Obviously Strategy-Proof Implementation of Top Trading Cycles

Peter Troyan*
Department of Economics
University of Virginia†

Abstract

Top trading cycles (TTC) is a classic algorithm for allocation problems without transfers (e.g., school choice), and while there is a rich theoretical literature extolling its virtues, its use in practical applications is much rarer. Anecdotal evidence suggests that one possible explanation is that TTC is difficult for participants to understand. This paper formalizes this intuition by asking whether it is possible to implement TTC in an *obviously* strategy-proof (OSP) way. We identify an acyclicity condition that is both necessary and sufficient for TTC to be OSP. The acyclicity condition is unlikely to be satisfied in many practical applications, which may explain why TTC has a reputation of being difficult to understand and is rarely used, despite its many theoretically appealing properties.

*I would like to thank Itai Ashlagi, Yannai Gonczarowski, Fuhito Kojima, and Marek Pycia for helpful conversations relating to this project, as well as the editor and anonymous referees for comments that have improved the paper.

†P.O. Box 400182, Charlottesville, VA, 22904. Email: troyan@virginia.edu

1 Introduction

An important consideration when designing institutions to allocate scare resources is the incentives given to agents to report their preferences truthfully. For example, in school choice, it is well-known that the so-called 'Boston' mechanism gives students (or their parents) very clear incentives to lie about their true preference, and can harm students who may be unable to strategize appropriately. Because of this, economists have generally advised against using manipulable mechanisms such as the Boston mechanism in favor of mechanisms that ensure it is always an optimal strategy for agents to report their true preferences; in other words, they recommend that the mechanism be *strategy-proof.*[1]

There are three canonical strategy-proof ways to assign indivisible objects in the absence of transfers that are used in practice: Serial Dictatorship (SD), Deferred Acceptance (DA), and Top Trading Cycles (TTC). However, just because a mechanism is strategy-proof in theory does not mean that participants will be truthful in practice. The agents participating in it must also *understand* that it is strategy-proof.

One formalization of what it means for participants to "understand" that a mechanism is strategy-proof is the concept of *obvious strategy-proofness (OSP)* due to Li (2017). A mechanism is obviously strategy-proof if each agent has an obviously dominant strategy, which, informally, is a strategy such that even the worst possible outcome from following it is better than the best case from any possible deviation. Li (2017) shows that obviously

[1]Pathak and Sönmez (2008) argue that non-strategy-proof mechanisms may harm parents who are unable to strategize well, and incentive considerations were key concerns for the New York City and Boston school districts when deciding upon which mechanism to use (Abdulkadiroğlu et al., 2005a,b). At the same time, strategy-proofness is not without costs, and recent literature has begun to re-examine non-strategy-proof mechanisms such as the Boston mechanism and shows that it may outperform strategy-proof mechanisms from an ex-ante perspective in some settings, at least under equilibrium play (see, for example, Miralles (2009), Abdulkadiroğlu et al. (2011), Featherstone and Niederle (2014), and Troyan (2012)).

dominant strategies are those that can be seen as optimal by a cognitively limited agent who cannot engage in contingent reasoning.

For no-transfer settings like those considered in this paper, Li (2017) shows that serial dictatorships are obviously strategy-proof when implemented dynamically, and provides experimental evidence that rates of truth-telling are higher when using such a dynamic implementation compared to a static one, which is strategy-proof, but not obviously so. For deferred acceptance, Ashlagi and Gonczarowski (2016) show that acyclicty on the non-strategic side of the market in the sense of Ergin (2002) is sufficient for DA to be OSP-implementable for the strategic (proposing) side, though it is not necessary.[2] They also provide an impossibility result that shows it is always possible to choose (cyclical) preferences on the non-strategic side such that no stable mechanism will be OSP for the strategic side. In the class of indivisible-good economies considered by Shapley and Scarf (1974), in which each agent is initially endowed with ownership of exactly one good but may prefer others, Li (2017) shows that the classic TTC mechanism (which determines a core allocation) is not obviously strategy-proof.[3]

In this paper, we consider a more general class of TTC mechanisms in which each object is associated with a strict priority relation over the agents. These priority relations are used to determine how "property rights" evolve throughout the mechanism, and each priority structure defines a TTC mechanism. Our main result provides a tight characterization for OSP-implementability of TTC using a condition on the priority structure called *weak acyclicity*. Essentially, our definition of a (strong) cycle is three agents, i, j, k, and three objects, x, y, z, such that each agent outranks (has higher

[2]Roth (1982a) and Dubins and Freedman (1981) were the first to show that DA is strategy-proof in the standard sense for the proposing side. It is also well-known that if both sides of the market are strategic (two-sided matching), then there is no mechanism that is stable and strategy-proof for all agents on both sides. We consider allocation problems in which only one side of the market is strategic, while the other side consists of objects to be consumed.

