} \subseteq O$, let $\hat{O}^*$ be the set of informative objects in the restricted problem $(\hat{I}, \hat{O}, \rhd_{\hat{O}})$ where $\hat{I} = \cup_{\hat{o}\in\hat{O}}(I_{\hat{o}} \cup K_{\hat{o}})$.

Moreover, we say that a set of uninformative objects $\bar{O} \subseteq O^{\mu}$ in $(I, O, \rhd_O)$ is *benign* if it satisfies: (1) $\bar{O}^* \neq \emptyset$, where $\bar{O}^*$ is the set of informative objects in the restricted problem $(\bar{I}, \bar{O}, \rhd_{\bar{O}})$ and (2) there is no super set $\bar{O}' \subset O$ of $\bar{O}$ that satisfies the first condition.[14]

The basic components of Γ^D are given by the tuple $(I, O, \rhd_O, J, \{I_o\}_{o\in O}, \{K_o\}_{o\in O}, O^*)$. Moreover, we use μ^n to keep track of all temporarily matched pairs at $\mathcal{P}^n$. According to Corollary 3.4, we can derive the reduced problem $(I, O, \rhd_O)$ from μ^n: If there exists $\mu' \in \mu^n$ that is anchored in the original problem, we remove all agents/objects in μ' from the market and the remaining agents/objects constitute the reduced problem.

Initially, let $(I, O, \rhd_O)$ be the original problem, $\mathcal{P}^0 = \mathcal{P}$, $\mu^0 = \emptyset$, $O^* = O$, $I_o = I$ for all $o \in O$ and $K_o = \emptyset$ for all $o \in O$. At any $\mathcal{P}^n$ for $n \geq 0$, the game Γ^D selects the next decision via the following algorithm:

Stage 1: If no agent has secure objects at $\mathcal{P}^n$,[15] move to Stage 2.

[14] Note that once O^{μ} is non-empty, the existence of a benign uninformative set $\bar{O} \in O^{\mu}$ is guaranteed. Concretely, take any $o \in O^{\mu}$ and let $\bar{O} = \{o\}$, then by definition o is an informative object in the restricted problem $(I_o \cup K_o, \bar{O}, \rhd_{\bar{O}})$ – the set $\bar{O}$ satisfies the first condition of a benign uninformative set. Then, we simply enlarge $\bar{O}$ by adding other objects in O^{μ} until the second condition is also satisfied.

[15] Since we already introduced the identification of secure objects in Proposition 3.3, we omit the details here.

Otherwise, select the smallest indexed $i \in I \setminus J$ among all such agents. Ask i to reveal whether she wants to secure one of her secure objects at $\mathcal{P}^n$ or whether her top choice is one of the objects that she cannot secure.

- If i reveals that a secure object o is her top choice, let $\mu^{n+1} = (i, o) \odot \mu^n$, update the remaining preference profiles $\mathcal{P}^{n+1}$, update the tuple $(I, O, \rhd_O, J, \{I_o\}_{o \in O}, \{K_o\}_{o \in O}, O^*)$ and repeat Stage 1.
- Otherwise, update the remaining preference profiles $\mathcal{P}^{n+1}$ and repeat Stage 1.

Stage 2: First, select an object $o^* \in O$ in the following way:

- If O^* is non-empty, consider only those objects $o \in O^*$ for which $i_o \notin J$ and i_o has changed most recently.[16] Among those objects, pick the smallest indexed one o^*.
- If O^* is empty, collect in $\{\bar{O}_t\}_t$ all benign uninformative sets. Then, find $\bar{O} \in \{\bar{O}_t\}_t$ in which there exist an object $\bar{o} \in \bar{O}^*$ such that for any other $\hat{O} \in \{\bar{O}_t\}_t$, if there is $\hat{o} \in \hat{O}$ with $i_{\bar{o}} \in I_{\hat{o}}$, then $\hat{o} \in \hat{O}^*$ and $i_{\bar{o}} = i_{\hat{o}}$ (If no such set exists, let $\bar{O} \in \{\bar{O}_t\}_t$ be arbitrary). Among those just mentioned objects in $\bar{O}^*$, consider only o for which $i_o \notin J$ and i_o has changed most recently, and pick the smallest indexed object o^*.

Second, based on the selected o^*, ask agent i_{o^*} whether o^* is her top choice among all objects in $O_{i_{o^*}}(\mathcal{P}^n)$.

- If i_{o^*} answers "Yes", let $\mu^{n+1} = (i_{o^*}, o^*) \odot \mu^n$, update the remaining preference profiles $\mathcal{P}^{n+1}$, update the tuple $(J, \{I_o\}_{o \in O}, \{K_o\}_{o \in O}, O^*)$ and repeat Stage 1.
- Otherwise, update the remaining preference profiles $\mathcal{P}^{n+1}$ and repeat Stage 1.

[16] Concretely, for each o, let n_o be the first node along back from n to the root in Γ^D where $i_o^{n_o} \neq i_o$. We consider the objects for which n_o is the closest to n.

The game terminates at node n where $O = \emptyset$ or μ^n contains all agents in the original problem. After termination, all assigned pairs in μ^n are finalized and all remaining agents stay unassigned.

We next present our conjecture on DA.

Conjecture 3.1. *The extensive-form revelation game* Γ^D *is a weakly optimal implementation of DA.*

To prove the conjecture, it seems promising to compare Γ^D with some $\hat{\Gamma}$ that we assume to induce a partition weakly coarser than $\mathbb{P}^{\Gamma^D}$ and show that $\mathbb{P}^{\hat{\Gamma}}$ is coarser than $\mathbb{P}^{\Gamma^D}$ only caused by obvious decisions made in Γ^D. In order to establish the just mentioned result, it is useful to proceed by induction on the number of nodes where Γ^D reaches Stage 2 and to show that the decisions provided by Γ^D and $\hat{\Gamma}$ have to be identical at these nodes.

Since the description of Stage 1 is straightforward, here we discuss Stage 2. We first consider when $O^* \neq \emptyset$ and $o^* \in O^*$ is selected in Stage 2. According to the definition of O^*, each $o \in O \setminus \{o^*\}$ might be finally assigned to some $k \in K_o$ after i_{o^*} reports o^* as her top choice at n. Note that in Stage 2, we always select object o^* for which i_{o^*} has changed most recently. This actually ensures that for each $o \in O$ and each $i \in I_o$ at n, after i_{o^*} chooses o^* at n, there is a possible continuation in Γ^D_n where o becomes unavailable to i before i reveals any information of her preference about o. In other words, for each agent $i \neq i_{o^*}$ who remains at n, if we deviate from Γ^D by asking i to play at n, then no matter what i reveals at n in the deviation, she would reveal less such information in some realizations of Γ^D_n where i_{o^*} selects o^* at n. Intuitively, this in turn contributes to that no other implementation induces a coarser partition than $\mathbb{P}^{\Gamma^D}$ caused by the non-obvious decision Γ^D picks at n.

Next, we consider when $O^* = \emptyset$ in Stage 2. Informally speaking, as Γ^D turns to agents according to priorities from top to bottom for each object, it is very likely that an object is informative if it is still unassigned. That is, $O^* = \emptyset$ is usually reached in scenarios where all objects are either temporarily assigned or revealed to be not the top choice by most agents. As for each benign uninformative set $\bar{O} \in \{\bar{O}_t\}_t$, it is the maximal set of uninformative objects that can form a restricted problem in which

the set of informative objects $\bar{O}^*$ is non-empty. Although the decision i_{o*} made at n cannot influence the assignments of all objects, it is vital in at least one restricted problem. Moreover, Γ^D prioritizes selecting an agent i_{o^*} whose decision will not be influenced by assignments in other restricted problems. In this sense, Γ^D guarantees a non-obvious decision at n that will not cause $\mathbb{P}^{\Gamma^D}$ to be more revealing than the partition induced by any other implementation.

3.6 Conclusion

We develop an optimality notion for sequential implementations of matching rules that selects an obvious decision whenever it exists and minimizes the amount of information revealed by agents. An optimal implementation guarantees to implement a rule in obviously dominant strategies in problems where that rule is OSP-implementable. In the absence of OSP-implementations, our optimality notion provides a promising solution that complies with obvious dominance whenever possible and incentivizes agents by minimizing decisions made.

We derive an optimal sequential implementation of TTC which only asks agents about top choices. However, we show that the two conditions for optimality might contradict each other in DA implementations. We thus propose a weaker optimality notion that while prioritizing obvious decisions, minimizes the amount of information elicited through non-obvious decisions. At last, we introduce a sequential revelation game that weakly optimally implements DA under truthful behavior.

Our results may serve as a starting point for further works in various frameworks where OSP-implementations are absent. For instance, one possible direction is to explore models beyond the unit-supply case. Also, it is worthwhile to investigate alternative notions of optimality that exist for arbitrary strategy-proof rules. As we introduce the (weakly) optimal implementations of TTC and DA through different algorithms, it would be interesting to design an algorithm that derives (weakly) optimal implementations for any strategy-proof rule.

3.A TTC AND DA

In this section, we give the formal definitions of TTC and agent-proposing DA studied in this chapter. The definition of TTC given below follows from Abdulkadiroğlu and Sönmez (2003). For any preference profile $\succeq \in \mathcal{P}$, TTC yields a matching via the following algorithm:

Step 1 Each agent $i \in I$ points to her favorite object (or herself) according to $\succeq_i$, and each object $o \in O$ points to the agent who has the highest priority on $\rhd_o$. Since I and O are finite, there is at least one **top trading cycle** $\{i_1, o_1, i_2, \ldots, i_k, o_k, i_1\}$ such that i_1 points to o_1, o_1 points to i_2, ..., and o_k points to i_1 while all element in this cycle are distinct. Remove all such cycles from the system by assigning each agent in these cycles to the object she points to.[17] Denote the remaining agents by I_1 and the remaining objects by O_1. If $I_1 \neq \emptyset$ and $O_1 \neq \emptyset$, move to Step 2; otherwise end the algorithm.

Step k, $k \geq 2$ Each agent $i \in I_{k-1}$ points to her favorite object in O_{k-1} (or herself) according to $\succeq_i$, and each object $o \in O_{k-1}$ points to the agent who has the highest priority among all agents in I_{k-1}. There is at least one top trading cycle. Remove all cycles from the system by assigning each agent in these cycles to the object she points to. Denote the remaining agents by I_k and the remaining objects by O_k. If $I_k \neq \emptyset$ and $O_k \neq \emptyset$, move to Step $k+1$; otherwise end the algorithm.

The algorithm terminates when no agent or no object left. The resulting matching is the collection of the assigned pairs at each step and the unassigned agents/objects after the last step.

Next, we induce the agent-proposing DA due to Gale and Shapley (1962). For any preference profile $\succeq \in \mathcal{P}$, the algorithm that computes the outcome of DA works as follows:

[17] Note that $\{i, i\}$ is also a top trading cycle, removing such cycle means assigning agent i to herself. This statement also applies to the following rounds.

Step 1 Each agent $i \in I$ proposes to her most preferred object (or herself) in $O \cup \{i\}$. Each object $o \in O$ considers all the proposals and tentatively accepts the candidates who apply to o with the highest priority at that object. The remaining proposals are rejected. Moreover, all agents that propose to themselves are regarded as accepted and assigned alone.

Step k, $k \geq 2$ Each agent who was rejected at step $k-1$ applies to her most preferred object (or herself) not yet applied to. Each object $o \in O$ considers all the new applicants together with the tentatively assigned agent at step $k-1$. Each object o now tentatively accepts the highest ranked applicant and rejects all others. Moreover, all agents that propose to themselves are regarded as accepted and assigned alone.

The algorithm terminates at the first step when no agent is rejected. The matching outcome is the tentative assignments at that step.

3.B Examples

In this section, we present two examples that illustrate the non-obvious decisions an optimal implementation of DA would avoid. In both examples, we consider a simplified problem where all preferences are full, and our insight can be easily generalized to problems where preferences are not full.

Example 3.4. There are three agents $I = \{i_1, i_2, i_3\}$ and three objects $O = \{o_1, o_2, o_3\}$. The priority structure $\rhd_O$ is given by the following table.

$\rhd_{o_1}$	$\rhd_{o_2}$	$\rhd_{o_3}$
i_1	i_1	i_2
i_3	i_3	i_1
i_2	i_2	i_3

We consider the following two implementations Γ, Γ' of DA. At the beginning, both Γ and Γ' ask i_1 (i_2) whether among $\{o_1, o_2, o_3\}$, her top choice is in $\{o_1, o_2\}$ ($\{o_3\}$). Consider the continuation of the games where both agents' answers are "No". Note that when i_1 answers "No", it implicitly means that she prefers o_3 most. Thus, candidates for the next player are i_2 and i_3.

Suppose that Γ picks i_2 to make decisions then. That is, i_2 is called on to choose from $\{o_1, o_2\}$. Notably, no matter what i_2 reports, it will not add any useful information to i_1 and i_3. Thus, at the next step, Γ needs to ask i_3 to choose from $\{o_1, o_2\}$.

Suppose that Γ' picks i_3 to make decisions after the "No" answers from i_1 and i_2. That is, Γ' asks i_3 to choose from $\{o_1, o_2\}$. Notably, the following part of the game will be exactly the same as Γ. However, Γ' induces a coarser partition than Γ in terms of i_2's decisions.

In conclusion, Γ is not optimal. More specifically, if i_1 and i_2 have revealed that their top choices are not the objects for which they have the highest priority, then i_2's revelation is irrelevant since how the assignments are determined depends on i_3's decision.

Intuitively, this indicates that an optimal implementation must avoid picking agents whose non-obvious decisions will not cause any other agent to reveal less information.

The next example explains that after selecting agents whose non-obvious decisions are informative, an optimal implementation should further care about the "right" amount of revelation from them.

Example 3.5. There are four agents $I = \{i_1, i_2, i_3, i_4\}$ and four objects $O = \{o_1, o_2, o_3, o_4\}$. The priority structure $\rhd_O$ is given by the following table.

$\rhd_{o_1}$	$\rhd_{o_2}$	$\rhd_{o_3}$	$\rhd_{o_4}$
i_1	i_2	i_1	i_2
i_3	i_3	i_2	i_1
i_4	i_4	i_4	i_4
i_2	i_1	i_3	i_3

Let $\succeq_{i_1}$: $o_2, o_3, \ldots$ and $\succeq_{i_2}$: $o_1, o_4, \ldots$. Suppose that both i_1 and i_2 tell the truth.

Consider an implementation Γ of DA that reaches i_3 after both i_1 and i_2 have revealed that their top choices are not the objects for which they have the highest priority. Specifically, Γ asks i_3 to choose her favorite object from $\{o_1, o_2, o_3, o_4\}$. However, if i_3 reports o_3 or o_4, the following game will be exactly the same as she only reveals that her top choice is in $\{o_3, o_4\}$. That is, revealing the exact top choice from $\{o_3, o_4\}$ will not ease the decisions of the next player. Moreover, only for some of the possible scenarios, it is necessary to know the exact top choice of i_3. For instance, if $\succeq_{i_4}$: $o_1, \ldots$, it is not necessary to know; if $\succeq_{i_4}$: $o_3, \ldots$, it is then necessary to know. Notably, even when it is necessary to know, it will not hurt others if we ask i_3 to reveal (which of o_3 and o_4 is her top choice) after i_4 reports her top choice.

In conclusion, the information about i_3's top choice between $\{o_3, o_4\}$ is only valuable for some realizations of the game. Moreover, it will not cause any agent to reveal more information if we ask i_3 to reveal her top choice between o_3 and o_4 later in this game when it becomes necessary.

