Analysis of One-to-One Matching Mechanisms via SAT Solving: Impossibilities for Universal Axioms

Ulle Endriss
Institute for Logic, Language and Computation
University of Amsterdam
The Netherlands

Abstract

We develop a powerful approach that makes modern SAT solving techniques available as a tool to support the axiomatic analysis of economic matching mechanisms. Our central result is a preservation theorem, establishing sufficient conditions under which the possibility of designing a matching mechanism meeting certain axiomatic requirements for a given number of agents carries over to all scenarios with strictly fewer agents. This allows us to obtain general results about matching by verifying claims for specific instances using a SAT solver. We use our approach to automatically derive elementary proofs for two new impossibility theorems: (i) a strong form of Roth's classical result regarding the impossibility of designing mechanisms that are both stable and strategyproof and (ii) a result establishing the impossibility of guaranteeing stability while also respecting a basic notion of cross-group fairness (so-called gender-indifference).

1 Introduction

The development of the *theory of matching*, going back to the seminal work of Gale and Shapley (1962), has been one of the grand success stories of Economics and Operations Research, leading to significant benefits to society at large. Examples include matching schemes developed for the allocation of students to schools, of resident doctors to hospitals, and of organ donors to patients (Economic Sciences Prize Committee 2012).

Motivated by the demands of diverse applications, many different variants of the basic model of matching have been considered. For instance, we may need to compute one-to-one matchings between agents on two sides of a market, or we may be able to match several agents to a single agent on the other side of the market. We may require agents to report complete preferences over potential partners or we may want to permit incomplete (truncated) preferences. Similarly, we may or may not want to allow agents to declare preferential indifferences. The list goes on. This makes it difficult to obtain a clear picture of the range of desiderata for which we can successfully design a matching mechanism. This difficulty motivates the idea of using tools from AI to automate some of the tasks of the economic theorist intent on analysing an intricate model of matching to better understand what opportunities there are (or are not) for designing good mechanisms. In this paper, we develop an approach to automatically search for impossibility theorems regarding the design of matching mechanisms that satisfy a number of desirable properties of interest (so-called *axioms*). We focus on the classical model of one-to-one matching, with two groups of n agents each, who we need to pair up with each other on the basis of their reported preferences.

Our approach makes use of SAT solving techniques, which have long been studied—and applied—in AI (Biere, Heule, and van Maaren 2009). SAT is the NP-complete problem of deciding whether a given propositional formula has a model. Despite this high complexity, in practice a modern SAT solver will often be able to compute an answer in a matter of seconds for formulas in CNF (conjunctive normal form) with millions of clauses. We use such formulas to encode the desirable properties of matching mechanisms (for a specific choice of n) we are interested in; a model then corresponds to a mechanism with those properties. If we obtain a negative answer and thus designing a mechanism of the desired kind is impossible for the specific value of n tested, then in some cases this observation can be generalised to an impossibility theorem for all larger values of n.

Related work. The original idea of using a SAT solver to automatically prove the "base case" on an impossibility theorem is due to Tang and Lin (2009) who used this approach to obtain an alternative proof of Arrow's Impossibility Theorem and a number of closely related results in social choice theory. Since then, this methodology has also been applied, amongst other things, to the analysis of preference extensions (Geist and Endriss 2011), resolute voting rules (Brandt, Geist, and Peters 2017), irresolute voting rules (Brandt and Geist 2016), and multiwinner voting rules (Peters 2018). Geist and Peters (2017) review the state of the art in this subfield of computational social choice (Brandt et al. 2016). While SAT solving can automate the proof of a "base case", deriving a fully fledged impossibility theorem still requires a manual proof of the corresponding "inductive step". Depending on the axioms involved, this can be difficult and/or tedious. Only for the (particularly simple) domain of preference extensions has it been possible so far to obtain a general meta-result that shows—for axioms that meet certain syntactic conditions when expressed in a suitable formal language—that any impossibility observed at a given domain size always extends to all larger domains sizes, thereby making manual proofs for specific such axioms obsolete (Geist and Endriss 2011).

While the present paper is the first instance where the SAT-based approach is used to support the axiomatic study of matching mechanisms, we note that Drummond, Perrault, and Bacchus (2015) have used SAT solving techniques to *implement* matching mechanisms in practice.

Contribution. We propose a formal language for expressing axioms of interest in the context of designing one-to-one matching mechanisms and show that for axioms that have a particular syntactic form (the *universal axioms*) the mathematical possibility of designing mechanisms satisfying those axioms is preserved whenever we reduce the number of agents involved. This *Preservation Theorem* allows us to fully automate the search for relevant impossibility theorems by delegating the verification of the base case to a SAT solver. We describe this approach and use it to derive elementary human-readable proofs of two new impossibility theorems. One of them is a strong form of a classical result on the impossibility of combining *stability* and *strategyproofness* (Roth 1982), and the other establishes the impossibility of combining *stability* with a fairness notion known as *gender-indifference* (Masarani and Gokturk 1989).

Paper outline. In Section 2 we recall the familiar model of one-to-one matching and discuss relevant axioms. Then, in Section 3 we present our language for encoding axioms and prove our Preservation Theorem, and in Section 4 we show how to apply our approach to automate the search for impossibility theorems. Section 5 concludes.

The code used to derive (and full documentation of) the computer-generated results mentioned in this paper are available as supplementary material (Endriss 2019).

2 Matching Mechanisms

In this section we recall the classical model of one-to-one matching markets introduced by Gale and Shapley (1962). We also review a number axioms describing intuitively desirable properties of one-to-one matching mechanisms.

2.1 The Model

For any given number $n \in \mathbb{N}$, let $A_n = \{a_1, \ldots, a_n\}$ and $B_n = \{b_1, \ldots, b_n\}$. For any set X, let $X!$ denote the set of all strict linear orders on X. We want to model scenarios involving two groups of n agents each, A_n and B_n. Each agent $a \in A_n$ reports her preferences in the form of a strict linear order on B_n, and each agent $b \in B_n$ reports her preferences in the form of a strict linear order on A_n. This results in a *profile* of preferences—an element of $B_n!^n \times A_n!^n$.