[3]Roth (1982b) was the first to show that TTC is strategy-proof in the standard sense.

priority than) the other two at one of the objects. When no such agents and objects exist, then the priority structure is weakly acyclic. We show that weak acyclicity is both necessary and sufficient for the associated TTC mechanism to be obviously strategy-proof.[4]

Standard serial dictatorships are recovered as a special case of our model by giving each object the exact same priority relation (equivalent to the serial dictatorship ordering), which is trivially weakly acyclic. More generally, the characterization result shows that obviously strategy-proof implementation is possible for some cases beyond this, as it allows for what we call a *dual dictatorship*. In a dual dictatorship, the set of objects to be allocated is initially divided among at most two agents. Each agent becomes the "dictator" over her endowment set, in the sense that if her most preferred object is in her endowment, she receives it and leaves the market. If neither agent is endowed with her most preferred object, then it must be endowed to the other dictator, in which case the agents swap these two objects and leave the market. The objects remaining are then passed down to at most two new agents, who become the next dual dictators, and the process is repeated. TTC under any weakly acyclic priority structure is equivalent to such a dual dictatorship, which is therefore obviously strategy-proof.[5]

One main takeaway from our results is that while we show that OSP implementation is formally possible beyond the already-known case of a simple serial dictatorship, dual dictatorships fully exhaust how far this can be extended. This is of particular importance because from a theoretical perspective, TTC is one of the most popular mechanisms for allocation problems

[4]We use the terminology weak acyclicity because the condition is reminiscent of other well-known acyclicity conditions due to Ergin (2002) and Kesten (2006), though it is weaker than both of them (that is, more priority structures satisfy our acyclicity condition than those of Ergin (2002) or Kesten (2006), in a set inclusion sense).

[5]Bogomolnaia et al. (2005) characterize a class of mechanisms called *bipolar serial dictatorships* as the unique class of mechanisms that satisfy certain natural axioms. While similar in spirit, bipolar serial dictatorships only allow the first two agents to be "bi-dictators," and after that, the mechanism proceeds as a standard serial dictatorship. Since our class of mechanisms is more general, we use the "dual dictatorship" terminology.

without transfers, and there is a long and rich literature devoted to analyzing and characterizing it and other closely related mechanisms.[6] In school choice, for example, TTC is one of the most commonly-proposed mechanism designs, starting with Abdulkadiroğlu and Sönmez (2003)'s extension of the original Shapley-Scarf mechanism to accommodate important features of school choice environments. Despite its appealing theoretical properties, the use of TTC in practice is rare.[7] One explanation often put forth is that TTC is a very difficult mechanism for both parents and school administrators to understand, even though it is strategy-proof at a theoretical level, and thus parents can do no better than to just report their true preferences. Pathak (2016) provides anecdotal evidence supporting this view after working with various school districts and concluding that the difficulty of explaining TTC is one of the main reasons it has not been adopted by more cities.[8] Our results provide a formal justification for these anecdotal observations: since very few priority structures observed in practice are weakly acyclic, TTC will generally not be obviously strategy-proof, and so the optimality of truth-telling may be difficult to recognize.

Most closely related to our paper is the aforementioned work of Ashlagi and Gonczarowski (2016). Much like the current paper, the main takeaway from Ashlagi and Gonczarowski (2016) is that OSP-implementation of one of the canonical strategy-proof matching mechanisms (in their case, DA) is not possible, except for very restrictive special cases.[9] While DA sometimes

[6]Besides the papers already mentioned, a partial and incomplete summary of the literature on TTC includes Ma (1994), Abdulkadiroğlu and Sönmez (1999), Pápai (2000), Morrill (2013), Hakimov and Kesten (2014), Morrill (2015), Dur and Ünver (2015), and Pycia and Ünver (2017).

[7]One prominent example is New Orleans, where TTC was used for one year, before being abandoned (see Abdulkadiroglu et al., 2017).

[8]This reasoning is also present in Leshno and Lo (2017) and Dur and Morrill (2017), who attempt to provide alternative "simpler" descriptions of TTC based on priority cutoffs and competitive equilibrium prices, respectively. Note, however, that these cutoffs/prices are only known ex-post, after everyone submits their preferences and the mechanism itself is run.

[9]Ashlagi and Gonczarowski (2016) provide a sufficient and a (different) necessary

has a reputation of being "easier" to understand than TTC, recent evidence suggests that this may not be the case empirically. Indeed, real-world agents often make mistakes under DA, both in the lab and in high-stakes decisions decisions in the field (see, e.g., Hassidim et al., 2015; Rees-Jones, 2017; Shorrer and Sóvágó, 2017).[10] Further, more recent work by Pycia and Troyan (2016) suggests that in matching environments, even some OSP mechanisms may not necessarily be simple to understand. They introduce a strengthening they call strong obvious strategy-proofness, and use this along with efficiency and fairness axioms to provide a characterization of the popular Random Priority mechanism (also called the Random Serial Dictatorship). Since our results show that TTC is unlikely to be even OSP, it will also fail to satisfy the definition of Pycia and Troyan (2016) as well.