Loosely speaking, when there are no obvious decisions for any agent, then it is unfair for the next player since she has to reveal information via non-obvious decisions. Our optimality notion minimizes such non-obvious revelations in two folds: First, it avoids selecting agents whose decisions are not informative at that stage. Second, it avoids asking active agents to reveal information more than necessary at that stage.

3.C Proof of Theorem 3.1

Denote TTC by f^T. Since we only consider the matching rule f^T throughout the proof, we refer to in the following $O_i(\tilde{\mathcal{P}}, f^T)$ as $O_i(\tilde{\mathcal{P}})$ for any $i \in I$ and $\tilde{\mathcal{P}} \in \mathcal{P}$. By construction Γ^T implements f^T under truthfully behavior. Thus, it suffices to show that Γ^T satisfies the condition 2.(a) and 2.(b) of Definition 3.5.

Fix an arbitrary node n of Γ^T. Let $i \in I$ denote the agent being called to take actions at n and let the remaining preference profiles at n be $\mathcal{P}^n$. Denote the decision i made at n by $\mathbb{P}_i^n$. Before formally proving the result, we first consider the partition of $\mathcal{P}^n$ induced by Γ^T, which we denote by $\mathbb{P}^{\Gamma_n^T}$ hereafter. Specifically, if a partition $\tilde{\mathbb{P}}$ of $\mathcal{P}^n$ is coarser than $\mathbb{P}^{\Gamma_n^T}$, then at least two elements $\hat{\mathcal{P}}^1, \hat{\mathcal{P}}^2 \in \mathbb{P}^{\Gamma_n^T}$ are contained in the same cell of $\tilde{\mathbb{P}}$. Accordingly, there must be (at least) one agent k whose decision $\mathbb{P}_k'$ in the subgame Γ_n^T causes $\hat{\mathcal{P}}^1, \hat{\mathcal{P}}^2$ to be apart. In this case, we say that the coarser part of $\tilde{\mathbb{P}}$ compared to $\mathbb{P}^{\Gamma_n^T}$, namely the union of $\hat{\mathcal{P}}^1$ and $\hat{\mathcal{P}}^2$, is caused by $\mathbb{P}_k'$ (a formal definition is presented in Definition 3.7).

Next, select an arbitrary extensive-form revelation game $\tilde{\Gamma}$ on $\mathcal{P}^n$ such that $\mathbb{P}^{\tilde{\Gamma}}$ implements f^T on $\mathcal{P}^n$ and $\mathbb{P}^{\tilde{\Gamma}}$ is weakly coarser than $\mathbb{P}^{\Gamma_n^T}$. In the following, we show that

(A) If $\mathbb{P}_i^n$ is not obvious, then there is no obvious decision at $\mathcal{P}^n$; and

(B) $\mathbb{P}^{\tilde{\Gamma}}$ cannot be coarser than $\mathbb{P}^{\Gamma_n^T}$ caused by the decision $\mathbb{P}_i^n$.

Note that since n is arbitrarily taken, showing statement (A) above is equivalent to showing condition 2.(a) of Definition 3.5 for Γ^T. Moreover, after we show statement (B), we will use it inductively to show that condition 2.(b) of Definition 3.5 also holds for Γ^T.

Now, we formally prove statement (A) and (B) in three stages separately.

Stage 1 Suppose that n is a node where Stage 1 is reached. Note that the set S_i at node n contains all options which i is secured to receive when she reports as her top choice. Thus, we have $S_i(\mathcal{P}^n) \subseteq S_i$. Let $S_i(\mathcal{P}^n) = \{o^1, \ldots, o^K\}$, then the decision provided by Γ^T to agent i at node n can be represented as $\mathbb{P}_i^n = (\mathcal{P}_i^{O_i(\mathcal{P}^n)\setminus S_i}, \mathcal{P}_i^{o^1}, \ldots, \mathcal{P}_i^{o^K})$, where $\mathcal{P}_i^{O'}$ denotes the set of all preference relations from $\mathcal{P}_i^n$ in which the top choice among $O_i(\mathcal{P}^n)$ belongs to $O' \subseteq O_i(\mathcal{P}^n)$. According to Proposition 3.1, the decisions $\mathbb{P}_i^n$ is obvious for i at $\mathcal{P}^n$. Therefore, we do not need to consider condition 2.(a), namely statement (A), at such node n.

It remains to be shown that statement (B) holds, and we proceed by contradiction. Suppose that there exist $\tilde{\mathcal{P}}^s \in \mathbb{P}^{\tilde{\Gamma}}$ and $\succeq, \succeq' \in \tilde{\mathcal{P}}^s$ such that $\succeq_i, \succeq_i'$ belong to different cells of $\mathbb{P}_i^n$. Then, we distinguish the following two cases.

- **Case 1:** $\succeq_{\mathbf{i}} \in \mathcal{P}_{\mathbf{i}}^{\mathbf{p}}, \succeq_{\mathbf{i}}' \in \mathcal{P}_{\mathbf{i}}^{\mathbf{q}}$ **for distinct** $\mathbf{p}, \mathbf{q} \in \mathbf{S_i}(\mathcal{P}^{\mathbf{n}})$.

  Since $p, q \in S_i(\mathcal{P}^n)$ and $\mathbb{P}^{\tilde{\Gamma}}$ is a partition of $\mathcal{P}^n$, it follows immediately that $f_i^T(\succeq_i, \tilde{\succeq}_{-i}) = p$ and $f_i^T(\succeq_i', \tilde{\succeq}_{-i}) = q$ for any $\tilde{\succeq}_{-i} \in \tilde{\mathcal{P}}_{-i}^s$. Recall that we consider Cartesian subsets. It then follows that $(\succeq_i, \succeq_{-i}), (\succeq_i', \succeq_{-i}) \in \tilde{\mathcal{P}}^s$. This implies that $\mathbb{P}^{\tilde{\Gamma}}$ does not implements f^T on $\mathcal{P}^n$ and we reach a contradiction.

- **Case 2:** $\succeq_{\mathbf{i}} \in \mathcal{P}_{\mathbf{i}}^{\mathbf{p}}$ **for some** $\mathbf{p} \in \mathbf{S_i}(\mathcal{P}^{\mathbf{n}})$ **and** $\succeq_{\mathbf{i}}' \in \mathcal{P}_{\mathbf{i}}^{\mathbf{O_i}(\mathcal{P}^{\mathbf{n}}) \setminus \mathbf{S_i}}$.

  Note that we can just relabel the two profiles if $\succeq_i, \succeq_i'$ are the other way around. We introduce the following necessary notation. Let $top(\succeq_i' |_{O_i(\mathcal{P}^n)}) = o$ for some $o \in O_i(\mathcal{P}^n)$. Also, fix any $\tilde{\succeq}_{-i} \in \tilde{\mathcal{P}}_{-i}^s$ and let $k^* \in I$ be such that $f_{k^*}^T(\succeq_i', \tilde{\succeq}_{-i}) = o$. For ease of presentation denote $f^T(\succeq_i', \tilde{\succeq}_{-i}) = \mu$ hereafter. Let $\{k^*, o, k_1, o_1 \ldots, k_t, o_t, k^*\}$ be the top priority cycle that assigns k^* to o in the matching μ and let $K = \{k^*, k_1, \ldots, k_t\}$ be the set of agents contained in that cycle. As a remark, K could be a singleton. Our strategy for Case 2 is as follow. First, we show that there exists $j \in K$ who has not revealed her ranking between p and μ_j at $\mathcal{P}^n$. Second, we show that $\tilde{\Gamma}$ asks such j to reveal more information than Γ_n^T does, with which we reach a contradiction to $\mathbb{P}^{\tilde{\Gamma}}$ being coarser than $\mathbb{P}^{\Gamma_n^T}$.

  We now show the first part, that is, we show that there exists $j \in K, \succ_j, \succ_j' \in \mathcal{P}^n$ such that $p \succeq_j \mu_j$ and $\mu_j \succeq_j' p$. We proceed by contradiction. First, assume that $p \succeq_k \mu_k$ for all $k \in K$ and all $\succeq_k \in \mathcal{P}_k^n$. Notably, since $p \in S_i(\mathcal{P}^n)$, it holds $p \notin S_k(\mathcal{P}^n)$ for any $k \in K$. Therefore, if k has revealed that she prefers p to μ_k, she must have pointed to some agent in I_i (who only owns p) in Stage 3 before n. In this case, however, we have $p \tilde{\succeq}_k \mu_k$ and we should find the cycle $\{i, o, k_1, p, \ldots, i\}$ when calculate $f^T(\succeq_i', \tilde{\succeq}_{-i})$. This contradicts to $\mu_i = p$. Next, assume that

$\mu_k \succeq_k p$ for all $k \in K$ and all $\succeq_k \in \mathcal{P}_k^n$. In this case, each $k \in K$ has revealed that her top choice is μ_k before node n. Note that such revelation only happens in Stage 2, according to Γ^T, these agents are assigned and removed then. Recall that o is involved in the same cycle with K when calculate $f^T(\succeq_i', \tilde{\succeq}_{-i})$, o should also be removed being assigned to k_1 before n. However, this contradicts to $o \in O_i(\mathcal{P}^n)$.

As a result, there exists $j \in K$ who has not revealed her ranking between p and μ_j at n. Moreover, from above arguments we can infer that the underlying $\tilde{\succeq}_{-i} \in \tilde{\mathcal{P}}_{-i}^s$ satisfies that $\mu_j \tilde{\succeq}_j p$. Before starting the second part, we make the following construction. Construct $\hat{\succeq}_j \in \mathcal{P}_j$ such that (1) $p \hat{\succeq}_j \mu_j$ and; (2) $o' \hat{\succeq}_j o''$ if and only if $o' \tilde{\succeq}_j o''$ for all $o', o'' \in O \cup \{j\} \setminus \{p\}$. Notice that in Γ^T the preferences are elicited from the top down to the bottom and that j has not revealed her preference between p and μ_j at n, we know that $\hat{\succeq}_j \in \mathcal{P}_j^n$.

We now formally show the second part for Case 2. More concretely, we show that $(\succeq_i, \tilde{\succeq}_j, \tilde{\succeq}_{-i,j})$ and $(\succeq_i, \hat{\succeq}_j, \tilde{\succeq}_{-i,j})$ are contained in different cells of $\mathbb{P}^{\tilde{\Gamma}}$ while they are contained in the same cell of $\mathbb{P}^{\Gamma_n^T}$. The former part follows if $(\succeq_i, \hat{\succeq}_j, \tilde{\succeq}_{-i,j}) \notin \tilde{\mathcal{P}}^s$. As argued above, since $p \hat{\succeq}_j \mu_j$, we have $f_i^T(\succeq_i', \hat{\succeq}_j, \tilde{\succeq}_{-i,j}) = o \neq \mu_i$. Note that we consider Cartesian domains and that $\tilde{\Gamma}$ implements f^T on $\mathcal{P}^n$, this implies that $(\succeq_i, \hat{\succeq}_j, \tilde{\succeq}_{-i,j}) \notin \tilde{\mathcal{P}}^s$. Thus, it remains to be shown that $(\succeq_i, \tilde{\succeq}_j, \tilde{\succeq}_{-i,j})$ and $(\succeq_i, \hat{\succeq}_j, \tilde{\succeq}_{-i,j})$ belong to same cell of $\mathbb{P}^{\Gamma_n^T}$. Towards this goal, let n_p be the immediate successor of n in Γ^T such that $\mathcal{P}^{n_p} = \mathcal{P}_i^p \times \mathcal{P}_{-i}^n$. According to Γ^T, p is removed from the market (being assigned to i) at n_p and thus $p \notin O_j(\mathcal{P}^{n_p})$. Notice that Γ^T only asks players to reveal information about their preferences over objects that are still available to them, we can infer that in $\Gamma_{n_p}^T$, agent j does not reveal any information about how she ranks p on her preferences. Since $\tilde{\succeq}_j$ and $\hat{\succeq}_j$ only differ in p's ranking, if we let $\mathcal{P}^* \in \mathbb{P}^{\Gamma_{n_p}^T}$ be such that $(\succeq_i, \tilde{\succeq}_j, \tilde{\succeq}_{-i,j}) \in \mathcal{P}^*$, we must have $(\succeq_i, \hat{\succeq}_j, \tilde{\succeq}_{-i,j}) \in \mathcal{P}^*$. We reach the desired contradiction and this completes the proof for Case 2.

As a result, we can claim that no extensive-form game that implements f^T on $\mathcal{P}^n$ induces a coarser partition than Γ_n^T does – statement (B) is satisfied. This completes the proof for Stage 1.

Stage 2 Suppose that n is a node where Stage 2 is reached. Note that in Stage 2, i is directly assigned to the object that she reports as favorite. Therefore, the decision is obvious and we do not need to consider statement (A) for such node n. Let i' be the agent to whom i points in G and let $\{o^1, \ldots, o^K\}$ be the objects that i' owns at I. Then, the decision at n can be represented as $\mathbb{P}_i^n = (\mathcal{P}_i^{o^1}, \ldots, \mathcal{P}_i^{o^K})$. Notably, we can use exactly the same reasoning as we used for Case 1 in Stage 1 to conclude that statement (B) also holds here. This finishes the check for Stage 2.

Stage 3 Suppose that n is a node where Stage 3 is reached. Let

$$J = \{i \in I : i \text{ has an outgoing edge in } G\}$$

be the agents in I who have already revealed who owns their top choices (among O) and let

$$L = \{i \in I \setminus J : S_i \neq \{i\}\}$$

be the agents who are not in J and who own at least one object at I. We first claim $|L| \geq 2$. Specifically, if $|L| = 1$, for the only agent $l \in L$, S_l contains all remaining objects and herself. This implies that in Stage 1, agent l must reveal that her top choice is in S_l and that Stage 3 will not be reached. Thus, $|L| \geq 2$ in Stage 3. Since i is the player at node n, we have $i \in L$ and we write $L = \{i, l_1, \ldots, l_{|L|-1}\}$. Moreover, for each $m < |L|$, let $S_m^* = S_{l_m} \setminus \{l_m\}$ be the set of objects which directly or indirectly point to l_m at n. Then, the decision provided by Γ^T to i at n can be represented as $\mathbb{P}_i^n = (\mathcal{P}_i^{S_1^*}, \ldots, \mathcal{P}_i^{S_{|L|-1}^*})$. Notably, if $|L| = 2$, then $\mathbb{P}_i^n = (\mathcal{P}_i^{S_1^*})$, implying that no decision is required from i. Therefore, in the rest of the proof, we only need to check the two target statements for cases where $|L| \geq 3$.

We first show that statement (A) holds, and the argument follows from Troyan (2019). Concretely, if $|L| \geq 3$, the priority among the remaining agents violates the *weak acyclicity* condition[18]. According to the proof of Theorem 1 of Troyan (2019), there is no extensive-form game that implements f^T on $\mathcal{P}^n$ and that provides the first player with obvious decisions. That is, there is no obvious decision available at node n – statement (A) is satisfied.

In the rest of Stage 3, we show that statement (B) holds. We use the same strategy as we used in Case 2 of Stage 1. Concretely, we show that if compared to in Γ_n^T, i reveals less information in terms of the decision $\mathbb{P}_i^n$ in $\tilde{\Gamma}$, then some agent $l \in L$ must reveal more information in $\tilde{\Gamma}$ than she does in Γ_n^T. In the following, assume that there is $\tilde{\mathcal{P}}^s \in \mathbb{P}^{\tilde{\Gamma}}$ and $\succeq^1, \succeq^2 \in \tilde{\mathcal{P}}^s$ such that $\succeq_i^1 \in \mathcal{P}_i^{t'}$ and $\succeq_i^2 \in \mathcal{P}_i^{t''}$ for distinct $\mathcal{P}_i^{t'}, \mathcal{P}_i^{t''} \in \mathbb{P}_i^n$.