On the basis of such a profile, we want to match each member of A_n with a member of B_n. The first—intriguing if moderately inappropriate—example for such a scenario in the original paper of Gale and Shapley (1962) is that of a “marriage market”, in which n men and n women need to get married off to each other. A more significant application is that of n job seekers and n employers (with one vacancy each) who need to get matched to each other.

A *matching* M is a subset of the set $A_n \times B_n$ of pairs with the property that $\#\{b \mid (a,b) \in M\} = 1$ for all $a \in A_n$ and $\#\{a \mid (a,b) \in M\} = 1$ for all $b \in B_n$.

A *matching mechanism* $\mu : B_n!^n \times A_n!^n \to 2^{A_n \times B_n}$ is a function mapping any given profile p to some matching $M = \mu(p)$. We call n the *dimension* of the mechanism μ (and also of the profiles p it can be applied to).

Example 1. Consider the following profile for $n = 3$:

$$\begin{array}{lcl@{\qquad}lcl} a_1 & : & b_1 \succ b_2 \succ b_3 & b_1 & : & a_2 \succ a_3 \succ a_1 \\ a_2 & : & b_2 \succ b_1 \succ b_3 & b_2 & : & a_3 \succ a_1 \succ a_2 \\ a_3 & : & b_1 \succ b_3 \succ b_2 & b_3 & : & a_1 \succ a_2 \succ a_3 \end{array}$$

Under the *serial dictatorship* with a_1 picking first and a_2 next, we obtain the matching $\{(a_1, b_1), (a_2, b_2), (a_3, b_3)\}$. Under the *deferred-acceptance mechanism* of Gale and Shapley (with agents in A_3 proposing to agents in B_3, who always accept the best proposal received so far), on the other hand, we obtain $\{(a_1, b_2), (a_2, b_1), (a_3, b_3)\}$. △

2.2 Axioms

What makes for a good matching mechanism? In game theory, and in economic theory more broadly, we approach such questions by means of the *axiomatic method*. This involves formulating intuitively appealing normative principles in precise mathematical terms, resulting in so-called *axioms*, and exploring what kind of mechanism will satisfy which axioms (Roth and Sotomayor 1990). The most important axiom in the literature on matching is *stability*. It encodes the desideratum that we would not want our mechanism to return a matching that is unstable, in the sense that two agents a and b—a so-called *blocking pair*—have an incentive to partner up with each other rather than the agents they have been assigned to by the mechanism.

Definition 1. *A mechanism μ is* ***stable*** *if for no profile p in which agent a prefers b to b' and agent b prefers a to a' it is the case that both $(a, b') \in \mu(p)$ and $(a', b) \in \mu(p)$.*

For instance, the aforementioned deferred-acceptance mechanism is known to be stable (Gale and Shapley 1962), while Example 1 shows that serial dictatorships are not stable (in the example, (a_3, b_1) is a blocking pair).

We are also going to consider the following (nonstandard) variant of stability, which only requires protection against deviating agents who are each others’ absolute favourite.

Definition 2. *A mechanism μ is* ***top-stable*** *if $(a, b) \in \mu(p)$ for every profile p in which agent a ranks b at the top of her preference order and agent b ranks a at the top of hers.*

Observe that stability entails top-stability (which is a much weaker requirement). If we have reason to believe that agents will reject the solution proposed by our mechanism only in the most extreme cases, then we may be content with trying to guarantee merely top-stability rather than the more demanding property of stability.

A central concern in all of game theory, including the theory of matching, is whether we can prevent strategic agents from misrepresenting their truthful preferences. For agent $i \in A_n \cup B_n$, we say that profiles p and p' are *i-variants* of each other if no agent other than i (and possibly not even i) changes her preferences when we move from p to p'.

Definition 3. *A mechanism μ is **one-way strategyproof** for agents of the first group if for no agent a in that group and no a-variants p and p' it is the case that, in p, a prefers b to b' for agents b and b' with $(a, b') \in \mu(p)$ and $(a, b) \in \mu(p')$.*

Note that if the condition above is violated, then in profile p agent a has an incentive to pretend that her actual preferences are as in profile p'. The axiom of one-way strategyproofness for agents of the second group is defined analogously. Now a mechanism μ is called *two-way strategyproof* if it is one-way strategyproof for agents of both groups.

A very different kind of concern is whether the mechanism we choose to implement is *fair*. For example, a well-known deficiency of the deferred-acceptance mechanism is that it treats the members of the first group more favourably than the members of the second group. An interesting proposal for an axiom designed to rule out such sources of unfairness is due to Masarani and Gokturk (1989). In reference to the interpretation of agents in A_n as men and agents in B_n as women, this axiom is known as “gender-indifference”. We say that profiles p and p' are *swap-variants* of each other, if a_i prefers b_j to $b_{j'}$ in p whenever b_i prefers a_j to $a_{j'}$ in p' (and analogously with the roles of a and b inverted).

Definition 4. *A mechanism μ is **gender-indifferent** if $(a_i, b_j) \in \mu(p)$ implies $(a_j, b_i) \in \mu(p')$ for any swap-variants p and p' and any two indices $i, j \in \mathbb{N}$.*

Thus, this axiom requires that, if we swap groups in the input consumed by μ, then this should result in a corresponding swap of the groups in the output returned by μ.

3 The Preservation Theorem

In this section we prove our Preservation Theorem, establishing that any matching mechanism μ^+ of dimension n with $n > 1$ that satisfies a given set of axioms meeting certain conditions can be transformed into a matching mechanism μ of dimension $n-1$ that satisfies the same set of axioms. The theorem applies to all axioms that can be expressed in (a particular fragment of) a particular formal language, which we also introduce in this section. The theorem is useful, because its contrapositive reading shows that any impossibility for designing a matching mechanism satisfying certain axioms that we might observe for a given specific dimension n immediately extends to a general impossibility theorem for every dimension greater or equal to n.