Besides those already discussed, there are several other related papers in the very new but rapidly growing literature on obvious strategy-proofness. Bade and Gonczarowski (2016) characterize the class of OSP and Pareto efficient mechanisms in a variety of settings, including object allocation, while Arribillaga et al. (2017) study OSP voting rules on the domain of single-peaked preferences. Mackenzie (2017) proves an analogue of the revelation principle for OSP implementation, and Zhang and Levin (2017a,b) provide decision-theoretic foundations for obvious dominance. The rapid expansion of this literature suggests that deepening our understanding of what mech-

condition for DA to be OSP-implementable, whereas we provide one condition that is both necessary and sufficient for TTC. The sufficient condition identified by Ashlagi and Gonczarowski (2016) is stronger than weak acyclicity, while the necessary condition is weaker than it. Weak acyclicty turns out not to be sufficient for OSP implementation of DA, and so the right condition to fill the "gap" of Ashlagi and Gonczarowski (2016) remains an open question.

[10]Note that OSP is not a metric that ranks mechanisms by degree of simplicity, but rather is a binary classification, and therefore we make no formal statements about one mechanism being "easier" than another, beyond saying that neither will be OSP-implementable except for very special cases. Also, while there is increasing empirical evidence of mistakes under DA, there is little field evidence on behavior under TTC, since it has rarely been used in practice. Chen and Sönmez (2006) present lab evidence of low rates of truth-telling under TTC.

anisms are simple to play is clearly an interesting research agenda, and we expect there to be even more follow up papers in the future.

2 Preliminaries

2.1 Model

The primitives of the model are $(I, X, \succ_X)$, where I is a set of **agents**, X is a (finite) set of indivisible **objects**, and $\succ_X = (\succ_x)_{x \in X}$ is a **priority structure**, where $\succ_x$ denotes the (strict) **priority relation** of object $x \in X$ over the set of agents I. Throughout the paper, we will use i, j, k, ℓ to denote generic agents in I, and x, y, z, w to denote generic objects in X. Each agent demands exactly one object, and $|I| = |X| = N$; we refer to N as the **size** of the market.[11] An **allocation** is a bijection $\mu : I \to X$, where $\mu(i)$ denotes agent i's assignment under allocation μ, and $\mu^{-1}(x)$ denotes the agent who is assigned to object $x \in X$. Let $\mathcal{A}$ denote the set of all allocations.

Each agent $i \in I$ has a strict **preference ranking** P_i over the elements of X; we use R_i to denote the weak preference relation corresponding to P_i.[12] Let $\mathcal{P}$ denote the set of all possible strict preference rankings over X. We will sometimes refer to $P_i \in \mathcal{P}$ as agent i's **type**. We use $P_I = (P_i)_{i \in I}$ to denote an entire **preference profile**, consisting of one preference relation for each agent. Let $\mathcal{P}^N$ denote the set of all possible preference profiles. Given a preference profile P_I, an allocation μ is said to be **Pareto efficient** if there is no other allocation ν such that $\nu(i) R_i \mu(i)$ for all $i \in I$, and $\nu(i) P_i \mu(i)$ for some $i \in I$.

A **social choice rule** is a mapping from preference profiles into allocations $f : \mathcal{P}^N \to \mathcal{A}$, where $f(P_I)$ denotes the allocation under preference profile P_I, and $f_i(P_I)$ denotes agent i's assignment under this allocation. A

[11] While we assume there is only one copy of each object, our results can be easily generalized to multiple copies.

[12] That is, $x R_i y$ if either $x P_i y$ or $x = y$.

choice rule is just a systematic way to determine what allocation should be implemented for every possible preference profile of the agents. An example of a choice rule is the top trading cycles rule (which will be defined formally below). A social choice rule is said to be **Pareto efficient** if $f(P_I)$ is a Pareto efficient allocation for all $P_I \in \mathcal{P}^N$.

Since agent preferences are their private information which must be reported, evaluating Pareto efficiency in this way is only meaningful if we are confident that we are able to elicit the true preferences of the agents. We say a social choice rule f is **strategy-proof** if $f_i(P_i', P_{-i}) R_i f_i(P_i, P_{-i})$ for all $i \in I, P_{-i} \in \mathcal{P}^{N-1}, P_i, P_i' \in \mathcal{P}$. Strategy-proofness of a social choice rule can be thought of as asking the agents to play a direct mechanism where each agent is asked to submit her preference ranking, and the mechanism translates these reports into outcomes via the social choice rule f. Strategy-proofness requires that in this direct mechanism, for any agent i, reporting her true preferences P_i always results in a weakly better outcome than any other report P_i', for all possible preference profiles P_{-i} that could be submitted by the other agents.