In this paragraph, we present some necessary notation. We denote $top(\succeq_i^1 |_{O_i(\mathcal{P}^n)}) = o_1$ and let $l_1 \in L$ be such that $o_1 \in S_{l_1}^*$. Similarly, we denote $top(\succeq_i^2 |_{O_i(\mathcal{P}^n)}) = o_2$ and let $l_2 \in L$ be such that $o_2 \in S_{l_2}^*$. By the structure of $\mathbb{P}_i^n$ it is clear $o_1 \neq o_2$ and $l_1 \neq l_2$. Since $\mathbb{P}^{\tilde{\Gamma}}$ implements f^T on $\mathcal{P}^n$, we must have $f^T(\succeq^1) = f^T(\succeq^2)$. Thus, there must be $\succeq \in \{\succeq^1, \succeq^2\}$ such that $f_i^T(\succeq) \neq top(\succeq_i |_{O_i(\mathcal{P}^n)})$. Suppose that (at least) $\succeq^1$ satisfies this inequality. Let $\{k_1, o_1, l_1, \ldots, k_t, o_t, k_1\}$ be the top priority cycle that involves o_1 in calculating $f^T(\succeq^1)$ and let $K_1 = \{l_1, k_1, \ldots, k_t\}$ be the set of agents contained in that cycle. Next, if $\succeq^2$ also satisfies $f_i^T(\succeq^2) \neq top(\succeq_i^2 |_{O_i(\mathcal{P}^n)})$, we define K_2 in a similar way. Otherwise, let $\{i, o_2, l_2 \ldots, k'_{t'}, o'_{t'}, i\}$ be the top priority cycle that involves i and o_2 in calculating $f^T(\succeq^2)$, and let $K_2 = \{k : k \in L \cap \{l_2, k'_1, \ldots k'_{t'}\}$ and $f_k^T(\succeq^2) \notin S_i\}$.[19] Finally, Let $K = K_1 \cup K_2$. Notably, we will select the target l, who reveals more information in $\tilde{\Gamma}$ than in Γ_n^T, from the set K.

Next, we present a common feature for agents in K which is useful for selecting the target agent. That is, for each $k \in K$, she has not revealed her ranking between any two objects in $O_k(\mathcal{P}^n) \setminus S_k^*$ at

[18] As defined by Troyan (2019), a *strong cycle* in a priority structure is described by three agents $i_1, i_2, i_3 \in I$ and three objects $o_1, o_2, o_3 \in O$ such that $i_1 \rhd_{o_1} i_2, i_3$ and $i_2 \rhd_{o_2} i_1, i_3$ and $i_3 \rhd_{o_3} i_1, i_2$. If there are no strong cycles, the priority structure is said to be *weak acyclicity*.

[19] That is, K_2 contains the set of agents who are contained in both L and the underlying top priority cycle while who do not point to the object $o'_{t'} \in S_i$ that points to i at n. Notably, K_2 could be empty.

n. Note that since all agents in L have not revealed their rankings between any two available objects, it will be sufficient to show $K \subseteq L$. Also, since K_2 is either selected in the same way as K_1 or selected directly from L, we only need to show $K_1 \subseteq L$. Looking carefully at the description of Γ^T, we can find that for each $j \in J$ at node n, agent j must (in)directly point to i in G. Suppose by contradiction that $J \cap K_1 \neq \emptyset$, then one of the following two scenarios is realized. First, i is assigned to o_1 at $f^T(\succ^1)$, which clearly contradicts to $o_1 \succeq_i^1 f_i^T(\succeq^1)$. Second, o_1 has already been allocated at $\mathcal{P}^n$, which contradicts to $o_1 \in O_i(\mathcal{P}^n)$. Thus, we can then infer that $J \cap K_1 = \emptyset$ and thus $K_1 \subseteq L$. As a result, we have $K \subseteq L$ and thus all agents in K have not revealed their ranking between any two available objects. This implies that fix any $p \in S_i$ and any $k \in K$, there exists $\succeq_k \in \mathcal{P}_k^n$ such that $top(\succeq_k |_{O_k(\mathcal{P}^n)}) = p$.

We now select the target agent l from K, where the selection is based on the game $\tilde{\Gamma}$. In the rooted tree $\tilde{R}$ of $\tilde{\Gamma}$, we denote the root by $\tilde{r}$, denote the terminal node which corresponds to $\tilde{\mathcal{P}}^s$ by $s \in L(\tilde{R})$ and let $\tilde{\mathcal{P}}^{n'}$ be the set of remaining preference profiles at any non-terminal node $n' \in N(\tilde{R})$. Since $\tilde{\Gamma}$ implements f^T on $\mathcal{P}^n$, we have $\tilde{\mathcal{P}}^{\tilde{r}} = \mathcal{P}^n$. As has already been shown above, for all $k \in K$, there exists $\succeq_k \in \tilde{\mathcal{P}}_k^{\tilde{r}}$ such that $top(\succeq_k |_{O_k(\mathcal{P}^n)}) = p$. Moreover, we must have that $f_k^T(\succeq^1) \succeq_k' p$ for all $k \in K$ and all $\succeq_k' \in \tilde{\mathcal{P}}^s$ since otherwise i should be contained in the same top trading cycle with some agents in K. This implies that $top(\succeq_k' |_{O_k(\mathcal{P}^n)}) \neq p$ for all $\succeq_k' \in \tilde{\mathcal{P}}^s$. Therefore, along the path from $\tilde{r}$ to s in $\tilde{\Gamma}$, we can find the last node $\tilde{n}'$ where p is still a possible top choice for all agents in K. Specifically, denote the player at $\tilde{n}'$ by $l \in K$ and denote the immediate successor of $\tilde{n}'$ on the underlying path by $\tilde{n}$. Then, we have that (1) for all $k \in K$, there exists $\succeq_k \in \tilde{\mathcal{P}}_k^{\tilde{n}'}$ such that $top(\succeq_k |_{O_k(\mathcal{P}^n)}) = p$ and (2) for all $\succeq_l \in \tilde{\mathcal{P}}_l^{\tilde{n}}$, $top(\succeq_l |_{O_l(\mathcal{P}^n)}) \neq p$.

Finally, we show that l reveals more about her preference in $\tilde{\Gamma}$ than in Γ_n^T. To do so, select any $l_a \in \{l_1, l_2\} \setminus \{l\}$ with $a \in \{1, 2\}$. In Γ_n^T, let n_a be the successor of n such that $\mathcal{P}^{n_a} = \mathcal{P}_i^{S_a^*} \times \mathcal{P}_{-i}^n$. Moreover, let n_p be the immediate successor of n_a such that $\mathcal{P}^{n_p} = \mathcal{P}_i^{S_a^*} \times \mathcal{P}_{l_a}^p \times \mathcal{P}_{-i,l_a}^n$, where $\mathcal{P}_{l_a}^p$ denotes the set of preferences from $\mathcal{P}_{l_a}^n$ in which the top choice among $O_{l_a}(\mathcal{P}^n)$ is p. Since $l_a \neq l$, we know that at node $\tilde{n}$ in $\tilde{\Gamma}$, her top choice among $O_{l_a}(\mathcal{P}^n)$ could still be p. Thus, select any $\succeq \in \tilde{\mathcal{P}}^{\tilde{n}}$

such that $\succeq_i = \succeq_i^a$ and $top(\succeq_{l_a}|_{O_{l_a}(\mathcal{P}^n)}) = p$, and we know that $\succeq \in \mathcal{P}^{n_p}$. Next, let $\hat{\succeq}_l$ be such that (1) $top(\hat{\succeq}_l|_{O_l(\mathcal{P}^n)}) = p$ and; (2) $o' \hat{\succeq}_l o''$ if and only if $o' \succeq_l o''$ for all $o', o'' \in O \cup \{l\} \setminus \{p\}$. By construction it is obvious that $\hat{\succeq}_l \notin \tilde{\mathcal{P}}_l^{\tilde{n}}$, which implies $(\hat{\succeq}_l, \succeq_{-l}) \notin \tilde{\mathcal{P}}^{\tilde{n}}$. Since $\succeq \in \tilde{\mathcal{P}}^{\tilde{n}}$, we can infer that $\succeq$ and $(\hat{\succeq}_l, \succeq_{-l})$ cannot be in the same cell of $\mathbb{P}^{\tilde{\Gamma}}$. However, since p is removed at n_p in Γ^T, $\Gamma^T_{n_p}$ will never ask l to reveal her rankings about p. Note that as $\succeq_l, \hat{\succeq}_l \in \mathcal{P}_l^{n_a} = \mathcal{P}_l^n$ and $\succeq_l, \hat{\succeq}_l$ only differ in p's ranking, $\succeq$ and $(\hat{\succeq}_l, \succeq_{-l})$ must belong to the same cell of $\mathbb{P}^{\Gamma^T_n}$. This completes the proof for statement (B) in Stage 3.

So far, we have shown statement (A) and statement (B). Since n is arbitrarily taken, statement (A) being true indicates that condition 2.(a) of Definition 3.5 is satisfied by Γ^T. It remains to be shown that condition 2.(b) of Definition 3.5 is also satisfied by Γ^T. Towards this goal, we show that $\tilde{\Gamma}$ induces the same partition as Γ^T_n does. Specifically, according to statement (B) and the assumption that $\mathbb{P}^{\tilde{\Gamma}}$ is weakly coarser than $\mathbb{P}^{\Gamma^T_n}$, $\tilde{\Gamma}$ must pick the same decision at its root $\tilde{r}$ as Γ^T_n picks at n. That is, $\tilde{r}$ in $\tilde{\Gamma}$ and n in Γ^T_n have the same number of immediate successors and for each immediate successor $\tilde{n}$ of $\tilde{r}$ in $\tilde{\Gamma}$, there exists an immediate successor n' of n in Γ^T_n such that the remaining preference profiles at these two nodes are the same. Notably, this ensures us to use the above arguments for statement (B) to $\tilde{n}$ and n', which leads to the result that at $\tilde{n}$, $\tilde{\Gamma}$ can only pick the same decision as Γ^T_n picks at n'. By inductively applying such argument to each node in Γ^T_n, we reach the conclusion that $\mathbb{P}^{\tilde{\Gamma}}$ must be the same as $\mathbb{P}^{\Gamma^T_n}$. Since $\tilde{\Gamma}$ is arbitrarily taken, we conclude that no other implementation of f^T on $\mathcal{P}^n$ induces a coarser partition than $\mathbb{P}^{\Gamma^T_n}$. Finally, since n is arbitrarily taken, this implies that condition 2.(b) of Definition 3.5 is satisfied by Γ^T. This completes the proof.

4

Partition Obviously Strategy-Proof Rules*

Obviously strategy-proof (OSP) mechanisms (Li, 2017) provide appealing incentive properties. However, it is rarely possible to implement strategy-proof matching rules via OSP mechanisms. In the context of one-to-one object allocation markets, this chapter studies the incentive criterion *partition obvious strategy-proofness (POSP)* due to Zhang and Levin (2017), which is weaker than OSP but stronger than strategy-proofness. Similar to OSP, this criterion takes agents' limited reasoning abilities into consideration. For implementations of strategy-proof rules via static games, I introduce a

*This chapter is based on Chen (2021). I am grateful to Alexander Westkamp, Christoph Schottmüller, Markus Möller and Marius Gramb for their insightful suggestions and comments. This chapter also benefits from comments made by the seminar participants at the University of Cologne.

self-invariance condition that is both necessary and sufficient for such implementations to be POSP. However, this result does not hold when extensive-form implementations of strategy-proof rules are considered.

4.1 Introduction

In centralized object allocation markets, one desirable criterion when designing the matching rules is strategy-proofness, which guarantees that it is in agents' best interest to report their preferences truthfully.[1] Nevertheless, empirical and experimental studies on strategy-proof rules (e.g., Chen and Sönmez (2006); Hassidim et al. (2017); Shorrer and Sóvágó (2018)) suggest that in practice, a significant fraction of agents deviate from being truthful. These findings indicate that the incentives provided by strategy-proof rules might not be straightforward to some agents.

Strengthening strategy-proofness, Li (2017) introduces *obvious strategy-proofness (OSP)* in his seminal paper. Loosely speaking, a mechanism is OSP if at any information set, the worst possible outcome from telling the truth is weakly better than the best possible outcome following any deviation. Li (2017) shows that the strategic incentives induced by OSP mechanisms can be understood even by agents who are unable to engage in any contingent reasoning.[2] However, popular strategy-proof matching rules, such as *Deferred Acceptance (DA)* and *Top Trading Cycles (TTC)*, are not OSP implementable in general (Li, 2017; Ashlagi and Gonczarowski, 2018; Troyan, 2019). Thus, a certain degree of contingent reasoning is essential for an agent to understand the incentives provided by these strategy-proof rules. In this chapter, I study degrees of contingent reasoning needed for agents' understanding of strategy-proofess, where the degree is interpreted by a partition-based incentive criterion

[1] Abdulkadiroğlu et al. (2005) shows that in Boston, the local committee has replaced the manipulable Boston mechanism by the remarkable deferred acceptance (DA) mechanism for the school choice program. Pathak and Sönmez (2008) argue that the Boston mechanism, which is not strategy-proof, may harm naive agents who are not good at strategizing.

[2] Generally speaking, contingent reasoning is the ability to reason state-by-state about all hypothetical scenarios.

that is weaker than OSP.

I set up a model with limited contingent reasoning in a standard one-to-one object allocation problem. Concretely, I assume that agents could partition all states of the world into different events and contingently reason event-by-event. Zhang and Levin (2017) provide decision-theoretic foundations for obvious dominance in such a setting and define a strategy to be *partition dominant* if it is optimal in each event of that partition. For the interest of this chapter, I focus on *partition obvious strategy-proofness (POSP)*, a concept extracted from partition dominance with acting truthfully being the optimal strategy. In particular, fix an agent and given her partition of all possible preferences of other agents, a rule is partition obviously strategy-proof for her if *within each event* of that partition, the worst possible outcome from telling the truth is weakly better than the best possible outcome from any deviation. The incentive provided by a POSP rule can be understood by agents with limited reasoning abilities featured by the underlying partitions. Note that our model nests no contingent reasoning and full contingent reasoning as two extreme cases, thus, the incentive provided by POSP is located between those provided by OSP and strategy-proofness.

I define a partition to be *self-invariant* under a matching rule for an agent if fix any that agent's reported preference, her assignment is unambiguously determined within each event of that partition. For any strategy-proof rule, I show that the self-invariance condition on partitions is both necessary and sufficient for the static implementation of that rule to be POSP (in the sense of the underlying partitions). That is, an agent can understand the incentive provided by a strategy-proof rule if and only if her reasoning ability ensures her to figure out what she exactly gets in each event. Moreover, under any strategy-proof rule and fix any agent, I find that there exists a unique lowest degree of reasoning ability for her to understand that being truthful is optimal, and I develop an algorithm that yields such coarsest self-invariant partition.

Furthermore, I study POSP in extensive-form settings and find that the self-invariant partitions are not necessary for POSP of sequential implementations of strategy-proof rules. The key drivers

regarding this result are that in extensive-form games, agents' decisions of reporting preferences are decomposed into several small parts and they receive information on others' preferences from the system as game goes on. These features help an agent stick to the optimal strategy even with contingent reasoning specified by a partition that is coarser than the coarsest self-invariant partition under that matching rule.