3.1 A Formal Language for Expressing Axioms

Recall the axioms capturing basic normative principles regarding the design of matching mechanisms reviewed in Section 2.2. They all talk about relationships between profiles and agents. To represent these and other axioms, we now introduce a formal language. Readers familiar with *many-sorted first-order logic* (see, e.g., Hodges 1997) will recognise this language as first-order logic with two sorts, one for profiles and one for agents (or more accurately: indices of agents), and a set of seven atomic propositions (to speak about preferences of agents in a given profile, about relationships between profiles, and about matchings). We stress that, in principle, there is room to extend this language further and to add additional atomic propositions for other types of basic relationships.

Let $\mathrm{Var}_{\mathrm{N}}$ and $\mathrm{Var}_{\mathrm{P}}$ be two disjoint (and sufficiently large) sets of *variables*. The set of all syntactically valid expressions φ of our language is defined by the following specification in Backus-Naur form. Here i represents variables in $\mathrm{Var}_{\mathrm{N}}$ and p represents variables in $\mathrm{Var}_{\mathrm{P}}$.

$$\begin{array}{rcl} \varphi & ::= & i \succ^{\mathrm{A}}_{p,i} i \mid i \succ^{\mathrm{B}}_{p,i} i \mid \mathit{top}^{\mathrm{A}}_{p,i} = i \mid \mathit{top}^{\mathrm{B}}_{p,i} = i \mid \\ & & p \sim^{\mathrm{A}}_{i} p \mid p \sim^{\mathrm{B}}_{i} p \mid p \rightleftarrows p \mid p \triangleright (i,i) \mid \\ & & \neg\varphi \mid \varphi \wedge \varphi \mid \forall_{\mathrm{N}} i.\varphi \mid \forall_{\mathrm{P}} p.\varphi \end{array}$$

The *interpretation* of formulas in our language depends on n, the dimension of the matching mechanisms we want to describe. Variables in $\mathrm{Var}_{\mathrm{N}}$ range over the set $\{1, \ldots, n\}$ of agent indices, and variables in $\mathrm{Var}_{\mathrm{P}}$ range over profiles. An *assignment* is a function $\alpha : \mathrm{Var}_{\mathrm{N}} \cup \mathrm{Var}_{\mathrm{P}} \to \mathcal{D}_n$, where $\mathcal{D}_n = \{1, \ldots, n\} \cup (B_n!^n \times A_n!^n)$ that ensures that $\alpha(i) \in \{1, \ldots, n\}$ for all $i \in \mathrm{Var}_{\mathrm{N}}$ and $\alpha(p) \in B_n!^n \times A_n!^n$ for all $p \in \mathrm{Var}_{\mathrm{P}}$. We often write p^α as a shorthand for $\alpha(p)$. For any $x \in \mathrm{Var}_{\mathrm{N}} \cup \mathrm{Var}_{\mathrm{P}}$, assignments α and α' are called *x-variants* of each other if $\alpha(y) = \alpha'(y)$ for every variable $y \in \mathrm{Var}_{\mathrm{N}} \cup \mathrm{Var}_{\mathrm{P}} \setminus \{x\}$. We write $\mu, \alpha \models \varphi$ to say that mechanism μ *satisfies* formula φ under assignment α. This notion of satisfaction is defined inductively as follows:

- $\mu, \alpha \models j \succ^{\mathrm{A}}_{p,i} j'$ if $a_{\alpha(i)}$ ranks $b_{\alpha(j)}$ above $b_{\alpha(j')}$ in p^α
- $\mu, \alpha \models i \succ^{\mathrm{B}}_{p,j} i'$ if $b_{\alpha(j)}$ ranks $a_{\alpha(i)}$ above $a_{\alpha(i')}$ in p^α
- $\mu, \alpha \models \mathit{top}^{\mathrm{A}}_{p,i} = j$ if $a_{\alpha(i)}$ ranks $b_{\alpha(j)}$ at the top in p^α
- $\mu, \alpha \models \mathit{top}^{\mathrm{B}}_{p,j} = i$ if $b_{\alpha(j)}$ ranks $a_{\alpha(i)}$ at the top in p^α
- $\mu, \alpha \models p \sim^{\mathrm{A}}_{i} p'$ if profiles p^α and p'^α are $a_{\alpha(i)}$-variants
- $\mu, \alpha \models p \sim^{\mathrm{B}}_{j} p'$ if profiles p^α and p'^α are $b_{\alpha(j)}$-variants
- $\mu, \alpha \models p \rightleftarrows p'$ if profiles p^α and p'^α are swap-variants
- $\mu, \alpha \models p \triangleright (i, j)$ if $(a_{\alpha(i)}, b_{\alpha(j)}) \in \mu(p^\alpha)$
- $\mu, \alpha \models \neg\varphi$ if $\mu, \alpha \models \varphi$ is not the case
- $\mu, \alpha \models \varphi \wedge \psi$ if both $\mu, \alpha \models \varphi$ and $\mu, \alpha \models \psi$ are the case
- $\mu, \alpha \models \forall_{\mathrm{N}} i.\varphi$ if $\mu, \alpha' \models \varphi$ for all i-variants α' of α
- $\mu, \alpha \models \forall_{\mathrm{P}} p.\varphi$ if $\mu, \alpha' \models \varphi$ for all p-variants α' of α

A formula φ is called a *sentence* if every variable occurring in φ is bound by a quantifier. Observe that for sentences φ the assignment α in a statement such as $\mu, \alpha \models \varphi$ plays no role, so we can simply write $\mu \models \varphi$ instead and say that mechanism μ satisfies sentence φ.

For notational convenience, we also introduce several additional operators that can be reduced to the core operators of our language in the usual manner: $\varphi \vee \psi$ is short for $\neg(\neg\varphi \wedge \neg\psi)$, $\varphi \to \psi$ is short for $\neg(\varphi \wedge \neg\psi)$, $\exists_{\mathrm{N}} i.\varphi$ is short for $\neg(\forall_{\mathrm{N}} i.\neg\varphi)$, and $\exists_{\mathrm{P}} p.\varphi$ is short for $\neg(\forall_{\mathrm{P}} p.\neg\varphi)$. Note that our language has neither constant nor function symbols.