2.2 Obvious strategy-proofness

Strategy-proofness is often an important property of social choice rules in many contexts, because they are strategically "simple", at least at a formal level. However, Li (2017) challenges the idea that strategy-proof mechanisms are necessarily strategically simple, and indeed, there is evidence that not all strategy-proof mechanisms are created equal: some yield higher rates of actual truth-telling in practice than others. He thus introduces the concept of an *obviously strategy-proof* mechanism as a refinement of strategy-proofness that distinguishes between strategy-proof mechanisms that are strategically simple, and those that may not be.

Contrary to the standard definition of strategy-proofness (which only applies to simple normal form games), obvious strategy-proofness is a solution

concept that applies to extensive form games. A strategy S_i for an agent i in an extensive form game is said to *obviously dominate* another strategy S_i' if, at every history at which i is called upon to play, the worst possible outcome from following S_i is at least as good as the best possible outcome from following S_i'. The formal definition given below is a simplification of Li's definition found in Ashlagi and Gonczarowski (2016). We first define a mechanism, and then define obvious strategy-proofness.

Definition 1. An (extensive-form allocation) **mechanism** G consists of:

1. A rooted tree R, where
   (a) r denotes the root node of R
   (b) $L(R)$ denotes the set of leaves (end nodes) of R
   (c) $V(R)$ denotes the set of internal nodes (vertices) of R
   (d) For $v \in V(R)$, $E(v)$ denotes the set of outgoing edges from v
   (e) $\bar{E}(R) = \cup_{v \in V(R)} E(v)$ denotes the set of edges of R
   (f) For $e \in \bar{E}(R)$, $\rho(e)$ denotes the source node of e
2. A function $h : L(R) \to \mathcal{A}$ from the leaves of R to allocations.
3. A function $\iota : V(R) \setminus L(R) \to I$ describing which agent is to act at each internal node
4. A function $\alpha : \bar{E}(R) \to 2^{\mathcal{P}}$ such that:
   (a) For all $e \neq e'$ such that $\rho(e) = \rho(e')$, $\alpha(e) \cap \alpha(e') = \emptyset$
   (b) For any node $v \in V(R)$, $\cup_{e \in E(v)} \alpha(e) = \alpha(e')$, where e' is the most recent edge along path from r to v such that $\iota(\rho(e')) = \iota(v)$, or, if no such edge edge exists, then $\cup_{e \in E(v)} \alpha(e) = \mathcal{P}$.

Intuitively, a mechanism is an extensive form game where at each node, one agent is called on to take an action. Each edge e outgoing from a node v is interpreted as one possible "action" available to agent $\iota(v)$ at v, and $\alpha(e)$ is interpreted as the set of types of agent $\iota(v)$ that are recommended to take action e at node v (in the rest of the paper, we use the terms "edge" and "action" interchangeably). Given a preference profile P_I, we can find the allocation implemented by any such mechanism under preferences P_I by following the edges that correspond to the preference profile at each node.[13] Given a preference profile P_I, if a node v is on this path, then P_I is said to **pass through** v. Two preference relations P_i and P_i' for agent i are said to **diverge** at a node $v \in V(R)$ if there exists two distinct edges e, e' outgoing from v such that $P_i \in \alpha(e)$ and $P_i' \in \alpha(e')$.

For any mechanism G, we use f^G to denote the social choice function implemented by G, where $f^G(P_I) \in \mathcal{A}$ is the allocation found via the above process. In addition, we use $f_i^G(P_I) \in X$ to denote agent i's object under allocation $f^G(P_I)$.

Definition 2. An extensive-form allocation mechanism G is said to be **obviously strategy-proof (OSP)** if, for all i, all nodes v such that $\iota(v) = i$, and every $P_I, P_I' \in \mathcal{P}^N$ such that P_i and P_i' diverge at v, we have $f_i^G(P_I) R_i f_i^G(P_I')$.

In words, obvious strategy-proofness means that, at each node v where an agent $\iota(v) = i$ is called to act, the worst possible outcome from taking the action associated with her true preferences (i.e., the action e such that $P_i \in \alpha(e)$ at node v) is at least as good as the best possible outcome from following any other action $e' \neq e$.

Definition 3. A social choice rule $f : \mathcal{P}^N \to \mathcal{A}$ is **OSP-implementable** if there exists an obviously strategy-proof mechanism G such that $f = f^G$.

[13] This assumes that whenever agent i has to choose, he chooses the action (edge) that corresponds to his true preferences P_i. The idea behind constructing an obviously strategy-proof mechanism is to ensure it is always optimal (in a precise sense) for i to do so.

3 TTC

3.1 Definition

The top trading cycles (TTC) algorithm is a classic algorithm first proposed by Shapley and Scarf (1974). The original algorithm applies to so-called *housing markets*, which are models where a set of agents have preferences over a set of objects (often called "houses" for concreteness), and there are no monetary transfers. Each agent begins by being endowed with one house. An outcome is a reallocation of the houses amongst the agents such that each receives exactly one house. While the original TTC algorithm applies only to housing markets where each agent begins by owning one object, more recently, TTC has attracted attention because it's basic intuition can be generalized to apply to a wide array of general allocation problems, such as kidney exchange, the allocation of public housing units, and school choice. Indeed, in their seminal paper, Abdulkadiroğlu and Sönmez (2003) generalize Shapley and Scarf's mechanism to school choice problems and propose it as an alternative to flawed mechanisms that are used by some school districts. The formal definition of TTC that we give follows Abdulkadiroğlu and Sönmez (2003).