Related Literature The present chapter mostly relates to Zhang and Levin (2017). In a general domain, they provide an axiomatic approach to account for agents' deficiencies in reasoning and classify a broad family of mechanisms which overcome such deficiencies and which are verified to be useful in the laboratory. This chapter is an application of their approach to matching markets and I restrict attention to truthful behaviors. Moreover, this chapter provides results on how strategy-proof rules are related to POSP, which is not a focus in Zhang and Levin (2017).

This work also relates to the growing line of research that is based on Li (2017)'s obvious strategy-proofness, which has already discussed in the related literature of Chapter 3. As a review, Pycia and Troyan (2021) introduce a family of simplicity standards and characterize simple mechanisms in broad domains under their *richness* assumption. In terms of obvious strategy-proofness, they introduce *millipede games* that characterize general OSP mechanisms without monetary transfers. Troyan (2019) focuses on TTC and characterizes all priorities that ensure TTC to be implementable in obviously dominant strategies. Ashlagi and Gonczarowski (2018) show that stable mechanisms (including DA) are only OSP-implementable in very restrictive environments, Thomas (2021) follows their study characterizing all such environments that make DA OSP-implementable. Bade and Gonczarowski (2016) study OSP-implementations of Pareto-efficient social choice rules in various domains. This chapter complements the prior works on OSP mechanisms since I provide potential ideas about how to train agents' reasoning abilities such that they conduct fewer deviations in implementations of strategy-proof rules that are not OSP.

Finally, this chapter is also in line with laboratory results (Echenique et al., 2016; Klijn et al., 2019; Bó and Hakimov, 2020a; Breitmoser and Schweighofer-Kodritsch, 2021) which observe higher rates of truth-telling in sequential implementations of strategy-proof rules compared to those static counterparts.

The rest of this chapter is organized as follows. Section 4.2 introduces the model, the definition of POSP, and the basic features of POSP rules. Section 4.3 studies the relationship between POSP and strategy-proof matching rules, including the main result. Section 4.4 shows that the main result does not hold in extensive-form settings. Section 4.5 concludes.

4.2 Model

This chapter employs the same notation and considers the same allocation problem $(I, O, \rhd_O, \succeq)$ as we defined in Chapter 3. Recall that

- I is a finite set of agents,
- O is a finite set of indivisible objects,
- $\rhd_O = (\rhd_o)_{o \in O}$ is a *priority structure*, where for each $o \in O$, $\rhd_o$ is a strict priority ordering of $I \cup \{o\}$, and
- $\succeq = (\succeq_i)_{i \in I}$ is a *preference profile*, where for each $i \in I$, $\succeq_i$ is a strict preference ranking of $O \cup \{i\}$.

For the following discussion, given a preference profile $\succeq$, we refer to $\succeq_i$ as the *type* of agent i. Moreover, I refer to the tuple $(I, O, \rhd_O, \mathcal{P})$ as a *market*.

4.2.1 Partition obvious strategy-proofness

In this subsection, I introduce partition obvious strategy-proofness in the context of object allocations. For the following discussion, fix a problem $(I, O, \rhd_O, \succeq)$ and a rule f. I first define the parti-

tion system, which I adapt from Zhang and Levin (2017) and is useful to set up a model of how agents reason in the game.

Definition 4.1. Let $\Sigma = \{\Sigma_i\}_{i\in I}$ be a *partition system* where each $\Sigma_i = \{\mathcal{P}^t_{-i}\}_{t=1}^T$ is a finite disjoint partition of $\mathcal{P}_{-i}$.

I next introduce a class of incentive properties built on partition systems. Inspired by Li (2017), the concept below requires truth-telling to stand out in the best-worst case comparison. Notably, I restrict attention to static games for now and the following definition applies exclusively to normal-form settings.[3]

Definition 4.2. Given a partition system $\Sigma = \{\Sigma_i\}_{i\in I}$, we say that

1. f is Σ_i-*obviously strategy-proof* (Σ_i-OSP) for agent i if for all $\mathcal{P}^t_{-i} \in \Sigma_i$ and all distinct $\succeq_i$, $\succeq_i' \in \mathcal{P}_i$,

$$\inf_{\succeq_{-i}\in\mathcal{P}^t_{-i}} f_i(\succeq_i, \succeq_{-i}) \succeq_i \sup_{\succeq_{-i}\in\mathcal{P}^t_{-i}} f_i(\succeq_i', \succeq_{-i})$$

   where inf and sup represent the worst and best possible assignments for i with respect to $\succeq_i$ respectively.

2. f is Σ-*obviously strategy-proof* (Σ-OSP) if f is Σ_i-OSP for all i.

In words, if a matching rule is Σ_i-OSP for agent i, then given any i's type, when she only considers types of others within one (and any) element of Σ_i, the worst possible outcome following truth-telling is no worse than the best possible outcome following any misreporting strategy. We say that f is *partition obviously strategy-proof (POSP)* if there exists some partition system Σ such that f is Σ-OSP.

For each agent i, each type profile $\succeq_{-i} \in \mathcal{P}_{-i}$ is a possible scenario that could realize and each element $\mathcal{P}^t_{-i} \in \Sigma_i$ represents a possible *event* that contains a series of scenarios. In Definition 4.2, I

[3] POSP also applies to extensive-form games, and the formal definition of POSP in extensive-form games can be found in Appendix 4.4

use a partition system Σ to describe the *reasoning abilities* of agents in I. Concretely, if the reasoning ability of an agent $i \in I$ is specified by Σ_i, then for any $\succeq_i \in \mathcal{P}_i$ and any event $\mathcal{P}^t_{-i} \in \Sigma_i$, i can figure out

$$\mathcal{M}(\succeq_i, \mathcal{P}^t_{-i}) = \{\mu \in \mathcal{M} \mid \exists \succeq_{-i} \in \mathcal{P}^t_{-i} \text{ s.t. } f(\succeq_i, \succeq_{-i}) = \mu\},$$

the set of matchings that could be potentially realized when she submits $\succeq_i$ to f within the event $\mathcal{P}^t_{-i}$. However, once $\mathcal{M}(\succeq_i, \mathcal{P}^t_{-i})$ is not a singleton, for any type profile $\succeq_{-i} \in \mathcal{P}^t_{-i}$, she is not able to calculate the exact matching $f(\succeq_i, \succeq_{-i})$. In other words, when an agent has deficient reasoning ability, she can only partition all possible scenarios into different events and reason event-by-event, but not scenario-by-scenario within each event. The limited reasoning ability of an agent can be interpreted as her bounded understanding about the relation between the resulting matchings and possible scenarios. The coarser the partition is, the more difficult contingent reasoning becomes.

Note that POSP bridges the well-studied strategy-proofness and obvious strategy-proofness as two extreme cases of reasoning abilities. Concretely, when a partition $\hat{\Sigma}_i$ is the coarsest, i.e., when it holds that $\hat{\Sigma}_i = \{\mathcal{P}_{-i}\}$, f is said to be *obviously strategy-proof (OSP)* for i when f is $\hat{\Sigma}_i$-OSP for i.[4] On the contrary, when a partition $\bar{\Sigma}_i$ is the finest, i.e., when each $\mathcal{P}^t_i \in \bar{\Sigma}_i$ is a singleton, f being $\bar{\Sigma}$-OSP is equivalent to f being strategy-proof.

4.2.2 Features of POSP rules

Next, I provide some general features of partition obviously strategy-proof matching rules which turn out to be useful for my analysis.

We first investigate the relationship between partition obvious strategy-proofness and obvious strategy-proofness. Fix a partition system Σ. For each $\Sigma_i = \{\mathcal{P}^t_{-i}\}_{t=1}^T$ and each t, let $\mathcal{P}^{t,i} = \mathcal{P}_i \times \mathcal{P}^t_{-i}$ denote the subset of $\mathcal{P}$ such that $\succeq \in \mathcal{P}^{t,i}$ if and only if $\succeq_{-i} \in \mathcal{P}^t_{-i}$. In addition, for any $\hat{\mathcal{P}} \subseteq \mathcal{P}$,

[4] The definition of OSP provided here is a simplification of Li (2017)'s formal definition, and it only applies to normal-form settings.

we say that f is OSP on $\hat{\mathcal{P}}$ for i if given any distinct $\succeq_i, \succeq_i' \in \hat{\mathcal{P}}_i$, it holds that with respect to $\succeq_i$, $\inf_{\succeq_{-i} \in \hat{\mathcal{P}}^t_{-i}} f_i(\succeq_i, \succeq_{-i}) \succeq_i \sup_{\succeq_{-i} \in \hat{\mathcal{P}}^t_{-i}} f_i(\succeq_i', \succeq_{-i})$. The next result shows that if f is Σ_i-OSP for i, then truth-telling is an obvious dominant strategy for i in each event of Σ_i.

Lemma 4.1. *If f is Σ-OSP, then for each i and each t such that $\mathcal{P}^t_{-i} \in \Sigma_i$, f is OSP for i on $\mathcal{P}^{t,i}$.*

Proof. The result follows immediately from Definition 4.2. □

Fix any $i \in I$ and any two partitions Σ_i and $\tilde{\Sigma}_i$. We say that $\tilde{\Sigma}_i$ is a decomposition of Σ_i if for each $\tilde{\mathcal{P}}'_{-i} \in \tilde{\Sigma}_i$, there exists $\mathcal{P}'_{-i} \in \Sigma_i$ such that $\tilde{\mathcal{P}}'_{-i} \subseteq \mathcal{P}'_{-i}$, where at least one inclusion is strict. Accordingly, a partition system $\tilde{\Sigma}$ is a decomposition of another partition system Σ if $\tilde{\Sigma}_i$ decomposes Σ_i for each $i \in I$.

Li (2017) shows that any obviously strategy-proof mechanism is also strategy-proof. The next result generalizes this statement to any two POSP matching rules where one of the two underlying partition systems is a decomposition of the other.

Lemma 4.2. *Let Σ and $\tilde{\Sigma}$ be two partition systems such that $\tilde{\Sigma}$ is a decomposition of Σ. If f is Σ-OSP, then f is $\tilde{\Sigma}$-OSP.*

Proof. If f is Σ-OSP, then $\inf_{\succeq_{-i} \in \mathcal{P}^t_{-i}} f_i(\succeq_i, \succeq_{-i}) \succeq_i \sup_{\succeq_{-i} \in \mathcal{P}^t_{-i}} f_i(\succeq_i', \succeq_{-i})$ for all $i \in I$, all $\mathcal{P}^t_{-i} \in \Sigma_i$ and all $\succeq_i, \succeq_i' \in \mathcal{P}_i$.

From now on, fix any $i \in I$, any $\tilde{\mathcal{P}}^{\tilde{t}}_{-i} \in \tilde{\Sigma}_i$ and any $\succeq_i \in \mathcal{P}_i$ as i's true type. First, consider that i submits her true preference $\succeq_i$ to f. Note that as $\tilde{\Sigma}$ is a decomposition of Σ, there is an event $\mathcal{P}^t_{-i} \in \Sigma$ such that $\tilde{\mathcal{P}}^{\tilde{t}}_{-i} \subseteq \mathcal{P}^t_{-i}$. This implies that

$$\cup_{\succeq_{-i} \in \tilde{\mathcal{P}}^{\tilde{t}}_{-i}} f_i(\succeq_i, \succeq_{-i}) \subseteq \cup_{\succeq_{-i} \in \mathcal{P}^t_{-i}} f_i(\succeq_i, \succeq_{-i}).$$

That is, with respect to $\succeq_i$, the worst outcome within $\cup_{\succeq_{-i} \in \tilde{\mathcal{P}}^{\tilde{t}}_{-i}} f_i(\succeq_i, \succeq_{-i})$ must be weakly better than the worst outcome within $\cup_{\succeq_{-i} \in \mathcal{P}^t_{-i}} f_i(\succeq_i, \succeq_{-i})$:

$$\inf_{\succeq_{-i}\in\tilde{\mathcal{P}}^{\tilde{t}}_{-i}} f_i(\succeq_i,\succeq_{-i}) \succeq_i \inf_{\succeq_{-i}\in\mathcal{P}^{t}_{-i}} f_i(\succeq_i,\succeq_{-i}).$$

Second, consider that i submits any misreport $\succeq'_i \in \mathcal{P}_i$ to f. In this case, a similar argument also applies to the best outcomes with respect to $\succeq_i$:

$$\sup_{\succeq_{-i}\in\mathcal{P}^{t}_{-i}} f_i(\succeq'_i,\succeq_{-i}) \succeq_i \sup_{\succeq_{-i}\in\tilde{\mathcal{P}}^{\tilde{t}}_{-i}} f_i(\succeq'_i,\succeq_{-i}).$$

Combining the above relations, we obtain

$$\inf_{\succeq_{-i}\in\tilde{\mathcal{P}}^{\tilde{t}}_{-i}} f_i(\succeq_i,\succeq_{-i}) \succeq_i \sup_{\succeq_{-i}\in\tilde{\mathcal{P}}^{\tilde{t}}_{-i}} f_i(\succeq'_i,\succeq_{-i}).$$

Since all i, $\succeq_i$, $\succeq'_{-i}$ and $\tilde{\mathcal{P}}^{\tilde{t}}_{-i}$ are arbitrarily taken, we reach the desired result that f is $\tilde{\Sigma}$-OSP. □

An immediate observation of Lemma 4.2 is that fix any Σ, if a matching rule is OSP, it is also Σ-OSP; moreover, if a matching rule is Σ-OSP, it is also strategy-proof.

We say that Σ provides stronger incentive than $\tilde{\Sigma}$ if whenever a rule is Σ-OSP, it is also $\tilde{\Sigma}$-OSP. Note that Lemma 4.2 is tight since in general, we cannot compare the incentives provided by two partition systems when none of them is a decomposition of the other. The following illustrative example shows that fix any Σ_i and $\tilde{\Sigma}_i$, if we only know that the number of events in Σ_i is less than the number of events in $\tilde{\Sigma}_i$, then f is not necessarily $\tilde{\Sigma}_i$-OSP when f is Σ-OSP.