Table 1 demonstrates how to encode the axioms defined in Section 2.2 in our language. Inspection of these encodings shows that, not only can all of these axioms be expressed in our language, but they also are naturally expressed using a specific syntactic form. Let us call an axiom *universal*, if

Stability	$\forall_{\mathrm{P}} p . \forall_{\mathrm{P}} p' . \forall_{\mathrm{N}} i . \forall_{\mathrm{N}} i' . \forall_{\mathrm{N}} j . \forall_{\mathrm{N}} j' \,.\, \big[\, (j \succ^{\mathrm{A}}_{p,i} j' \,\wedge\, i \succ^{\mathrm{B}}_{p,j} i') \;\rightarrow\; \neg(p \rhd (i,j') \,\wedge\, p \rhd (i',j)) \,\big]$
Top-stability	$\forall_{\mathrm{P}} p . \forall_{\mathrm{N}} i . \forall_{\mathrm{N}} j \,.\, \big[\, (top^{\mathrm{A}}_{p,i} = j \,\wedge\, top^{\mathrm{B}}_{p,j} = i) \;\rightarrow\; (p \rhd (i,j)) \,\big]$
One-way strategyproofness (for A)	$\forall_{\mathrm{P}} p . \forall_{\mathrm{P}} p' . \forall_{\mathrm{N}} i . \forall_{\mathrm{N}} j . \forall_{\mathrm{N}} j' \,.\, \big[\, (p \sim^{\mathrm{A}}_{i} p' \,\wedge\, j \succ^{\mathrm{A}}_{p,i} j') \;\rightarrow\; \neg(p \rhd (i,j') \,\wedge\, p' \rhd (i,j)) \,\big]$
Gender-indifference	$\forall_{\mathrm{P}} p . \forall_{\mathrm{P}} p' . \forall_{\mathrm{N}} i . \forall_{\mathrm{N}} j \,.\, [\, (p \rightleftarrows p') \;\rightarrow\; (p \rhd (i,j) \;\rightarrow\; p' \rhd (j,i)) \,]$

Table 1: Formalisation of common axioms

it can be expressed by a sentence in our language that is of the form $\forall \vec{x} . \varphi(\vec{x})$, where $\vec{x}$ is a sequence of variables (each bound by the appropriate quantifier) and $\varphi(\vec{x})$ is a quantifier-free formula involving only variables occurring in $\vec{x}$.

Lemma 1. *The set of universal axioms is closed under the operation of conjunction.*

Proof. Immediate from the fact that $(\forall \vec{x} . \varphi) \wedge (\forall \vec{y} . \psi)$ is equivalent to $\forall \vec{x} \vec{y} . (\varphi \wedge \psi)$ whenever the sequences of variables $\vec{x}$ and $\vec{y}$ do not overlap, together with the fact that we can freely rename variables for any given sentence. □

Proposition 2. *The axioms of stability, top-stability, (one-way and two-way) strategyproofness, and gender-indifference are all universal axioms.*

Proof. Encodings of stability, top-stability, one-way strategyproofness, and gender-indifference in the required universal form are given in Table 1. For two-way strategyproofness, which is the conjunction of two formulas of the kind used to encode one-way strategyproofness, the claim now follows from Lemma 1. □

While our formal language—and specifically its universal fragment—thus are sufficiently expressive to encode a range of important axioms of practical interest, there also are clear limitations (and, as we are going to see, these limitations are intended, as they allow us to focus on axioms that are "well-behaved" in a sense to be made precise).

3.2 Preservation of Universal Axioms

We are now ready to formulate our central result. It is similar in nature to (one direction of) the classical Łoś-Tarski Theorem, a basic staple of model theory (Hodges 1997), on the preservation of first-order $\forall_1$-formulas in substructures.

Theorem 3 (Preservation Theorem)**.** *Let μ^+ be a top-stable matching mechanism of dimension $n > 1$ that satisfies all axioms in a given set Φ of universal axioms. Then there also exists a top-stable matching mechanism μ of dimension $n-1$ that satisfies all axioms in Φ.*

Proof. Keeping in mind that (i) a given matching mechanism satisfies all the axioms in a set Φ if and only if it satisfies their conjunction, and given that (ii) we have seen that the family of universal axioms is closed under taking conjunctions, it suffices to prove the claim for a single universal axiom φ. So let μ^+ be an arbitrary top-stable mechanism of dimension $n > 1$ and let φ be a universal axiom such that $\mu^+ \models \varphi$. We need to construct a top-stable mechanism μ of dimension $n-1$ such that $\mu \models \varphi$.

We are going to define μ by fixing a procedure for extending any given profile p of dimension $n-1$ to a profile p^+ of dimension n in such a way that we can first apply μ^+ to p^+ and then project the result from $A_n \times B_n$ down to $A_{n-1} \times B_{n-1}$, before returning it as $\mu(p)$. So let p be any profile of dimension $n-1$. We construct p^+ as follows. We first add b_n to the bottom of each preference order of each of the agents in A_{n-1} and a_n to the bottom of the preference orders of each of the agents in B_{n-1}. We then fix the preference order of the additional agent a_n as $b_n \succ b_{n-1} \succ \cdots \succ b_1$ and that of the additional agent b_n as $a_n \succ a_{n-1} \succ \cdots \succ a_1$. Here is a schematic illustration of this construction of the new profile p^+:

$$\begin{array}{llll}
a_1 & : \; \square \succ \cdots \succ \square \succ b_n & b_1 & : \; \square \succ \cdots \succ \square \succ a_n \\
a_2 & : \; \square \succ \cdots \succ \square \succ b_n & b_2 & : \; \square \succ \cdots \succ \square \succ a_n \\
\vdots & \;\;\, \vdots \qquad\qquad\;\; \vdots \quad \vdots & \vdots & \;\;\, \vdots \qquad\qquad\;\; \vdots \quad \vdots \\
a_{n-1} & : \; \square \succ \cdots \succ \square \succ b_n & b_{n-1} & : \; \square \succ \cdots \succ \square \succ a_n \\
a_n & : \; b_n \succ \cdots \succ b_2 \succ b_1 & b_n & : \; a_n \succ \cdots \succ a_2 \succ a_1
\end{array}$$

As μ^+ is top-stable by assumption, we are guaranteed that $(a_n, b_n) \in \mu^+(p^+)$. So we can define the new mechanism μ by letting $\mu(p) = \mu^+(p^+) \setminus \{(a_n, b_n)\}$ for any given profile p of dimension $n-1$. Let us now show that $\mu \models \varphi$ indeed holds for the mechanism μ thus defined.