Definition 4. Given a priority structure $\succ_X$, the **top trading cycles** social choice rule, $T^{\succ_X} : \mathcal{P}^N \to \mathcal{A}$, is defined by the following algorithm for any preference profile P_I:

Step 1 Each agent i points to his most-preferred object according to P_i, and each object x points to the agent who has highest priority according to $\succ_x$. There is at least one cycle. Each agent in a cycle is assigned to the object he is pointing to, and the agent and that object are removed. If any agents and objects remain, continue to step 2; otherwise, end the algorithm.

Step $k, k \geq 2$ Each agent i who remains points to his most preferred object according to P_i *among those objects that remain*, and each object points to the highest priority agent according to $\succ_x$ *among those* agent*s that remain*. There is at least one cycle. Each agent in a cycle is assigned to the object he is pointing to, and the agent and that object are removed. If any agents and objects remain, continue to step $k+1$; otherwise, end the algorithm.

The algorithm ends whenever there are no agents/objects remaining.

There are two important special cases that deserve particular attention, as we will return to them later. Let $Top(\succ_x) = \{i \in I : i \succ_x j \text{ for all } j \in I \backslash \{i\}\}$ denote the agent who has the highest priority at x according to $\succ_x$.

(SD) *For all* $x, y \in X$, $\succ_x = \succ_y$: In this case, $T^{\succ_X}$ reduces to the familiar *serial dictatorship (SD)* algorithm, in which agents are ordered according to the common priority relation and each is allowed to pick, in this order, her favorite good that is still available when it is her time to choose.

(HM) *For all* $x \neq y$, $Top(\succ_x) \neq Top(\succ_y)$: In this case, $T^{\succ_X}$ reduces to the simplified TTC algorithm defined by Shapley and Scarf (1974) for a *housing market (HM)*, where each agent begins by "owning" the object where she has highest priority.

The main question we ask is under what priority structures $\succ_X$ is the TTC social choice rule $T^{\succ_X}$ OSP-implementable. It is known that if $\succ_X$ satisfies condition (SD) above, then it is OSP-implementable, while if $T^{\succ_X}$ satisfies condition (HM), then it is not OSP-implementable (Li (2017)). In the next section, we first present an example that satisfies neither (SD) nor (HM), and for which $T^{\succ_X}$ is OSP-implementable. Then, we discuss the key features of $\succ_X$ driving this result, and provide a general necessary and sufficient condition on $\succ_X$ for OSP-implementability to hold.

3.2 Example

We start with an example of a nontrivial (i.e., $T^{\succ_X}$ does not reduce to a simple serial dictatorship) priority structure such that $T^{\succ_X}$ is OSP-implementable. This example is also important because it highlights a method for OSP-implementing TTC that we will use in the main proof below.[14]

Example 1. Consider a market $(I, X, \succ_X)$ that consists of three agents $I = \{i, j, k\}$ and three objects $X = \{x, y, z\}$. The priority relation $\succ_x$ for each $x \in X$ is given in the following table:

$\succ_x$	$\succ_y$	$\succ_z$
i	i	j
j	k	i
k	j	k

Now, given any preference profile P_I, the following mechanism G implements the TTC rule $T^{\succ_X}$ and is obviously strategy-proof:

1. Ask i if her top choice is x, y, or z. If i responds x, assign her to x, and go to step 1(a); if i responds y, assign her to y and go to step 1(b). If i responds z, go to step 2.

   (a) Ask j if she prefers z to y. If yes, assign j to z and k to y, and end the mechanism. If no, go to step 1(b)(i)

      i. Ask agent k if she prefers y to z. If yes, assign k to y and j to z; if no, assign k to z and j to y, and end the mechanism.

   (b) Ask j if she prefers x to z. If yes, assign j to x and k to z; otherwise, assign j to z and k to x, and end the mechanism.

[14] While this method of OSP-implementing TTC is inspired by Ashlagi and Gonczarowski (2016), who study OSP-implementability of stable matching mechanisms, we are not concerned with stability, and TTC choice rules are not in this class.