Example 4.1. Consider a market with $I = \{i, j, k\}$ and $O = \{a, b, c\}$. Each agent $i' \in I$ has two possible types denoted by $\succeq^1_{i'}$ and $\succeq^2_{i'}$. Specifically, for agent i, let $c \succeq^1_i b \succeq^1_i a$ and $b \succeq^2_i a \succeq^2_i c$. Fix a matching rule f and in the following, I only present i's assignments under f since the remaining assignments are irrelevant for the discussion. Concretely, suppose that f assigns i to the following objects for each preference profile:

	$\succeq_i^1$	$\succeq_i^2$
$(\succeq_j^1, \succeq_k^1)$	c	a
$(\succeq_j^1, \succeq_k^2)$	b	b
$(\succeq_j^2, \succeq_k^1)$	c	a
$(\succeq_j^2, \succeq_k^2)$	c	b

Next, consider two partitions for agent i: $\Sigma_i = \{\mathcal{P}^1_{-i}, \mathcal{P}^2_{-i}\}$ and $\tilde{\Sigma}_i = \{\tilde{\mathcal{P}}^1_{-i}, \tilde{\mathcal{P}}^2_{-i}, \tilde{\mathcal{P}}^3_{-i}\}$ where the events in these two partitions are given as follows:

$$\mathcal{P}^1_{-i} = \{(\succeq_j^1, \succeq_k^1), (\succeq_j^2, \succeq_k^1)\} \ \mathcal{P}^2_{-i} = \{(\succeq_j^1, \succeq_k^2), (\succeq_j^2, \succeq_k^2)\}$$

$$\tilde{\mathcal{P}}^1_{-i} = \{(\succeq_j^1, \succeq_k^1)\} \ \tilde{\mathcal{P}}^2_{-i} = \{(\succeq_j^2, \succeq_k^2)\} \ \tilde{\mathcal{P}}^3_{-i} = \{(\succeq_j^1, \succeq_k^2), (\succeq_j^2, \succeq_k^1)\}$$

Note that f is Σ_i-OSP for agent i, since in both events $\mathcal{P}^1_{-i}$ and $\mathcal{P}^2_{-i}$ and for each type of i, the worst possible outcomes of truth-telling are at least as good as the best possible outcomes of misreporting. However, f is not $\tilde{\Sigma}_i$-OSP for agent i. To see this, it is enough to check for $\tilde{\mathcal{P}}^3_{-i}$. In this event, the possible outcomes for i are $\{b, c\}$ and $\{b, a\}$ when she reports $\succeq_i^1$ and $\succeq_i^2$, respectively. Since $b \succeq_i^2 a$, truth-telling is not obviously strategy-proof for i in this event. By Lemma 4.1, f is thus not $\tilde{\Sigma}_i$-OSP for i.

4.3 POSP and strategy-proof rules

In this section, I focus on the requirement a partition system Σ should fulfill such that a strategy-proof matching rule is also Σ-OSP. In other words, I study the levels of reasoning ability necessary for agents to understand that a matching rule is strategy-proof.

4.3.1 Self-invariant partition

Towards this goal, I first introduce a class of partitions which specify agents' reasoning abilities that minimize the uncertainty about their own assignments in each event.

Definition 4.3. A partition $\Sigma_i = \{\mathcal{P}^t_{-i}\}_{t=1}^T$ is *self-invariant under* f if for all $\succeq_i \in \mathcal{P}_i$, all $\mathcal{P}^t_{-i} \in \Sigma_i$ and all pairs $\succeq_{-i}, \succeq'_{-i} \in \mathcal{P}^t_{-i}, f_i(\succeq_i, \succeq_{-i}) = f_i(\succeq_i, \succeq'_{-i})$. A partition system Σ is self-invariant under f if each $\Sigma_i \in \Sigma$ is self-invariant under f.

In words, a partition Σ_i is self-invariant under a matching rule if given any i's preference, i's assignments under that rule are the same within each event of Σ_i. The idea behind Definition 4.3 is that when an agent has reasoning ability specified by a self-invariant partition, then she could reason to the degree such that her assignment is unambiguously determined in each event. That is, it only allows the underlying agent to have uncertainty about the assignments of other agents.

The next noteworthy example provides an intuition about how a self-invariant partition system Σ under f relates to Σ-OSP of f. For simplicity, I will argue in the example under the assumption that all objects are acceptable to all agents in $\mathcal{P}$ (i.e., that I only consider preference relations that are full). The generalization to preferences with unacceptable objects is straightforward.

Example 4.2. There are three agents $I = \{i, j, k\}$ and three objects $O = \{a, b, c\}$. Consider first a sub-domain of preference profiles $\tilde{\mathcal{P}} \subset \mathcal{P}$ that is given as follows:

$\succeq_i^1$	$\succeq_i^2$	$\succeq_i^3$	$\succeq_i^4$	$\succeq_i^5$	$\succeq_i^6$	$\succeq_j^1$	$\succeq_j^2$	$\succeq_k^1$	$\succeq_k^2$
a	a	b	b	c	c	a	a	a	b
b	c	a	c	a	b	b	c	b	a
c	b	c	a	b	a	c	b	c	c

The following table depicts the priority structure $\rhd_O$:

$\rhd_a$	$\rhd_b$	$\rhd_c$
i	j	k
j	k	i
k	i	j

Moreover, denote the Top Trading Cycles (TTC) matching rule by f^{TTC}. As has been shown in Abdulkadiroğlu and Sönmez (2003), f^{TTC} is strategy-proof.

The table below presents the assignments of agent i under f^{TTC} for each preference profile in $\tilde{\mathcal{P}}$. It is apparent from the table that given any of agent i's type, her assignment is determined in $\tilde{\mathcal{P}}_{-i}$. Since $\tilde{\mathcal{P}}_i = \mathcal{P}_i$, we can infer that the event $\tilde{\mathcal{P}}_{-i}$ is self-invariant under f^{TTC} for i.

	$\succeq_i^1$	$\succeq_i^2$	$\succeq_i^3$	$\succeq_i^4$	$\succeq_i^5$	$\succeq_i^6$
$(\succeq_j^1, \succeq_k^1)$	a	a	b	b	c	c
$(\succeq_j^1, \succeq_k^2)$	a	a	b	b	c	c
$(\succeq_j^2, \succeq_k^1)$	a	a	b	b	c	c
$(\succeq_j^2, \succeq_k^2)$	a	a	b	b	c	c

I now argue that f^{TTC} is OSP for i on $\tilde{\mathcal{P}}$. To see this, note that agent i is guaranteed to get the object that she reports as her favorite on $\tilde{\mathcal{P}}$. Therefore, for any type of agent i, the worst possible outcome under truth-telling is at least as good as the best possible outcome under any misrepresentation. Truth-telling is thus an obviously dominant strategy and f^{TTC} is OSP for i on $\tilde{\mathcal{P}}$.

Next, consider another sub-domain $\hat{\mathcal{P}}$ that comprises the same sets of preferences for i, k and that satisfies $\hat{\mathcal{P}}_j = \tilde{\mathcal{P}}_j \cup \{\succeq_j^3\}$ with $\succeq_j^3$: $c \succeq b \succeq a$. We list below the assignments for i under f^{TTC} given the type profile $(\succeq_j^3, \succeq_k^2)$ that is contained in $\hat{\mathcal{P}}$.

	$\succeq_i^1$	$\succeq_i^2$	$\succeq_i^3$	$\succeq_i^4$	$\succeq_i^5$	$\succeq_i^6$
$(\succeq_j^3, \succeq_k^2)$	a	a	a	a	a	a

By checking i's assignment when she submits $\succeq_i^3$, it is apparent that $\hat{\mathcal{P}}$ is not self-invariant under f^{TTC} for i. Moreover, in the sub-domain $\hat{\mathcal{P}}$ and given that the true type of i is $\succeq_i^3$, the worst possible outcome under truth-telling is being assigned to a. However, this assignment is strictly worse than the best possible outcome by misreporting $\succeq_i^4$ (i.e., getting b). Therefore, f^{TTC} is not OSP for i on $\hat{\mathcal{P}}$.

Finally, consider the full domain of preference profiles $\mathcal{P}$. I argue how to construct a partition Σ_i of $\mathcal{P}_{-i}$ such that f^{TTC} is Σ_i-OSP for agent i. Recall that f^{TTC} is OSP for i on $\tilde{\mathcal{P}}$ but not on $\hat{\mathcal{P}}$. According to Lemma 4.1, it is reasonable to have $\tilde{\mathcal{P}}_{-i}$ as an event in Σ_i. In fact, as I present in Appendix 4.A, if we further partition the remaining scenarios $\mathcal{P}_{-i} \setminus \tilde{\mathcal{P}}_{-i}$ such that each event is self-invariant under f^{TTC} for i, we reach the desired partition Σ_i.

Generalizing the findings in Example 4.2, the next and main result of this section shows that self-invariant partition systems under a strategy-proof rule are necessary and sufficient for that rule to be POSP.

Theorem 4.1. *A strategy-proof matching rule f is Σ-OSP if and only if Σ is self-invariant under f.*

Proof. I first show the "if" part, which follows directly from definitions. Select any $i \in I$, $\succeq_i, \succeq_i' \in \mathcal{P}_i$, $\mathcal{P}_{-i}^t \in \Sigma_i$ and $\succeq_{-i}^* \in \mathcal{P}_{-i}^t$. First, since the partition system Σ is self-invariant under f, it is true that with respect to $\succeq_i$,

$$\inf_{\succeq_{-i} \in \mathcal{P}_{-i}^t} f_i(\succeq_i, \succeq_{-i}) = \sup_{\succeq_{-i} \in \mathcal{P}_{-i}^t} f_i(\succeq_i, \succeq_{-i}) = f_i(\succeq_i, \succeq_{-i}^*)$$

and

$$\inf_{\succeq_{-i} \in \mathcal{P}_{-i}^t} f_i(\succeq_i', \succeq_{-i}) = \sup_{\succeq_{-i} \in \mathcal{P}_{-i}^t} f_i(\succeq_i', \succeq_{-i}) = f_i(\succeq_i', \succeq_{-i}^*).$$

Next, since f is strategy-proof, it holds

$$f_i(\succeq_i, \succeq_{-i}^*) \succeq_i f_i(\succeq_i', \succeq_{-i}^*).$$

Combing the three relations above, we conclude

$$\inf_{\succeq_{-i}\in\mathcal{P}^t_{-i}} f_i(\succeq_i,\succeq_{-i}) = f_i(\succeq_i,\succeq^*_{-i}) \succeq_i f_i(\succeq'_i,\succeq^*_{-i}) = \sup_{\succeq_{-i}\in\mathcal{P}^t_{-i}} f_i(\succeq'_i,\succeq_{-i}).$$

Note that since i, $\succeq_i$, $\succeq'_i$ and $\mathcal{P}^t_{-i}$ are arbitrarily taken, f is by definition Σ-OSP.

To show the "only if" part, I prove the contrapositive statement: If the partition system Σ is not self-invariant under f, then f is not Σ-OSP.

Suppose that Σ is not self-invariant under f. Then, there exist $\Sigma_i \in \Sigma$, $\mathcal{P}^*_{-i} \in \Sigma_i$, $\succeq^1_{-i}, \succeq^2_{-i} \in \mathcal{P}^*_{-i}$ and $\succeq_i \in \mathcal{P}_i$ such that $f_i(\succeq_i,\succeq^1_{-i}) \neq f_i(\succeq_i,\succeq^2_{-i})$. Since I focus on strict preferences, it follows that either $f_i(\succeq_i,\succeq^1_{-i}) \succ_i f_i(\succeq_i,\succeq^2_{-i})$ or $f_i(\succeq_i,\succeq^2_{-i}) \succ_i f_i(\succeq_i,\succeq^1_{-i})$. For the rest of the proof, I assume w.l.o.g. that $f_i(\succeq_i,\succeq^1_{-i}) \succ_i f_i(\succeq_i,\succeq^2_{-i})$, since we just need to relabel the two profiles otherwise and the following argument would still work. Select any $\succeq'_i \in \mathcal{P}_i$ such that (1) $f_i(\succeq_i,\succeq^1_{-i})$ ranks top on $\succeq'_i$ and (2) $\succeq'_i \neq \succeq_i$.

I first argue that $f_i(\succeq'_i,\succeq^1_{-i}) = f_i(\succeq_i,\succeq^1_{-i})$. If not, and note that as $f_i(\succeq_i,\succeq^1_{-i})$ ranks top on $\succeq'_i$, it follows that $f_i(\succeq_i,\succeq^1_{-i}) \succeq'_i f_i(\succeq'_i,\succeq^1_{-i})$. However, this contradicts to the fact that f is strategy-proof. Next, since $f(\succeq'_i,\succeq^1_{-i}) = f_i(\succeq_i,\succeq^1_{-i})$ and $\succeq^1_{-i} \in \mathcal{P}^*_{-i}$, we can infer that with respect to $\succeq_i$,

$$\sup_{\succeq_{-i}\in\mathcal{P}^*_{-i}} f_i(\succeq'_i,\succeq_{-i}) \succeq_i f_i(\succeq_i,\succeq^1_{-i}).$$

Also, since $\succeq^2_{-i} \in \mathcal{P}^*_{-i}$, it is clear that

$$f_i(\succeq_i,\succeq^2_{-i}) \succeq_i \inf_{\succeq_{-i}\in\mathcal{P}^*_{-i}} f_i(\succeq_i,\succeq_{-i}).$$

Combining the just described two relations with the assumption $f_i(\succeq_i,\succeq^1_{-i}) \succ_i f_i(\succeq_i,\succeq^2_{-i})$, we can

claim that there exist $\succeq_i, \succeq_i' \in \mathcal{P}_i$ and $\mathcal{P}_{-i}^* \in \Sigma_i$ such that with respect to $\succeq_i$,

$$\sup_{\succeq_{-i} \in \mathcal{P}_{-i}^*} f_i(\succeq_i', \succeq_{-i}) \succ_i \inf_{\succeq_{-i} \in \mathcal{P}_{-i}^*} f_i(\succeq_i, \succeq_{-i}).$$

This means that f is not Σ_i-OSP for agent i. Since $\Sigma_i \in \Sigma$, f is thus not Σ-OSP, and this completes the proof. □

Theorem 4.1 states that agents will avoid strategic mistakes under strategy-proof matching rules if and only if the uncertainty about their own assignments is avoided. On the one hand, the result is intuitive since I focus on ordinal preferences where agents only care about their own assignments. Therefore, agents with self-invariant partitions stick to truth-telling even if they cannot figure out others' assignments. On the other hand, Theorem 4.1 implies that any small uncertainty about their own results might cause agents to deviate. Note that so far I restrict attention to static games where agents report their full preferences at once. As I will show in Section 4.4, the requirements put on by Theorem 4.1 are loosened when extensive-form games are included.

Theorem 4.1 characterizes all partition systems which guarantee the POSP of a strategy-proof matching rule. In fact, it might be interesting to focus on the coarsest ones since they refer to the least necessary levels of reasoning abilities for agents to understand the strategy-proofness of the underlying rule. As I will construct the coarsest partition systems in the next section, I show here that for each strategy-proof rule, the coarsest self-invariant partition system is unique.

Proposition 4.1. *For any strategy-proof matching rule f, there exists a partition system Σ^f such that (1) f is Σ^f-OSP and that (2) for any $\Sigma \neq \Sigma^f$ satisfying f being Σ-OSP, Σ is a decomposition of Σ^f.*

Proof. Since possible partitions are finite in a given problem, for each agent $i \in I$, select a partition Σ_i^f such that f is Σ_i^f-OSP for i and that $|\Sigma_i| \geq |\Sigma_i^f|$ for any other partition Σ_i which satisfies that f

is Σ_i-OSP. Let $\Sigma^f = \{\Sigma_i^f\}_{i\in I}$, and it is obvious that f is Σ^f-OSP. In the following, we show that the selected partition system Σ^f also fulfills the second condition in Proposition 4.1.

As has been already shown by Lemma 4.2, given any decomposition Σ of Σ^f, it must be true that f is Σ-OSP. To complete the proof, it remains to show that for every Σ' that is not a decomposition of Σ^f, f is not Σ'-OSP. I proceed by contradiction. Suppose that there exists Σ' such that f is Σ'-OSP and that Σ' is not a decomposition of Σ^f. Then, there exists at least one agent $i \in I$, one event $\mathcal{P}'_{-i} \in \Sigma'_i$ and two type profiles $\succeq^1_{-i}, \succeq^2_{-i} \in \mathcal{P}'_{-i}$ such that $\succeq^1_{-i} \in \mathcal{P}^{f,1}_{-i}$ and $\succeq^2_{-i} \in \mathcal{P}^{f,2}_{-i}$ for two distinct $\mathcal{P}^{f,1}_{-i}, \mathcal{P}^{f,2}_{-i} \in \Sigma^f_i$. I aim at a contradiction to the selection of Σ^f.