As φ is universal, it can be written as $\forall \vec{x} . \psi$, where $\forall \vec{x}$ is a sequence of universal quantifications and ψ is a quantifier-free formula (involving only variables that occur in the sequence $\vec{x}$). By definition, $\mu^+ \models \varphi$ if and only if $\mu^+, \alpha^+ \models \psi$ for *all* assignments α^+ from variables to elements of the domain $\mathcal{D}_n$. In particular, $\mu^+ \models \varphi$ entails $\mu^+, \alpha^+ \models \psi$ for all those assignments α^+ that happen to map every variable $i \in \mathrm{Var}_{\mathrm{N}}$ to an element of $\{1, \ldots, n-1\}$ and every variable $p \in \mathrm{Var}_{\mathrm{P}}$ to a profile (of dimension n) that could have been produced by our construction above. Let us call any such assignment (mapping into $\mathcal{D}_n$) a *restricted assignment*.

Now observe that there exists a natural bijection between the set of all restricted assignments α^+ mapping into $\mathcal{D}_n$ and the set of *all* assignments α mapping into $\mathcal{D}_{n-1}$:

- $\alpha^+(i) = \alpha(i)$ for all $i \in \mathrm{Var}_{\mathrm{N}}$
- $\alpha^+(p) = (\alpha(p))^+$ for all $p \in \mathrm{Var}_{\mathrm{P}}$

The latter means that we can obtain the interpretation (which is a profile) of the variable p under assignment α^+ by first retrieving the interpretation of p under assignment α and then performing our construction on that profile of dimension $n-1$ to turn it into a profile of dimension n.

As $\mu \models \varphi$ if and only if $\mu, \alpha \models \psi$ for all assignments α mapping into $\mathcal{D}_{n-1}$, we are done proving $\mu \models \varphi$ if we can show that $\mu^+, \alpha^+ \models \psi$ if and only if $\mu, \alpha \models \psi$ for every restricted assignment α^+ and its counterpart α.[1] We do so by induction on the size of ψ. There is one base case to consider for every kind of atomic proposition in the language:

- For $\psi = (j \succ^{\text{A}}_{p,i} j')$ the claim holds, because no agent a_i indexed by some $i \in \{1, \ldots, n-1\}$ changes her relative ranking of agents b_j and $b_{j'}$ indexed by some $j, j' \in \{1, \ldots, n-1\}$ when we move between a profile of dimension $n-1$ and the corresponding profile of dimension n. Analogous arguments apply to $(j \succ^{\text{B}}_{p,i} j')$, $(top^{\text{A}}_{p,i} = j)$, $(top^{\text{B}}_{p,j} = i)$, $(p \sim^{\text{A}}_{i} p')$, $(p \sim^{\text{B}}_{j} p')$, and $(p \rightleftarrows p')$.
- For $\psi = (p \rhd (i,j))$ the claim holds, because μ has been defined so as to agree with μ^+ on all pairings involving agents with indices in $\{1, \ldots, n-1\}$.

To complete the proof, we require one inductive step for every propositional operator in the language (which are negation and conjunction). This part is straightforward: First, ($\mu^+, \alpha^+ \models \psi$ if and only if $\mu, \alpha \models \psi$) immediately entails ($\mu^+, \alpha^+ \models \neg\psi$ if and only if $\mu, \alpha \models \neg\psi$). Second, once we have established ($\mu^+, \alpha^+ \models \psi$ if and only if $\mu, \alpha \models \psi$) and ($\mu^+, \alpha^+ \models \psi'$ if and only if $\mu, \alpha \models \psi'$), this immediately entails ($\mu^+, \alpha^+ \models \psi \wedge \psi'$ if and only if $\mu, \alpha \models \psi \wedge \psi'$).

It remains for us to verify that μ is top-stable. But given that top-stability is a universal axiom, this follows immediately from our general result about μ satisfying all universal axioms that are satisfied by μ^+. □

Some readers may find this result unsurprising. It certainly could be argued that the claim made is obvious as far as *specific* universal axioms are concerned, such as stability. Indeed, the kind of reasoning we have used here is implicit in some results in the literature, where it is taken to be self-evident that, whenever we can design a mechanism for n that satisfies the axiom of interest, then we certainly can do so for $n-1$ as well.[2] The power of the Preservation Theorem lies in the fact that it applies to *every* combination of axioms we can express in the universal fragment of our language. At this point, it may be instructive to consider examples for (other) axioms for which our theorem does *not* apply.

Example 2. An axiom of interest we are *not* able to express in our language is known as *peer-indifference* (Masarani and Gokturk 1989). It postulates that the outcome returned by a mechanism should not change if we swap two agents belonging to the same group. Modelling this axiom would require the introduction of an additional kind of atomic proposition that allows us to state that profile p' can be obtained from profile p by swapping two of the agents in the first group (and a similar atomic proposition for swapping two agents in the second group). And indeed, peer-indifference is not always preserved when moving to smaller scenarios. For instance, even though for $n = 3$ there exists a mechanism that is both peer-indifferent and gender-indifferent, for $n = 2$ this is not the case anymore.[3] △

Example 3. Consider the following axiom—which we *can* express in our language, albeit not in its universal fragment:

$$\begin{array}{l} \forall_{\text{P}} p. \exists_{\text{N}} i. \forall_{\text{N}} j. [\, (top^{\text{A}}_{p,i} = j) \;\rightarrow\; (p \rhd (i,j)) \,] \wedge \\ \forall_{\text{P}} p. \exists_{\text{N}} j. \forall_{\text{N}} i. [\, (top^{\text{B}}_{p,j} = i) \;\rightarrow\; (p \rhd (i,j)) \,] \end{array}$$

It encodes the requirement that for every profile there should be at least one agent from each group who gets assigned to her most preferred partner. This is clearly possible for $n = 3$ but impossible for $n = 2$. △

We conclude this section with a useful reformulation of the Preservation Theorem that shows that impossibilities are preserved as we move up (while possibilities, as we have seen, are preserved as we move down).