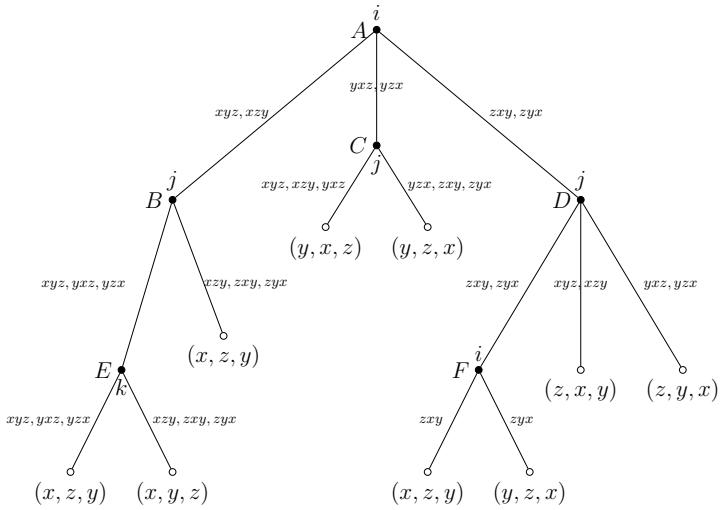

Figure 1: Tree representation of the mechanism G from Example 1.

2. Ask j if her top choice is x, y, or z. If j responds x, assign i to z, j to x, and k to y; if j responds y, assign i to z, j to y, and k to x; if j responds z, assign her to z, and go to 2(a).

   (a) Ask i if she prefers x to y. If yes, assign i to x and k to y, and end the mechanism. If no, then assign i to y and k to x, and end the mechanism.

The tree corresponding to this game is in Figure 1. In the figure, the edges are labeled with the types of agents who follow that edge. To shorten notation, we use "xyz" to represent the type with preferences $xPyPz$. The labels of the end nodes represent the matching implemented at that history, where the first entry corresponds to agent i, the second to agent j, and the third to agent k; e.g., (x, y, z) means agent i receives x, j receives y, and k receives z. The labels of the internal nodes represent the agent who acts at that node (i,j, or k). In addition, for easy reference, we label each internal node with a unique letter (A, B, C, D, E, F).

To see that this mechanism is OSP, it is easiest to look at the tree. We must check at every node, and for every type instructed to follow a certain edge at that node, the worst possible outcome she can receive is no worse than the best possible outcome from any other edge. The argument is most complex for agent i, and so we focus on her. At node F, there are two types of agent i who may reach node F, type zxy and zyx. At F, i's action is effectively a choice between x and y, and so it is clearly obviously dominant for type zxy to follow the left edge and type zyx to follow the right edge. So, consider node A, if i's first choice is x, then following the left edge and middle edge, respectively, guarantees she gets her top choice, no matter what j or k do, and so these strategies are clearly obviously dominant for these types. Finally, consider agent i of type zxy at node A. If she follows the strategy indicated, at node A, her set of possible outcomes is $\{x, z\}$, and so the worst possible outcome is x.[15] If she were to follow the left edge, she is guaranteed x for sure, which is equivalent, and if she is to follow the middle node, she is guaranteed y for sure, which is worse than x. Thus, following the strategy indicated is obviously dominant. A similar argument shows the same for type zyx. It is then simple to check that $f^G(\cdot) = T^{\succ x}(\cdot)$.

4 Results

4.1 Weak acyclicity

The example satisfies neither of the previous special cases SD or HM, and yet is OSP-implementable. We can generalize this example to provide a necessary and sufficient condition on the priority structure for OSP-implementability to hold.[16]

[15] Note in particular that y is *not* a possible outcome, because if node F is reached, agent i of type zxy follows the left edge and receives x.

[16] We call our condition weak acyclicity because it is reminiscent of acyclicity conditions that have arisen in other contexts (see Ergin (2002) and Kesten (2006)). Weak acyclicity is formally weaker than both of these, in that Kesten-acyclicity implies Ergin-acyclicity,

Definition 5. A priority structure $\succ_X$ is said to have a **strong cycle** if there are three objects $x, y, z \in X$ and three agents $i, j, k \in I$ such that: $i \succ_x j, k$, $j \succ_y i, k$ and $k \succ_z i, j$. If $\succ_X$ contains no strong cycles, then we say $\succ_X$ is **weakly acyclic**.

Weak acyclicity relates very clearly to the special cases we have seen thus far (HM, SD, and the example from the previous section). Under HM, every agent begins by "owning" exactly one object. This can be mapped into our model by constructing the priority structure $\succ_X$ such that for every x, the highest-ranked agent according to $\succ_x$ is the agent who initially owns x, and letting the rest of $\succ_x$ be defined arbitrarily. Defining the priority structure in this way, we see that $Top(\succ_x) \neq Top(\succ_y)$, and so for any housing market, the induced priority structure always contains a strong cycle. At the other extreme, consider a serial dictatorship, which can be mapped to our model by constructing the priority structure such that $\succ_x = \succ_y$ for all $x, y \in X$. It is clear that any such priority structure is trivially weakly acyclic. We summarize these two facts in the following observations.