From now on, fix any $\succeq_i \in \mathcal{P}_i$. Since f is Σ'-OSP, by Theorem 4.1 we know that Σ'_i is self-invariant under f. Therefore, we have

$$f_i(\succeq_i, \succeq^1_{-i}) = f_i(\succeq_i, \succeq^2_{-i}).$$

Similarly, since f is Σ^f-OSP, we know that Σ^f_i is self-invariant. Thus, for all $\succeq'_{-i} \in \mathcal{P}^{f,1}_{-i}$, we have

$$f_i(\succeq_i, \succeq'_{-i}) = f_i(\succeq_i, \succeq^1_{-i}),$$

and for all $\succeq''_{-i} \in \mathcal{P}^{f,2}_{-i}$, we have

$$f_i(\succeq_i, \succeq''_{-i}) = f_i(\succeq_i, \succeq^2_{-i}).$$

Let $\mathcal{P}^f_{-i} = \mathcal{P}^{f,1}_{-i} \cup \mathcal{P}^{f,2}_{-i}$. Note that $\succeq_i$ is arbitrarily taken, we conclude from the above findings that for all $\succeq'_i \in \mathcal{P}_i$ and all $\succeq_{-i}, \tilde{\succeq}_{-i} \in \mathcal{P}^f_{-i}$,

$$f_i(\succeq'_i, \succeq_{-i}) = f_i(\succeq'_i, \tilde{\succeq}_{-i}).$$

Then, we construct Σ^* such that $\Sigma^*_i = \Sigma^f_i \cup \mathcal{P}^f_{-i} \setminus \{\mathcal{P}^{f,1}_{-i}, \mathcal{P}^{f,2}_{-i}\}$ and $\Sigma^*_j = \Sigma^f_j$ for $j \neq i$. By construction, Σ^* is self-invariant under f. Again by Theorem 4.1, f is Σ^*-OSP. However, note that as $|\Sigma^*_i| = |\Sigma^f_i| - 1$,

this contradicts to how Σ^f is selected. □

The next and final result of this subsection is a characterization for self-invariant partitions. Fix a strategy-proof matching rule f and a type profile $\succeq_{-i} \in \mathcal{P}_{-i}$. Let

$$O_i(f, \succeq_{-i}) = \{o \in O \mid \exists \succeq_i \in \mathcal{P}_i : f_i(\succeq_i, \succeq_{-i}) = o\}$$

denote the set of objects that agent i could potentially receive under f when the type profile of others is known as $\succeq_{-i}$.

Proposition 4.2. *A partition Σ_i is self-invariant under a strategy-proof rule f if and only if for each $\mathcal{P}^t_{-i} \in \Sigma_i$, it is true that $O_i(f, \succeq_{-i}) = O_i(f, \succeq'_{-i})$ for all pairs $\succeq_{-i}, \succeq'_{-i} \in \mathcal{P}^t_{-i}$.*

Proof. The "only if" part follows directly from the definition: If Σ_i is self-invariant under f, then for any two type profiles $\succeq_{-i}, \succeq'_{-i}$ that are contained in the same event $\mathcal{P}^t_{-i} \in \Sigma_i$, we have that $f_i(\succeq_i, \succeq_{-i}) = f_i(\succeq_i, \succeq'_{-i})$ for all $\succeq_i \in \mathcal{P}_i$. It then follows that $O_i(f, \succeq_{-i}) = O_i(f, \succeq'_{-i})$.

I prove the "if" part by showing the contrapositive statement. Suppose that Σ_i is not self-invariant, that is, there exist $\succeq_i \in \mathcal{P}_i$, $\mathcal{P}^t_{-i} \in \Sigma_i$ and $\succeq_{-i}, \succeq'_{-i} \in \mathcal{P}^t_{-i}$ such that $f_i(\succeq_i, \succeq_{-i}) \neq f_i(\succeq_i, \succeq'_{-i})$. Denote the two assignments of i by $f_i(\succeq_i, \succeq_{-i}) = o$ and $f_i(\succeq_i, \succeq'_{-i}) = o'$, respectively. For the rest of the proof, I assume w.l.o.g. that $o \succ_i o'$, since we can just relabel the two profiles otherwise. First, note that since $f_i(\succeq_i, \succeq_{-i}) = o$, we have $o \in O_i(f, \succeq_{-i})$ by the definition of O_i. Second, since f is strategy-proof, it must hold that $o' = f_i(\succeq_i, \succeq'_{-i}) \succeq_i f_i(\succeq'_i, \succeq'_{-i})$ for all $\succeq'_i \in \mathcal{P}_i$. Therefore, for any $\tilde{o} \in O_i(f, \succeq'_{-i})$, we have $o' \succeq_i \tilde{o}$, which implies $o \notin O_i(f, \succeq'_{-i})$. Since $o \in O_i(f, \succeq_{-i})$ and $o \notin O_i(f, \succeq'_{-i})$, we obtain the desired result that $O_i(f, \succeq_{-i}) \neq O_i(f, \succeq'_{-i})$. This completes the proof. □

Proposition 4.2 shows that for each agent with a self-invariant partition, the set of objects that remains possible to her in each event is constant. As a remark, Proposition 4.2 provides an important

implication for the coarsest partition system Σ^f introduced in Proposition 4.1. That is, for all $\Sigma_i^f \in \Sigma^f$, all $\mathcal{P}_{-i}^t, \mathcal{P}_{-i}^s \in \Sigma_i^f$ (with $t \neq s$) and all $\succeq_{-i}^t \in \mathcal{P}_{-i}^t$, $\succeq_{-i}^s \in \mathcal{P}_{-i}^s$, it is true $O_i(f, \succeq_{-i}^t) \neq O_i(f, \succeq_{-i}^s)$. In fact, Proposition 4.2, along with this implication, turn out to be useful soon when I construct the unique coarsest self-invariant partition system under a strategy-proof matching rule.

4.3.2 Coarsest self-invariant partition system under TTC

In this subsection, I focus on TTC. More concretely, I introduce an algorithm which yields the coarsest self-invariant partition system under f^{TTC} for a problem $(I, O, \rhd_O, \mathcal{P})$. Building on the insights from Proposition 4.2, this algorithm computes the desired partition system by classifying the sets $\{O_i(f^{TTC}, \succeq_{-i}) \mid \succeq_{-i} \in \mathcal{P}_{-i}\}$ for each $i \in I$ via the following processes.

Round 0 Let $\mathcal{P}'_{-i} = \mathcal{P}_{-i}$ and let Σ_i^{TTC} be the finest partition of $\mathcal{P}_{-i}$ where each event is a singleton. Move to Round 1.

Round 1 If $\mathcal{P}'_{-i}$ is non-empty, select any $\succeq_{-i} \in \mathcal{P}'_{-i}$ and move to Round 2. Otherwise, terminate the algorithm.

Round 2 Run TTC without i's report. Concretely, go through the following processes.

Round 2.1 Each agent $i' \in I \setminus \{i\}$ points to her favorite object (or herself) according to $\succeq_{-i}$, and each object $o \in O$ points to the agent who has the highest priority on $\rhd_o$. If there are no cycles, move to Round 3. Otherwise, remove the cycles from the system, denote the remaining agents by I_1, denote the remaining objects by O_1 and move to Round 2.2.

Round 2.k, $k \geq 2$ Each agent $i' \in I_{k-1} \setminus \{i\}$ points to her favorite object in O_{k-1} (or herself) according to $\succeq_{-i}$, and each object $o \in O_{k-1}$ points to the agent who has the highest priority among agents in I_{k-1}. If there are no cycles, move to Round 3. Otherwise, remove the cycles from the system, denote the remaining agents by I_k, denote the remaining objects by O_k and move to Round 2.$(k+1)$.

Round 3 Let $O(\succeq_{-i})$ comprise the objects that remain in the system after the last step of Round 2. If there exists $\succeq'_{-i} \in \mathcal{P}_{-i} \setminus \mathcal{P}'_{-i}$ such that $O(\succeq_{-i}) = O(\succeq'_{-i})$, add $\succeq_{-i}$ to the event in Σ_i^{TTC} that contains $\succeq'_{-i}$. Otherwise, keep the singleton event in Σ_i^{TTC} that contains $\succeq_{-i}$. Then, remove $\succeq_{-i}$ from $\mathcal{P}'_{-i}$ and move back to Round 1.

The algorithm goes through the above three rounds for $|\mathcal{P}_{-i}|$ times and terminates with the partition Σ_i^{TTC}. After running the algorithm for all $i \in I$, we obtain a partition system $\Sigma^{TTC} = \{\Sigma_i^{TTC}\}_{i \in I}$. The next result shows that Σ^{TTC} is the desired partition system.

Lemma 4.3. *The partition Σ_i^{TTC} is the coarsest self-invariant partition for i under f^{TTC}.*

Proof. To see this, let us first look at Round 3. Note that in this round, the algorithm guarantees that for any event $\mathcal{P}^t_{-i} \in \Sigma^{TTC}$, all types $\succeq_{-i} \in \mathcal{P}^t_{-i}$ yield the same set of remaining objects at the end of Round 2. Let $O(\mathcal{P}^t_{-i}) = O(\succeq_{-i})$ for any $\succeq_{-i} \in \mathcal{P}^t_{-i}$. Moreover, Round 3 also guarantees that any two events $\mathcal{P}^t_{-i}, \mathcal{P}^s_{-i} \in \Sigma_i^{TTC}$ have different values of remaining objects: $O(\mathcal{P}^t_{-i}) \neq O(\mathcal{P}^s_{-i})$.

As explained in the discussion of Proposition 4.2, the desired result follows once we have that $O(\succeq_{-i}) = O_i(f^{TTC}, \succeq_{-i})$. Thus, I show next that $O(\succeq_{-i}) = O_i(f^{TTC}, \succeq_{-i})$. Under TTC, agents and objects are allocated when they are contained in a top trading cycle. During the process of TTC, once we find a top trading cycle that does not contain i while i remains unassigned, this cycle becomes part of the final matching no matter what i reports (Roth, 1982). In other words, for objects contained in such cycles, i has no chance to be assigned to them regardless of her submitted type. Notably, each cycle removed in Round 2 of the algorithm belongs to the cycles mentioned above. This implies that $o \in O \setminus O_i(f^{TTC}, \succeq_{-i})$ for each $o \in O \setminus O(\succeq_{-i})$. That is, it holds $O_i(f^{TTC}, \succeq_{-i}) \subseteq O(\succeq_{-i})$. Next, select any $o \in O(\succeq_{-i})$. Since o remains unassigned after Round 2, then it is not contained in any cycle. Instead, o must be part of a *top trading chain*[5] that ends up with an object pointing to i.

[5] A top trading chain is an ordered list of *distinct* objects and agents where each agent points to her top choice and where each object points to agent with the top priority.

In this case, if i points to o, a top trading cycle will immediately be formed and TTC will assign i to o. Thus, $o \in O_i(f^{TTC}, \succeq_{-i})$, which further implies $O(\succeq_{-i}) \subseteq O_i(f^{TTC}, \succeq_{-i})$. In conclusion, we have $O(\succeq_{-i}) = O_i(f^{TTC}, \succeq_{-i})$, and thus Σ^{TTC} is the coarsest self-invariant partition system under f^{TTC}. □

Loosely speaking, the main benefit of the above algorithm for TTC is that it can compute the set $O_i(f^{TTC}, \succeq_{-i})$ without considering the types of agent i. Looking carefully at the discussions above, we see that this benefit comes from the fact that TTC is non-bossy (Pápai, 2000).[6] In Appendix 4.C, I present an algorithm which computes the coarsest partition system for any strategy-proof matching rule. Being more general than the above algorithm for TTC, it runs in more rounds but still improves upon cycling through all preference profiles.

4.4 Extensive-form games

Like obvious strategy-proofness, POSP is a solution concept that also applies to extensive-form games. Therefore, I now extend the analysis to extensive-form settings. Since the formal definition of an extensive-form revelation game is familiar to most readers, I relegate it to Appendix 4.B. Also, since the necessary adaptations to the definition of POSP are standard, I relegate the adapted version and the definition of a POSP implementation to Appendix 4.B.

Notably, it turns out that with necessary adjustments in the respective proofs, Lemma 4.1, Lemma 4.2, Proposition 4.1 and Proposition 4.2 are still true after including extensive-form games. However, as the next result shows, Theorem 4.1 fails to hold in this more general setting. Concretely, the self-invariance of a partition system Σ becomes not necessary for Σ-OSP implementation of TTC.

Theorem 4.2. *There exists a Σ-OSP implementation of TTC where Σ is not self-invariant under TTC.*

[6] As defined by Satterthwaite and Sonnenschein (1981), non-bossiness of a rule requires that no agent can change others' assignments without changing her own assignment by misreporting her types.

The proof below is constructive. Troyan (2019) shows that TTC is OSP-implementable if and only if the underlying priority structure is *weakly acyclic*.[7] Since Lemma 4.2 holds in extensive-form game settings, we can infer that when the priority structure is weakly acyclic, TTC is Σ-OSP implementable for any partition system Σ. Therefore, in the following, I consider an example where the priority structure is not weakly acyclic, and construct a partition system Σ that is not self-invariant under TTC but guarantees the Σ-OSP implementation of TTC.

Proof. Consider a market with three agents $I = \{i, j, k\}$ and three objects $O = \{a, b, c\}$. For simplicity suppose that all objects are acceptable. Agent i has the full domain of preferences while the preferences of agent j and k are given in the following table:

$\succeq_j^1$	$\succeq_j^2$	$\succeq_k^1$	$\succeq_k^2$
a	b	a	c
b	a	b	a
c	c	c	b

The priority structure $\rhd_O$ is as follows

$\rhd_a$	$\rhd_b$	$\rhd_c$
i	j	k
j	k	i
k	i	j

Note that $\rhd_O$ is not weakly acyclic since there are three agents who rank top at all objects. Suppose that TTC is implemented in this market, and consider the following partition $\Sigma_i = \{\mathcal{P}_{-i}^t\}_{t=1}^3$ with

$$\mathcal{P}_{-i}^1 = \{(\succeq_j^1, \succeq_k^1), (\succeq_j^1, \succeq_k^2)\},\ \ \mathcal{P}_{-i}^2 = \{(\succeq_j^2, \succeq_k^1)\},\ \ \mathcal{P}_{-i}^3 = \{(\succeq_j^2, \succeq_k^2)\}.$$

[7] As defined in Troyan (2019), a *strong cycle* in a priority structure $\rhd_O$ is described by three agents $i, j, k \in I$ and three objects $a, b, c \in C$ such that $i \rhd_a j, k, j \rhd_b i, k$ and $k \rhd_c i, j$. If there are no strong cycles, the priority structure is said to be *weakly acyclic*.

It is easily checked that $O_i(f, \succeq_j^1, \succeq_k^1) = \{a, b, c\}$ and $O_i(f, \succeq_j^1, \succeq_k^2) = \{a, b\}$. Next, note that since $(\succeq_j^1, \succeq_k^1), (\succeq_j^1, \succeq_k^2) \in \mathcal{P}_{-i}^1$ and $O_i(f, \succeq_j^1, \succeq_k^1) \neq O_i(f, \succeq_j^1, \succeq_k^2)$, by Proposition 4.2, Σ_i is not self-invariant under TTC.

Now consider the following extensive-form mechanism Γ that implements TTC in this market:

Step 1 Ask agent i to choose her top choice from $\{a, b, c\}$. If i responds a, assign i to a, j to b and k to c, then stop the mechanism; if i answers b, move to Step 2; if i answers c, move to Step 3.