Corollary 4. *If there exists no matching mechanism of dimension k that satisfies all axioms in a given set Φ of universal axioms, then the same is true for all dimensions $n \geq k$.*

4 Automated Search for Impossibility Results

In this section we describe our approach for automating the search for impossibility theorems via SAT solving and then present human-readable proofs for two such theorems. In fact, the use of SAT solving techniques is not restricted to proving impossibility theorems, and we briefly comment on some further uses of this technology along the way.

4.1 Approach

By Corollary 4, to prove an impossibility theorems for a set of universal axioms that applies to all dimensions $n \geq k$, it suffices to prove it for k. We can do the latter by translating the axioms into propositional formulas (which is possible for a fixed dimension, by rewriting universally quantified formulas as conjunctions, and so forth) and then checking the conjunction of all those formulas for satisfiability.

The translation to propositional logic (in CNF) is straightforward and similar to prior work using SAT solving for the analysis of voting rules (Geist and Peters 2017). For every profile p and every pair of indices $i, j \in \{1, \ldots, n\}$ we introduce a propositional variable $x_{p \rhd (i,j)}$, which should be set to *true* if and only if a_i should get matched with b_j in profile p.[4] For example, stability can be expressed as follows:

$$\bigwedge_{p} \bigwedge_{i} \bigwedge_{j} \bigwedge_{\substack{i' \text{ s.t. } p \text{ has} \\ a_i \succ_{b_j} a_{i'}}} \bigwedge_{\substack{j' \text{ s.t. } p \text{ has} \\ b_j \succ_{a_i} b_{j'}}} \big(\neg x_{p \rhd (i,j')} \vee \neg x_{p \rhd (i',j)}\big)$$

Using a simple script (written in PYTHON) we can generate the formula in CNF corresponding to each of the axioms in

[1] In fact, we only need to show that $\mu^+, \alpha^+ \models \psi$ implies $\mu, \alpha \models \psi$. However, in the proof that follows, the step covering the negation operator relies on the other direction being true as well.

[2] This is the case, for instance, for the original proof of the result by Roth (1982) we are going to discuss in Section 4.2.

[3] These two claims can be easily verified using the techniques we are going to introduce in Section 4.1 (Endriss 2019).

[4] As there are $(n!)^{2n}$ profiles, the number of variables required is $(n!)^{2n} \cdot n^2$. For $n = 4$ this figure is roughly $1.76 \cdot 10^{12}$, meaning that this approach will be hardly feasible for dimensions $n \geq 4$. Fortunately, inspection of the social choice literature suggests that many (though of course not all) interesting phenomena in this domain occur when changing a relevant parameter from 2 to 3.

$$\begin{tikzcd}
p_4 = \begin{pmatrix} \underline{213} & 123 \\ 132 & \underline{123} \\ 312 & 213 \end{pmatrix} & & & & p_7 = \begin{pmatrix} \underline{123} & \underline{123} \\ 132 & 123 \\ 312 & 231 \end{pmatrix} \\
p_5 = \begin{pmatrix} \underline{321} & 123 \\ 132 & 123 \\ 312 & \underline{123} \end{pmatrix} & p_1 = \begin{pmatrix} 321 & 123 \\ 132 & 123 \\ 312 & 213 \end{pmatrix} \arrow[ul, "a_1"'] \arrow[l, "b_3"'] \arrow[dl, "b_1"] & p_0 = \begin{pmatrix} 312 & 123 \\ 132 & 123 \\ 312 & 213 \end{pmatrix} \arrow[l, "a_1"'] \arrow[r, "b_3"] \arrow[d, "a_1"] & p_2 = \begin{pmatrix} 312 & 123 \\ 132 & 123 \\ 312 & 231 \end{pmatrix} \arrow[ur, "a_1"] \arrow[r, "a_2"] \arrow[dr, "b_3"'] & p_8 = \begin{pmatrix} 312 & 123 \\ \underline{312} & 123 \\ 312 & \underline{231} \end{pmatrix} \\
p_6 = \begin{pmatrix} 321 & \underline{213} \\ \underline{132} & 123 \\ 312 & 213 \end{pmatrix} & & p_3 = \begin{pmatrix} \underline{123} & \underline{123} \\ 132 & 123 \\ 312 & 213 \end{pmatrix} & & p_9 = \begin{pmatrix} 312 & 123 \\ 132 & 123 \\ \underline{312} & \underline{312} \end{pmatrix}
\end{tikzcd}$$

Figure 1: Profiles involved in the proof of Theorem 5

Table 1 (Endriss 2019). Similarly, we can generate one further formula in CNF to encode the requirement that every a is matched to at least one b, and every b is matched to at most one a. We can now use a SAT solver, such as PicoSAT (Biere 2008), to check whether the conjunction of this latter formula and the formulas encoding a set of axioms of interest is satisfiable for, say, $n = 3$. If it is not, then we have found an impossibility theorem for $n \geq 3$.

At this point the reader may raise an objection or two: What if our script does not in fact generate the correct formula in CNF? What if the SAT solver has a bug? While it is not unreasonable to assume that the latter objection is unlikely to have much impact on the correctness of the results generated by our approach and while it is possible, at least in principle, to proof-read a script for generating the CNF in the same way as one would proof-read a paper, the accepted gold standard for certifying the correctness of a mathematical statement still is, and always will be, an elementary human-readable proof.

The good news is that SAT solving technology offers a solution to this problem. We can use a tool such as PicoMUS, which is part of the PicoSAT distribution, to automatically extract a *minimal unsatisfiable subset* (MUS) from a given unsatisfiable set of clauses. Provided it is sufficiently small, we can then turn this MUS into a human-readable proof.