Observation 1. *If $N \geq 3$ and $Top(\succ_x) \neq Top(\succ_y)$ for all distinct x, y, then $\succ_X$ contains a strong cycle.*

Observation 2. *If $\succ_x = \succ_y$ for all $x, y \in X$, then $\succ_X$ is weakly acyclic.*

The example from the previous section lies in between these two extremes: the priority structure is weakly acyclic, but does not reduce trivially to a serial dictatorship (this is why the results of Li, 2017 do not apply).

4.2 Main characterization

Our main result below shows that the construction in the example is no coincidence, as weak acyclicity exactly characterizes the class of priority struc-

which in turn implies weak acyclicity, though the converses do not hold. Thus, in particular, our characterization shows that neither Kesten-acyclicity nor Ergin-acyclicity are necessary for OSP implementation of TTC.

tures for which TTC is OSP-implementable.

Theorem 1. *The top trading cycles social choice rule, $T^{\succ_X}$, is OSP-implementable if and only if $\succ_X$ is weakly acyclic.*

The proof of this theorem will immediately follow, but we first note some useful properties of weakly acyclic priority structures. Let

$$\mathcal{T}(\succ_X) = \{i \in I : Top(\succ_x) = i \text{ for some } x \in X\}.$$

In words, $\mathcal{T}(\succ_X)$ is the set of agents who have the highest priority for at least one object under $\succ_X$. Also, starting from a given market $(I, X, \succ_X)$, consider a submarket $(I', X', \succ_{X'})$ that consists of a subset $I' \subset I$ of the initial agents and a subset $X' \subset X$ of the initial objects, and defines $\succ'_{X'}$ as follows:

$$\text{For all } j, k \in I' \text{ and all } y \in X' : j \succ'_y k \iff j \succ_y k.$$

In words, $\succ'_{X'}$ just deletes the agents in $I \setminus I'$ from the priority relation of every remaining object $x \in X'$, keeping the relative ordering of all other agents unchanged. We call any $\succ'_{X'}$ defined in this way a **contraction** of the original $\succ_X$.

We can now present three properties that will be useful in the proof.

(P1) If $\succ_X$ is weakly acyclic, then $|\mathcal{T}(\succ_X)| \leq 2$.

(P2) Let $\succ'_{X'}$ be a contraction of $\succ_X$. If $\succ_X$ is weakly acyclic, then $\succ'_{X'}$ is also weakly acyclic.

(P3) If $\succ_X$ is weakly acyclic, then $|\mathcal{T}(\succ'_{X'})| \leq 2$ for all contractions $\succ'_{X'}$ of $\succ_X$.

With (P1)-(P3) in hand, we now present the proof of Theorem 1.

Proof of Theorem 1. First we show the "if" part. To do so, we use induction on the market size n. First, consider $n = 2$.[17] Label the two objects x and y, and assume each has a different agent who has top priority: $i \succ_x j$ and $j \succ_y i$.[18] The following mechanism, labeled G^2, OSP-implements $T^{\succ_{X(n)}}$:

1. Ask agent i if she prefers x to y. If yes, assign i to x and j to y, and end the mechanism; otherwise, go to step 2.

2. Ask agent j if she prefers y to x. If yes, assign j to y and i to x. Otherwise, assign i to y and j to x.

It is simple to check that the above mechanism OSP-implements $T^{\succ_{X(n)}}$ for $n = 2$.

Next, consider a market of size N. By (P1) we have $|\mathcal{T}(\succ_{X(N)})| \leq 2$. First, consider the case $|\mathcal{T}(\succ_{X(N)})| = 1$, and, wlog, let $\mathcal{T}(\succ_{X(N)}) = \{i\}$. Define the mechanism G^N as follows:

1. Ask i which object is her top choice and assign her to this object.

2. Consider the contracted market $(I(N-1), X(N-1), \succ_{X(N-1)})$ where agent i is removed from the market together with the good she was assigned. This submarket is of size $N-1$, and, by (P2), $\succ_{X(N-1)}$ is weakly acyclic. By the inductive hypothesis, there exists a mechanism G^{N-1} that OSP-implements $T^{\succ_{X(N-1)}}$ on this submarket. Run mechanism G^{N-1} on the submarket.

It is clear that the grand mechanism G^N implements the TTC rule $T^{\succ_{X(N)}}$. In addition, this mechanism is obviously strategy-proof for agent i, since she simply receives her favorite object. For all other agents, their first decision node comes after i has been assigned, and so their strategic decision is

[17] With only two objects and two agents, the priority relation is trivially weakly acyclic.

[18] If this were not true, then TTC reduces to a serial dictatorship, which we already know is OSP-implementable.

equivalent to that under the mechanism that OSP implements $T^{\succ_{X(N-1)}}$ on the submarket, which, by induction is OSP. Thus, the above mechanism is obviously strategy-proof for all agents, and so $T^{\succ_{X(N)}}$ is OSP-implementable.