Step 2 Ask j if she prefers a to b. If yes, assign i to b, j to a and k to c, then end the mechanism; otherwise, assign j to b and move to Step 2.1.

Step 2.1 Ask agent i if she prefers a to c. If yes, assign i to a and k to c, then end the mechanism. Otherwise, ask agent k if she prefers a to c. If k answers yes, assign i to c, assign k to a and end the mechanism; if k answers no, assign i to a, assign k to c and end the mechanism.

Step 3 Ask k if she prefers a to c. If yes, assign i to c, j to b and k to a, then end the mechanism; otherwise, assign k to c and move to Step 3.1.

Step 3.1 Ask agent i if she prefers a to b. If yes, assign i to a and j to b, then end the mechanism. Otherwise, ask j if she prefers a to b. If j answers yes, assign i to b, assign j to a and end the mechanism; if j answers no, assign i to a, assign j to b and end the mechanism.

I now argue that Γ is Σ_i-OSP for agent i. Concretely, I show that in each event of Σ_i, at each Step where i plays, her worst possible outcome under truth-telling is weakly better than her best possible outcome under misreporting.

I first check for the event $\mathcal{P}_{-i}^1$, that is, consider the two profiles $(\succeq_j^1, \succeq_k^1)$ and $(\succeq_j^1, \succeq_k^2)$. In this event, if i reports a or b as her top choice, she is ensured to receive it. Therefore, truth-telling is an obviously dominant strategy for i in cases where she truly prefers a or b most. It remains to consider

the case where the true preferences of agent i are $\succeq_i^1$: $c \succeq_i^1 a \succeq_i^1 b$ or $\succeq_i^2$: $c \succeq_i^2 b \succeq_i^2 a$. In these two cases, we first check whether i has the incentive to misreport at Step 1. Note that truth-telling of $\succeq_i^1$ and $\succeq_i^2$ refer to the same action at Step 1. Therefore, if i reports truthfully at Step 1, then under $(\succeq_j^1, \succeq_k^1)$, she will receive her favorite object; and under $(\succeq_j^1, \succeq_k^2)$ she will receive her second favorite object at Step 3.1. However, if i misreports at Step 1, she will be assigned to a or b whichever she misreports as her favorite. That is, the best possible assignment under misreporting is her second choice. Therefore, no matter whether i is of type $\succeq_i^1$ or $\succeq_i^2$, reporting truthfully obviously dominates any misreport at Step 1. Next, I check whether agent i has the incentive to report $\succeq_i^2$ when she is of type $\succeq_i^1$ (or the other way around) at Step 3.1. Once Step 3.1 is reached, agent i knows that c is already allocated to k. Then, in event $\mathcal{P}_{-i}^1$, she receives a by reporting $\succeq_i^1$ while she receives b by reporting $\succeq_i^2$. Notice that $a \succeq_i^1 b$ and $b \succeq_i^2 a$, she thus has no incentive to misreport at Step 3.1. In conclusion, when agent i only considers event $\mathcal{P}_{-i}^1$, truth-telling is an obviously dominant strategy.

As for the remaining events $\mathcal{P}_{-i}^2$ and $\mathcal{P}_{-i}^3$, since both events are singleton, the argument follows directly from strategy-proofness of TTC. I thus omit the details here.

So far, I can claim that Γ is Σ_i-OSP for agent i. Next, since Γ implements TTC and TTC is strategy-proof, it is easy to construct two self-invariant partitions under TTC, Σ_j and Σ_k for agent j and agent k, respectively. Let $\Sigma = \{\Sigma_i, \Sigma_j, \Sigma_k\}$. Since Σ_i is not self-invariant under TTC, Σ is not self-invariant under TTC. However, by construction it follows that Γ is Σ-OSP. Thus, Σ is a partition system which is not self-invariant under TTC but guarantees Σ-OSP implementation of TTC. This completes the proof. □

Notably, the self-invariance condition on a partition Σ under f is still sufficient for Σ-OSP implementation of f. To provide an intuition, recall first Theorem 4.1: If the reasoning ability of an agent corresponds to a self-invariant partition, then even if she is asked to report her entire preference ranking without any information about others' types, she will stick to truth-telling. Next, note that in

extensive-form games, agents are usually asked to reveal their types step by step. Such a way of reporting brings agents two sources of benefits. First, it allows agents to break down their decisions into small parts, and each of them requires less reasoning ability. Second, at each step when they (partially) reveal preferences, they receive information about preferences of others through the game. In conclusion, with more information and less reasoning load, agents with self-invariant partitions under f still have the incentive to truthfully report their preferences in extensive-form games that implement f.

4.5 Conclusion

In a standard one-to-one object-allocation setting, I model agents' bounded contingent reasoning and define an incentive property that lies between OSP and strategy-proofness. I study the degree of reasoning which is necessary and sufficient for an agent to understand that a matching rule is strategy-proof. This chapter opens up several avenues for future study. For instance, it is still an open question how much reasoning abilities are needed to understand a sequential form of a strategy-proof matching rule. Also, it would be interesting to regard POSP as a criterion to study and compare the performances of different sequential implementations of a matching rule.

4.A Supplement for Example 4.2

I now construct Σ_i such that in each event and given any of i's type, the assignment of i is fixed under TTC. The construction of Σ_i and i's assignments under TTC for all preference profiles are listed in Table 4.1. Concretely, the types of j and k are divided into four different events, namely $\mathcal{P}^1_{-i}$ to $\mathcal{P}^4_{-i}$. Let $\Sigma_i = \{\mathcal{P}^t_{-i}\}^4_{t=1}$. It is apparent from the table that agent i receives the same outcome in each event given any her own type – Σ_i is self-invariant for i under f^{TTC}. Also, we can conclude from the table that f^{TTC} is Σ_i-OSP for agent i.

Events	$\succeq_j$	$\succeq_k$		$f^{TTC}(\succeq)$					
			$\succeq_i$	*abc*	*acb*	*bac*	*bca*	*cab*	*cba*
$\mathcal{P}^1_{-i}$	*abc*	*abc*		*a*	*a*	*b*	*b*	*c*	*c*
	abc	*acb*		*a*	*a*	*b*	*b*	*c*	*c*
	acb	*abc*		*a*	*a*	*b*	*b*	*c*	*c*
	acb	*acb*		*a*	*a*	*b*	*b*	*c*	*c*
	abc	*bac*		*a*	*a*	*b*	*b*	*c*	*c*
	abc	*bca*		*a*	*a*	*b*	*b*	*c*	*c*
	acb	*bac*		*a*	*a*	*b*	*b*	*c*	*c*
	acb	*bca*		*a*	*a*	*b*	*b*	*c*	*c*
	cab	*abc*		*a*	*a*	*b*	*b*	*c*	*c*
	cab	*acb*		*a*	*a*	*b*	*b*	*c*	*c*
	cba	*abc*		*a*	*a*	*b*	*b*	*c*	*c*
	cba	*acb*		*a*	*a*	*b*	*b*	*c*	*c*
$\mathcal{P}^2_{-i}$	*abc*	*cab*		*a*	*a*	*b*	*b*	*a*	*b*
	abc	*cba*		*a*	*a*	*b*	*b*	*a*	*b*
	acb	*cab*		*a*	*a*	*b*	*b*	*a*	*b*
	acb	*cba*		*a*	*a*	*b*	*b*	*a*	*b*
	cab	*cab*		*a*	*a*	*b*	*b*	*a*	*b*
	cab	*cba*		*a*	*a*	*b*	*b*	*a*	*b*
$\mathcal{P}^3_{-i}$	*bac*	*abc*		*a*	*a*	*a*	*c*	*c*	*c*
	bac	*acb*		*a*	*a*	*a*	*c*	*c*	*c*
	bca	*abc*		*a*	*a*	*a*	*c*	*c*	*c*
	bca	*acb*		*a*	*a*	*a*	*c*	*c*	*c*
	bac	*bac*		*a*	*a*	*a*	*c*	*c*	*c*
	bca	*bac*		*a*	*a*	*a*	*c*	*c*	*c*
$\mathcal{P}^4_{-i}$	*bac*	*bca*		*a*	*a*	*a*	*a*	*a*	*a*
	bca	*bca*		*a*	*a*	*a*	*a*	*a*	*a*
	bac	*cab*		*a*	*a*	*a*	*a*	*a*	*a*
	bac	*cab*		*a*	*a*	*a*	*a*	*a*	*a*
	bca	*cba*		*a*	*a*	*a*	*a*	*a*	*a*
	bca	*cba*		*a*	*a*	*a*	*a*	*a*	*a*
	cab	*bac*		*a*	*a*	*a*	*a*	*a*	*a*
	cab	*bca*		*a*	*a*	*a*	*a*	*a*	*a*
	cba	*bac*		*a*	*a*	*a*	*a*	*a*	*a*
	cba	*bca*		*a*	*a*	*a*	*a*	*a*	*a*
	cba	*cab*		*a*	*a*	*a*	*a*	*a*	*a*
	cba	*cba*		*a*	*a*	*a*	*a*	*a*	*a*

Table 4.1: All scenarios for agents i

4.B POSP IN EXTENSIVE-FORM SETTINGS

In this section, I introduce partition obvious strategy-proofness in extensive-form settings. I start by formally introducing the extensive-form revelation games with imperfect information.

Definition 4.4. An *extensive-form revelation game with imperfect information* Γ consists of:

1. A rooted game tree R where:
   (a) r is the root node of R
   (b) $N(R)$ is the set of non-terminal nodes of R, where $r \in N(R)$
   (c) $L(R)$ is the set of terminal nodes of R
   (d) $E(R) = \{E(n)\}_{n \in N(R)}$ is the set of edges of R, where $E(n)$ is the set of edges originating from node n; given any edge $e \in E(R)$, we denote the origin of e by $n(e)$
2. A map τ: $L(R) \to \mathcal{M}$ from the terminal nodes to matchings.
3. A map π: $N(R) \to I$ from non-terminal nodes to agents.
4. For each $n \in N(R)$, a map φ_n: $E(n) \to 2^{\mathcal{P}_{\pi(n)}} \setminus \{\emptyset\}$ from edges to sets of preference relations for agent $\pi(n)$ such that:
   (a) for any two distinct $e, e' \in E(n)$, $\varphi_n(e) \cap \varphi_n(e') = \emptyset$,
   (b) if e^* is the first edge along the path from n back to the root r such that $\pi(n(e^*)) = \pi(n)$, then $\cup_{e \in E(n)} \varphi(e) = \varphi(e^*)$; if no such edge exists, then $\cup_{e \in E(n)} \varphi(e) = \mathcal{P}_{\pi(n)}$.
5. The collection of information sets $\mathcal{K} = \{\mathcal{K}_i\}_{i \in I}$ where each $\mathcal{K}_i$ is a partition of $\{n \mid \pi(n) = i\}$. Specifically, n_1 and n_2 are in the same cell of $\mathcal{K}_i$ if and only if the following conditions hold. For any $e_1 \in E(n_1)$, there is an edge $e_2 \in E(n_2)$ such that $\varphi(e_1) = \varphi(e_2)$ and $|E(n_1)| = |E(n_2)|$. Each cell of $\mathcal{K}_i$ is called an information set for agent i.

For the following discussion, fix an extensive-form revelation game Γ. Intuitively, at each internal node of Γ, an agent is called to take action, and the actions are interpreted by the edges. Specifically, an edge e outgoing from a node n is one possible action that agent $\pi(n)$ could take at n, and $\varphi_n(e)$ specifies the types of agent $\pi(n)$ that are recommended to take this action. Upon reaching any information set $K \in \mathcal{K}_i \in \mathcal{K}$, let $\mathcal{P}^K \subseteq \mathcal{P}$ be the set of remaining preference profiles, i.e. for each $i \in I$, $\mathcal{P}_i^K = \varphi(e_i)$ where e_i is the most recent edge along the path back from the information set K to r such that $\pi(n(e_i)) = i$ (and $\mathcal{P}_i^K = \mathcal{P}_i$ if no such edge exists). Note that at any node $n \in K$, the set of remaining preference profiles $\mathcal{P}^n$ satisfies $\mathcal{P}^n = \mathcal{P}^K$.

I now define agents' strategies in the game. Fix an agent i, and a *strategy* for i in Γ is a function $s_i : \mathcal{K}_i \times \mathcal{P}_i \to \mathcal{P}_i$ such that for all $K \in \mathcal{K}_i$ and all $\succeq_i \in \mathcal{P}_i$, it holds $s_i(K, \succeq_i) \in \cup_{e \in E(n)} \varphi_n(e)$ with any $n \in K$. When $s_i(K, \succeq_i) = \succeq_i'$, it means that at information set K, i will choose the action corresponding to the edge $e \in E(n)$ with $\succeq_i' \in \varphi(e)$ if her true type is $\succeq_i$. We use $s = (s_i)_{i \in I}$ to denote the strategy profile of all agents. Specifically, a strategy profile s^* is said to be the *truthful behavior* if $s_i^*(K, \succeq_i) = \succeq_i$ for all $i \in I$, $K \in \mathcal{K}_i$ and $\succeq_i \in \mathcal{P}_i$. That is, agents always choose edges that contain their true preferences under truthful behavior. With the notation by hand, I now define what it means for an extensive-form game to implement a matching rule.

Definition 4.5. An extensive-form revelation game with imperfect information Γ *implements rule f under truthful behavior* if for any terminal node $l \in L(R)$ and any $\succeq \in \mathcal{P}^n$, it holds $\tau(l) = f(\succeq)$.

Next, I introduce a few more notation that are necessary for defining POSP. Fix a partition system Σ. For any $K \in \mathcal{K}_i \in \mathcal{K}$ and any $\Sigma_i \in \Sigma$, let

$$\Sigma_i^K = \{\mathcal{P}_{-i}^{t,K}\}_{t=1}^n \text{ where } \mathcal{P}_{-i}^{t,K} = \mathcal{P}_{-i}^t \cap \mathcal{P}_{-i}^K$$

be the *conditional partition* for i at K. Loosely speaking, the conditional partition Σ_i^K partitions the set $\mathcal{P}_{-i}^K$ based on how these profiles are separated in Σ_i. Moreover, two preferences $\succeq_i, \succeq_i' \in \mathcal{P}_i$

are said to *diverge at node n*, if there exist two distinct edges $e, \tilde{e} \in E(n)$ such that $\succeq_i \in \varphi_n(e)$ and $\succeq_i' \in \varphi_n(\tilde{e})$. Accordingly, $\succeq_i, \succeq_i'$ are said to *diverge at the information set K*, if there exist two distinct edges $e, \tilde{e} \in \{E(n)\}_{n \in K}$ such that $\succeq_i \in \varphi_n(e)$, $\succeq_i' \in \varphi_n(\tilde{e})$ and $\varphi_n(e) \neq \varphi_n(\tilde{e})$.

We are now ready to define POSP in the more general setting.

Definition 4.6. Let Γ implement f under truthful behavior.

1. The revelation game Γ is Σ_i*-obviously strategy-proof* (Σ_i-OSP) for agent i when for all $K \in \mathcal{K}_i$, all $\mathcal{P}_{-i}^{t,K} \in \Sigma_i^K$ and all $\succeq_i, \succeq_i' \in \mathcal{P}_i$ such that $\succeq_i, \succeq_i'$ diverge at K,

$$\inf_{\succeq_{-i} \in \mathcal{P}_{-i}^{t,K}} f_i(\succeq_i, \succeq_{-i}) \succeq_i \sup_{\succeq_{-i} \in \mathcal{P}_{-i}^{t,K}} f_i(\succeq_i', \succeq_{-i})$$

   where $\inf f_i$ ($\sup f_i$) represents the outcome which ranks lowest (highest) on $\succeq_i$.