4.2 Results

By a seminal result in the theory of matching due to Roth (1982), it is impossible to design a one-to-one matching mechanism that is both stable and two-way strategyproof. To be precise, Roth proved this impossibility theorem for a richer model, in which agents are permitted to report so-called *incomplete preferences*. Intuitively speaking, increasing the range of profiles we want our mechanism to operate on might increase the chance of running into an impossibility. Indeed, while Roth's result holds even for $n = 2$ for the model he considers, this is not the case for our model: there exists a matching mechanism for $n = 2$ that is both stable and two-way strategyproof. This is well known and easy to verify. In fact, with a SAT solver at hand, we can automatically design such a mechanism by simply running the solver on a formula in CNF that encodes the requirements of stability and strategyproofness, and then inspecting the model returned (Endriss 2019). We can also count the number of distinct models of this formula. This reveals that there are exactly four distinct mechanisms for $n = 2$ that are both stable and two-way strategyproof (Endriss 2019).

However, for $n \geq 3$ it is well known that Roth's result applies even when all preferences must be complete, as is the case for our model. We are now going to prove a strengthened variant of this result: the impossibility continues to apply even when we weaken stability to top-stability.

Theorem 5. *For $n \geq 3$, there exists no matching mechanism that is top-stable and two-way strategyproof.*

Proof. As we have seen that both top-stability (TS) and strategyproofness (SP) are universal axioms, due to Corollary 4, we are done if we can prove the claim for $n = 3$. Running PicoMUS on the corresponding CNF returns an MUS of 23 clauses that together make reference to ten different profiles. These profiles are shown in Figure 1, using a simplified notation to describe preferences. For example, in profile p_3 agent a_1 reports $b_1 \succ b_2 \succ b_3$, while agent b_3 reports $a_2 \succ a_1 \succ a_3$. Six of the clauses in the MUS are instances of TS, applied to the three profiles on the very left and the three profiles on the very right. Nine clauses are instances of SP; each of them corresponds to one of the nine arrows in the picture (pointing from the truthful profile to the profile we obtain when the agent whose name is used to label the arrow tries to manipulate). The remaining seven clauses encode (relevant aspects of) the fact that any agent must be matched with exactly one other agent.

Inspection of Figure 1 reveals that it is impossible to assign a partner to agent a_1 in profile p_0 without violating at least one of our requirements. Let us now make this line of argument explicit. We first show that $(a_1, b_1) \notin \mu(p_0)$ must be the case by exploring the profiles on the left:

- $(a_1, b_2) \in \mu(p_4)$ by TS
- $(a_1, b_2) \in \mu(p_4)$ implies $(a_1, b_1) \notin \mu(p_1)$ by SP
- $(a_1, b_3) \in \mu(p_5)$ by TS
- $(a_1, b_3) \in \mu(p_5)$ implies $(a_3, b_3) \notin \mu(p_1)$ by SP
- $(a_2, b_1) \in \mu(p_6)$ by TS
- $(a_2, b_1) \in \mu(p_6)$ implies $(a_3, b_1) \notin \mu(p_1)$ by SP
- $(a_1, b_3) \in \mu(p_1)$ follows as the only remaining option
- $(a_1, b_3) \in \mu(p_1)$ implies $(a_1, b_1) \notin \mu(p_0)$ by SP

For instance, the second step above follows when we consider the possibility of agent a_1 manipulating in profile p_1 (by moving to p_4) to get matched with b_2 instead of b_1. Next, we use analogous reasoning to establish $(a_1, b_3) \not\in \mu(p_0)$:

- $(a_1, b_1) \in \mu(p_7)$ by TS
- $(a_1, b_1) \in \mu(p_7)$ implies $(a_1, b_2) \not\in \mu(p_2)$ by SP
- $(a_2, b_3) \in \mu(p_8)$ by TS
- $(a_2, b_3) \in \mu(p_8)$ implies $(a_2, b_2) \not\in \mu(p_2)$ by SP
- $(a_3, b_3) \in \mu(p_9)$ by TS
- $(a_3, b_3) \in \mu(p_9)$ implies $(a_1, b_3) \not\in \mu(p_2)$ by SP
- $(a_2, b_3) \in \mu(p_2)$ follows as the only remaining option
- $(a_2, b_3) \in \mu(p_2)$ implies $(a_1, b_3) \not\in \mu(p_0)$ by SP

Finally, we get that $(a_1, b_2) \not\in \mu(p_0)$ as a consequence of $(a_1, b_1) \in \mu(p_3)$ and $(a_1, b_1) \in \mu(p_3)$ implying $(a_1, b_2) \not\in \mu(p_0)$, which are instance of TS and SP, respectively. □

Our second impossibility theorem shows that we cannot design a stable mechanism that would be fair in the sense of treating the two groups of agents in a symmetric manner.

Theorem 6. *For $n \geq 3$, there exists no matching mechanism that is stable and gender-indifferent.*

Proof. As we have seen that both stability and gender-indifference are universal axioms (and as stability entails top-stability), due to Corollary 4, we are done if we can prove the claim for $n = 3$. Running PicoMUS on the corresponding CNF returns an MUS of 13 clauses. Seven of these clauses encode the fact that any matching mechanism μ must be well-defined on the following profile (which we call p):

$$\begin{array}{lll} a_1 : b_2 \succ b_3 \succ b_1 & \quad & b_1 : a_2 \succ a_3 \succ a_1 \\ a_2 : b_3 \succ b_1 \succ b_2 & \quad & b_2 : a_3 \succ a_1 \succ a_2 \\ a_3 : b_1 \succ b_2 \succ b_3 & \quad & b_3 : a_1 \succ a_2 \succ a_3 \end{array}$$

Note that p is symmetric in the sense that, if we swap A and B (the operation at the heart of the definition of gender-indifference), then we end up with the very same profile p.

Note that (for $n = 3$) there are six possible matchings a mechanism μ might return for a given profile. The remaining six clauses in the MUS each exclude one of those matchings as a possibility. Four of them are instances of stability and two are instances of gender-indifference. Indeed, stability rules out the following four matchings:

- $\{(a_1, b_1), (a_2, b_2), (a_3, b_3)\}$ is blocked by (a_1, b_2)
- $\{(a_1, b_1), (a_2, b_3), (a_3, b_2)\}$ is blocked by (a_1, b_3)
- $\{(a_1, b_2), (a_2, b_1), (a_3, b_3)\}$ is blocked by (a_3, b_2)
- $\{(a_1, b_3), (a_2, b_2), (a_3, b_1)\}$ is blocked by (a_2, b_1)

But the remaining two matchings are inconsistent with the requirements imposed by gender-indifference:

- $\{(a_1, b_2), (a_2, b_3), (a_3, b_1)\}$ is not admissible, because if we match a_1 to b_2, then we also must match b_1 to a_2.
- $\{(a_1, b_3), (a_2, b_1), (a_3, b_2)\}$ is not admissible, because if we match a_2 to b_1, then we also must match b_2 to a_1.

Thus, any mechanism that is stable and gender-indifferent will have to remain undefined on p. In other words, there exists no mechanism that meets all our requirements. □

	# Clauses	# Variables	Build	SAT	MUS
Theorem 5	$4,805,568$	$419,904$	$34s$	$1s$	$64s$
Theorem 6	$1,399,680$	$419,904$	$9s$	$2s$	$19s$

Table 2: Resources required to prove impossibility theorems

Theorem 6 may appear to contradict a result due to Pini et al. (2011) who claim to have found a general approach for turning any given matching mechanism that is stable into a matching mechanism that is both stable and gender-indifferent. The source of this mismatch between results is that Pini et al. work with a different definition of gender-indifference: under their alternative definition, μ is gender-indifferent if $\mu(p) = \mu(p')$ whenever p' can be obtained from profile p by swapping A and B. So their definition is different from the one originally proposed by Masarani and Gokturk (1989), which is the one we use in this paper and which arguably is the most appropriate definition. While the definition of Pini et al. does encode some notion of what one might want to call "gender-independence" (the outcome does not depend on which group is called A and which group is called B), it arguably does not quite qualify as a fairness property, as it does not exclude the possibility that one group is greatly favoured when computing outcomes.

Table 2 summaries the computational resources required to prove the two impossibility theorems discussed in this section (to automatically prove the base case for $n = 3$). Besides the size of the formula φ in CNF that needs to be analysed, we report the time it takes to execute the Python script to build φ, the time it takes to run PicoSAT to verify that φ is not satisfiable, and the time it takes to run PicoMUS to compute an MUS from which we can read off a human-verifiable proof. All runtimes have been measured on a mid-range desktop machine (running an Intel Core i5-7500 processor at 3.40GHz with 8GB of memory).

Finally, we note that Theorem 5 and Theorem 6 are maximally strong, in the sense that neither can be strengthened further by lowering the bound to $n \geq 2$ or by omitting one of the axioms involved. Furthermore, Theorem 5 cannot be strengthened by weakening two-way strategyproofness to one-way strategyproofness, and Theorem 6 cannot be strengthened by weakening stability to top-stability. Each of these claims can be verified at the press of a button by running a SAT solver on the corresponding formula and finding that formula to be satisfiable (Endriss 2019).

5 Conclusion

We have extended the approach for proving impossibility theorems with the help of a SAT solver developed in the field of computational social choice over the past decade to the new domain of matching. Our main result is a meta-result regarding the approach developed: the Preservation Theorem shows that we can reduce the proof of any conjectured impossibility theorem involving only axioms that have a universal form to a proof for a fixed domain size—and the latter typically can be fully automated using SAT solving technology. Finally, we have used our approach to derive two new impossibility theorems for one-to-one matching mechanisms that are of some interest in their own right.

Future work should be directed towards the challenge of extending our approach to a wider range of scenarios in which matching mechanisms are used, notably the case of one-to-many matching. Another worthwhile challenge would be to look for applications of the approach beyond the task of proving impossibility theorems (such as the task of automatically designing mechanisms with attractive properties), which we have only briefly hinted at in this paper.

References

Biere, A.; Heule, M.; and van Maaren, H., eds. 2009. *Handbook of Satisfiability*. IOS Press.

Biere, A. 2008. PicoSAT essentials. *Journal on Satisfiability, Boolean Modeling and Computation* 4:75–97.

Brandt, F., and Geist, C. 2016. Finding strategyproof social choice functions via SAT solving. *Journal of Artificial Intelligence Research* 55:565–602.

Brandt, F.; Conitzer, V.; Endriss, U.; Lang, J.; and Procaccia, A. D., eds. 2016. *Handbook of Computational Social Choice*. Cambridge University Press.

Brandt, F.; Geist, C.; and Peters, D. 2017. Optimal bounds for the no-show paradox via SAT solving. *Mathematical Social Sciences* 90:18–27.

Drummond, J.; Perrault, A.; and Bacchus, F. 2015. SAT is an effective and complete method for solving stable matching problems with couples. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI-2015)*.

Economic Sciences Prize Committee. 2012. Stable allocation and the practice of market design. Technical report, Royal Swedish Academy of Sciences.

Endriss, U. 2019. Software and data for "Analysis of one-to-one matching mechanisms via SAT solving: Impossibilities for universal axioms". Zenodo. http://doi.org/10.5281/zenodo.3547826.

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

Geist, C., and Endriss, U. 2011. Automated search for impossibility theorems in social choice theory: Ranking sets of objects. *Journal of Artificial Intelligence Research* 40:143–174.

Geist, C., and Peters, D. 2017. Computer-aided methods for social choice theory. In Endriss, U., ed., *Trends in Computational Social Choice*. AI Access.

Hodges, W. 1997. *A Shorter Model Theory*. Cambridge University Press.

Masarani, F., and Gokturk, S. S. 1989. On the existence of fair matching algorithms. *Theory and Decision* 26(3):305–322.

Peters, D. 2018. Proportionality and strategyproofness in multiwinner elections. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS-2018)*.

Pini, M. S.; Rossi, F.; Venable, K. B.; and Walsh, T. 2011. Manipulation complexity and gender neutrality in stable marriage procedures. *Journal of Autonomous Agents and Multiagent Systems* 22(1):183–199.

Roth, A. E., and Sotomayor, M. A. O. 1990. *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*. Cambridge University Press.

Roth, A. E. 1982. The economics of matching: Stability and incentives. *Mathematics of Operations Research* 7(4):617–628.

Tang, P., and Lin, F. 2009. Computer-aided proofs of Arrow's and other impossibility theorems. *Artificial Intelligence* 173(11):1041–1053.