2. The game Γ is Σ*-obviously strategy-proof* (Σ-OSP) if for each $\Sigma_i \in \Sigma$, Γ is Σ_i-OSP for agent i.

In words, Γ is Σ_i-OSP for i if at each information set, for each conditional partition, for every type instructed to follow a certain edge at that information set, when agent i only considers the types of others in one certain cell of the conditional partition, the worst possible outcome she can receive is weakly better than the best outcome from any other edge. Finally, the definition of Σ-OSP implementation follows immediately from the above concepts.

Definition 4.7. An extensive-form revelation game with imperfect information Γ is said to be a Σ-*OSP implementation* of a matching rule f if Γ implements f under truthful behavior and Γ is Σ-OSP.

4.C Coarsest self-invariant partition system

In this section, I introduce an algorithm that computes the coarsest self-invariant partition system under any strategy-proof matching rule. Fix a problem $(I, O, \succeq, \rhd_O)$ and a strategy-proof matching

rule f. Following Proposition 4.2, I reach the goal by classifying the sets $\{O_i(f, \succeq_{-i}) \mid \succeq_{-i} \in \mathcal{P}_{-i}\}$. Specifically, to provide a convenient way of figuring out each $O_i(f, \succeq_{-i})$, I introduce the following result which holds with respect to any f.

Lemma 4.4. *Fix any $i \in I$ and any $\succeq_{-i} \in \mathcal{P}_{-i}$. For any object $o' \in O \backslash O_i(f, \succeq_{-i})$, agent i cannot receive o' no matter what she reports. Moreover, for any $o \in O_i(f, \succeq_{-i})$, agent i is guaranteed to be matched with o once she ranks o as her top choice.*

Proof. The first part is obvious from the definition, and thus omitted. I show the second part by contradiction. Suppose that there exists $o \in O_i(f, \succeq_{-i})$ such that for $\succeq_i$ on which o ranks highest, $f_i(\succeq_i, \succeq_{-i}) \neq o$. Notice that as $o \in O_i(f, \succeq_{-i})$, there exists a preference $\succeq_i' \in \mathcal{P}_i$ such that $f_i(\succeq_i', \succeq_{-i}) = o$. However, since that o ranks highest on $\succeq_i$ and that $\succeq_i$ is a strict ranking, we must have $f_i(\succeq_i', \succeq_{-i}) \succeq_i f_i(\succeq_i, \succeq_{-i})$, which contradicts the fact that f is strategy-proof. □

Lemma 4.4 clarifies that under any strategy-proof matching rule, to find out whether an object is available to an agent when the others' types are given, it is enough to check if that agent receives that object when she only lists it as acceptable. In the following, I introduce the target algorithm based on the characteristics provided in Lemma 4.4. Initially, fix any $i \in I$, let $\mathcal{P}'_{-i} = \mathcal{P}_{-i}$ and let Σ_i^f be the finest partition of $\mathcal{P}_{-i}$.

Round 1 If $\mathcal{P}'_{-i}$ is empty, terminate the algorithm. Otherwise, choose any $\succeq_{-i} \in \mathcal{P}'_{-i}$, let $O^* = O$, let $O(\succeq_{-i}) = \emptyset$ and move to Round 2.

Round 2 If O^* is empty, move to Round 4. Otherwise, choose any $o \in O^*$, let $\succeq_i^o$ be the preference where agent i only ranks o as acceptable,[8] let $\succeq^o = (\succeq_i^o, \succeq_{-i})$ and move to Round 3.

Round 3 Calculate $f(\succeq^o)$, let $O(\succeq_{-i}) = O(\succeq_{-i}) \cup \{f_i(\succeq^o)\}$, remove o from O^* and move back to Round 2.

[8] That is, $\succeq_i^o$ exhibits the following ranking: $o \succeq_i^o i \succeq_i^o \dots$

Round 4 If there exists $\succeq'_{-i} \in \mathcal{P}_{-i} \setminus \mathcal{P}'_{-i}$ such that $O(\succeq_{-i}) = O(\succeq'_{-i})$, add $\succeq_{-i}$ to the event in Σ_i^f that contains $\succeq'_{-i}$. Otherwise, keep the singleton event in Σ_i^f that contains $\succeq_{-i}$. Then, remove $\succeq_{-i}$ from $\mathcal{P}'_{-i}$ and move back to Round 1.

Let $\Sigma^f = \{\Sigma_i^f\}_{i \in I}$ be the partition system found by running the above algorithm for each $i \in I$. Notably, by Proposition 4.2 and Lemma 4.4, Σ^f is the coarsest self-invariant partition system under f.

Bibliography

Atila Abdulkadiroğlu and Tayfun Sönmez. School choice: A mechanism design approach. *American Economic Review*, 93(3):729–747, 2003.

Atila Abdulkadiroğlu, Parag A Pathak, Alvin E Roth, and Tayfun Sönmez. The boston public school match. *American Economic Review*, 95(2):368–371, 2005.

Atila Abdulkadiroğlu, Parag A. Pathak, and Alvin E. Roth. Strategy-proofness versus efficiency in matching with indifferences: Redesigning the nyc high school match. *American Economic Review*, 99(5):1954–78, 2009.

Mohammad Akbarpour and Shengwu Li. Credible auctions: A trilemma. *Econometrica*, 88(2): 425–467, 2020.

Samson Alva and Vikram Manjunath. Stable-dominating rules. Working paper, University of Ottawa, 2019.

Itai Ashlagi and Yannai A Gonczarowski. Stable matching mechanisms are not obviously strategy-proof. *Journal of Economic Theory*, 177:405–425, 2018.

Itai Ashlagi, Mark Braverman, Yash Kanoria, and Peng Shi. Clearing matching markets efficiently: informative signals and match recommendations. *Management Science*, 66(5):2163–2193, 2020.

Eduardo M Azevedo and Jacob D Leshno. A supply and demand framework for two-sided matching markets. *Journal of Political Economy*, 124(5):1235–1268, 2016.

Sophie Bade and Yannai A Gonczarowski. Gibbard-satterthwaite success stories and obvious strategyproofness. *arXiv preprint arXiv:1610.04873*, 2016.

Michel Balinski and Tayfun Sönmez. A tale of two mechanisms: student placement. *Journal of Economic theory*, 84(1):73–94, 1999.

Keisuke Bando. On the existence of a strictly strong nash equilibrium under the student-optimal deferred acceptance algorithm. *Games and Economic Behavior*, 87:269 – 287, 2014.

Inacio Bo and Rustamdjan Hakimov. The iterative deferred acceptance mechanism. *Available at SSRN 2881880*, 2019.

Inácio Bó and Rustamdjan Hakimov. Iterative versus standard deferred acceptance: Experimental evidence. *The Economic Journal*, 130(626):356–392, 2020a.

Inácio Bó and Rustamdjan Hakimov. Pick-an-object mechanisms. *Available at SSRN*, 2020b.

Tilman Börgers and Jiangtao Li. Strategically simple mechanisms. *Econometrica*, 87(6):2003–2035, 2019.

Yves Breitmoser and Sebastian Schweighofer-Kodritsch. Obviousness around the clock. *Experimental Economics*, pages 1–31, 2021.

Peter Chen, Michael Egesdal, Marek Pycia, and M Bumin Yenmez. Quantile stable mechanisms, 2014.

Yan Chen and Onur Kesten. Chinese college admissions and school choice reforms: A theoretical analysis. *Journal of Political Economy*, 125(1):99–139, 2017.

Yan Chen and Tayfun Sönmez. Improving efficiency of on-campus housing: An experimental study. *American Economic Review*, 92(5):1669–1686, 2002.

Yan Chen and Tayfun Sönmez. School choice: an experimental study. *Journal of Economic theory*, 127(1):202–231, 2006.

Yiqiu Chen. Partition obviously strategy-proof rules. Working paper, University of Cologne, 2021.

Yiqiu Chen and Markus Möller. Regret-free truth-telling in school choice with consent. Working paper, University of Cologne, 2021.

Yiqiu Chen and Alexander Westkamp. Optimal sequential implementation. Working paper, University of Cologne, 2021.

Lester E Dubins and David A Freedman. Machiavelli and the gale-shapley algorithm. *The American Mathematical Monthly*, 88(7):485–494, 1981.

Umut Dur, Robert G Hammond, and Thayer Morrill. Identifying the harm of manipulable school-choice mechanisms. *American Economic Journal: Economic Policy*, 10(1):187–213, 2018.

Umut Dur, A Arda Gitmez, and Özgür Yılmaz. School choice under partial fairness. *Theoretical Economics*, 14(4):1309–1346, 2019.

Federico Echenique, Alistair J Wilson, and Leeat Yariv. Clearinghouses for two-sided matching: An experimental study. *Quantitative Economics*, 7(2):449–482, 2016.

Lars Ehlers. Truncation strategies in matching markets. *Mathematics of Operations Research*, 33(2): 327–335, 2008.

Lars Ehlers and Thayer Morrill. (Il)legal Assignments in School Choice. *The Review of Economic Studies*, 08 2019.

Richard Engelbrecht-Wiggans. The effect of regret on optimal bidding in auctions. *Management Science*, 35(6):685–692, 1989.

Aytek Erdil and Haluk Ergin. What's the matter with tie-breaking? improving efficiency in school choice. *American Economic Review*, 98(3):669–89, 2008.

MA Fernandez. Deferred acceptance and regret-free truthtelling. Working paper, John Hopkins University, 2020.

Emel Filiz-Ozbay and Erkut Y Ozbay. Auctions with anticipated regret: Theory and experiment. *American Economic Review*, 97(4):1407–1418, 2007.

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

Thomas Gilovich and Victoria Husted Medvec. The experience of regret: what, when, and why. *Psychological review*, 102(2):379, 1995.

Yannai A Gonczarowski, Noam Nisan, Rafail Ostrovsky, and Will Rosenbaum. A stable marriage requires communication. *Games and Economic Behavior*, 118:626–647, 2019.

B Gong and Y Liang. A dynamic college admission mechanism in inner mongolia: Theory and experiment. *Working Paper*, 2016.

Julien Grenet, Yinghua He, and Dorothea Kübler. Decentralizing centralized matching markets: Implications from early offers in university admissions. Technical report, WZB Discussion Paper, 2019.

Guillaume Haeringer and Vincent Iehlé. Gradual college admission. *Université Paris-Dauphine Research Paper*, (3488038), 2019.

Rustamdjan Hakimov and Madhav Raghavan. Transparency in centralised allocation. Technical report, 2020.

Avinatan Hassidim, Assaf Romm, and Ran I Shorrer. " strategic" behavior in a strategy-proof environment. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pages 763–764, 2016.

Avinatan Hassidim, Déborah Marciano, Assaf Romm, and Ran I Shorrer. The mechanism is truthful, why aren't you? *American Economic Review*, 107(5):220–24, 2017.

Nicole Immorlica, Jacob Leshno, Irene Lo, and Brendan Lucier. Information acquisition in matching markets: The role of price discovery. *Available at SSRN*, 2020.

Onur Kesten. School choice with consent. *The Quarterly Journal of Economics*, 125(3):1297–1348, 2010.

Bettina Klaus and Flip Klijn. Median stable matching for college admissions. *International Journal of Game Theory*, 34(1):1, 2006.

Flip Klijn, Joana Pais, and Marc Vorsatz. Static versus dynamic deferred acceptance in school choice: Theory and experiment. *Games and Economic Behavior*, 113:147–163, 2019.

Fuhito Kojima and Mihai Manea. Axioms for deferred acceptance. *Econometrica*, 78(2):633–653, 2010.

Hyukjun Kwon and Ran I. Shorrer. Justified-envy-minimal efficient mechanisms for priority-based matching. *Microeconomics: Welfare Economics & Collective Decision-Making eJournal*, 2019.

Shengwu Li. Obviously strategy-proof mechanisms. *American Economic Review*, 107(11):3257–87, 2017.

Andrew Mackenzie. A revelation principle for obviously strategy-proof implementation. *Games and Economic Behavior*, 124:512–533, 2020.

Noam Nisan and Ilya Segal. The communication requirements of efficient allocations and supporting prices. *Journal of Economic Theory*, 129(1):192–224, 2006.

Joana Pais and Ágnes Pintér. School choice and information: An experimental study on matching mechanisms. *Games and Economic Behavior*, 64(1):303–328, 2008.

Szilvia Pápai. Strategyproof assignment by hierarchical exchange. *Econometrica*, 68(6):1403–1433, 2000.

Parag A Pathak and Tayfun Sönmez. Leveling the playing field: Sincere and sophisticated players in the boston mechanism. *American Economic Review*, 98(4):1636–52, 2008.

Parag A Pathak, Tayfun Sönmez, M Utku Ünver, and M Bumin Yenmez. Fair allocation of vaccines, ventilators and antiviral treatments: leaving no ethical value behind in health care rationing. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 785–786, 2021.

Marek Pycia and Peter Troyan. A theory of simplicity in games and mechanism design. *University of Zurich, Department of Economics, Working Paper*, (393), 2021.

Alex Rees-Jones. Suboptimal behavior in strategy-proof mechanisms: Evidence from the residency match. *Games and Economic Behavior*, 108:317–330, 2018.

Alvin E Roth. Incentive compatibility in a market with indivisible goods. *Economics letters*, 9(2): 127–132, 1982.

Alvin E Roth and Elliott Peranson. The redesign of the matching market for american physicians: Some engineering aspects of economic design. *American economic review*, 89(4):748–780, 1999.

Alvin E Roth and Uriel G Rothblum. Truncation strategies in matching markets—in search of advice for participants. *Econometrica*, 67(1):21–43, 1999.

Alvin E Roth, Tayfun Sönmez, and M Utku Ünver. Kidney exchange. *The Quarterly Journal of Economics*, 119(2):457–488, 2004.

Mark A Satterthwaite and Hugo Sonnenschein. Strategy-proof allocation mechanisms at differentiable points. *The Review of Economic Studies*, 48(4):587–597, 1981.

Ran I Shorrer and Sándor Sóvágó. Obvious mistakes in a strategically simple college admissions environment: Causes and consequences. *Available at SSRN 2993538*, 2018.

Qianfeng Tang and Jingsheng Yu. A new perspective on kesten's school choice with consent idea. *Journal of Economic Theory*, 154:543–561, 2014.

Chung-Piaw Teo and Jay Sethuraman. The geometry of fractional stable matchings and its applications. *Mathematics of Operations Research*, 23(4):874–891, 1998.

Clayton Thomas. Classification of priorities such that deferred acceptance is obviously strategyproof. *arXiv preprint arXiv:2011.12367*, 2021.

Peter Troyan. Obviously strategy-proof implementation of top trading cycles. *International Economic Review*, 60(3):1249–1261, 2019.

Peter Troyan and Thayer Morrill. Obvious manipulations. *Journal of Economic Theory*, 185:104970, 2020.

Robert Wilson. Game-theoretic analyses of trading processes. In *Advances in Economic Theory, Fifth World Congress*, pages 33–70, 1987.

Marcel Zeelenberg and Rik Pieters. A theory of regret regulation 1.0. *Journal of Consumer psychology*, 17(1):3–18, 2007.

Luyao Zhang and Dan Levin. Partition obvious preference and mechanism design: Theory and experiment. *Available at SSRN 2927190*, 2017.