The remaining case to consider is $|\mathcal{T}(\succ_{X(N)})| = 2$. Let $\mathcal{T}(\succ_{X(N)}) = \{i, j\}$. Divide the objects into those for which i has top priority, $X_i(N) = \{x \in X(N) : Top(x) = i\}$ and those for which j has top priority $X_j(N) = \{x \in X^N : Top(x) = j\}$. Implement the following mechanism G^N:

1. For each $x \in X_i(N)$, ask agent i if her top choice is x. If i ever answers yes for some x, assign her to this x, and go to 1(a). Otherwise, skip to step 2.

   (a) We now have a submarket of size $N-1$ with a weakly acyclic priority structure $\succ_{X(N-1)}$. By the inductive hypothesis, $T^{\succ_{X(N-1)}}$ is OSP-implementable. Using a similar argument as above, this implies that $T^{\succ_{X(N)}}$ is also OSP-implementable.

2. For each $y \in X_j(N)$, ask agent j if her top choice is y. If yes, assign her to this y, and go to 2(a). Otherwise, skip to step 3.

   (a) We now have a submarket of size $N-1$ with a weakly acyclic priority structure $\succ_{X(N-1)}$. By the inductive hypothesis, $T^{\succ_{X(N-1)}}$ is OSP-implementable. Using a similar argument as above, this implies that $T^{\succ_{X(N)}}$ is also OSP-implementable.

3. If the answers to both (1) and (2) are "No", then i's top choice belongs to $X_j(N)$, and j's top choice belongs to $X_i(N)$. Ask i for her top choice, $x(i)$, and j for her top choice $x(j)$. Assign i to $x(j)$ and j to $x(i)$, and go to step 3(a).

   (a) We now have a submarket of size $N-2$ with a weakly acyclic priority structure $\succ_{X(N-2)}$. By the inductive hypothesis, $T^{\succ_{X(N-2)}}$

is OSP-implementable. Using a similar argument as above, this implies that $T^{\succ_{X(N)}}$ is also OSP-implementable.

Thus, we conclude via induction that $T^{\succ_{X(N)}}$ is OSP-implementable for all N, which completes the proof of the "if" part of the theorem.

Last, we must show the "only if" part: if $T^{\succ_X}$ is OSP-implementable, then $\succ_X$ is weakly acyclic. To do so, we prove the contrapositive: if $\succ_X$ contains a strong cycle, then TTC is not OSP-implementable. Proposition 5 in Li (2017) shows this for the special case of a Shapley-Scarf housing market (HM), and we generalize this to any cyclic priority structure. Choose a strong cycle, and label the agents i, j, k and objects x, y, and z such that $i \succ_x j, k$, $j \succ_y i, k$ and $k \succ_z i, j$.[19]

If choice rule $f : \mathcal{P}^N \to \mathcal{A}$ is not OSP-implementable on some limited domain $\tilde{\mathcal{P}}^N \subset \mathcal{P}^N$, then it is also not OSP-implementable on the full domain $\mathcal{P}^N$ (Li (2017)). Thus, consider a limited preference domain, $\tilde{\mathcal{P}}^N \subseteq \mathcal{P}^N$, defined as follows. Each $\ell \in I \setminus \{i, j, k\}$, has only one possible type, $\hat{P}$, that ranks all objects $w \in C \setminus \{x, y, z\}$ above x, y, and z (other than this, the ordering of the objects is arbitrary). For agents i, j, and k, each has two possible types, defined as follows (the dots indicate that all preferences below the top three are irrelevant, and can be chosen arbitrarily):

Agent i	
P_i	P_i'
y	z
z	y
x	x
$\vdots$	$\vdots$

Agent j	
P_j	P_j'
x	z
z	x
y	y
$\vdots$	$\vdots$

Agent k	
P_k	P_k'
x	y
y	x
z	z
$\vdots$	$\vdots$

Note that, on any such preference profile $P_I \in \tilde{P}^N$, the TTC rule $T^{\succ_X}$ must assign agents $\{i, j, k\}$ to objects $\{x, y, z\}$, and must assign all $\ell \neq i, j, k$

[19] If there are only two agents/objects, then the priority structure is trivially weakly acyclic, and hence, OSP-implementable.

to some other $w \neq x, y, z$. Consider some G that OSP implements $T^{\succ x}$, and, given a strategy profile S, let G' denote the *pruning* of G with respect to S. The pruned mechanism G' eliminates any actions (edges) that are never taken by any type of any agent, and, if G is OSP, then so is G'.[20] Thus, we only need to show that any mechanism "pruned" with respect to the truthful strategy profile cannot be OSP.

Assume an OSP mechanism G that implements $T^{\succ x}$ did exist, and restrict attention to its pruning G'. Consider the first node v that has two actions, and, without loss of generality, assume $\iota(v) = i$.[21] Assume first that i follows the action (or edge) corresponding to her true preferences P_i. If j follows the edge to associated with P'_j and k choosing the edge associated with P'_k, then, i receives x. On the other hand, consider i choosing the (other) action corresponding to P'_i, and k choosing the action associated with P_k. In this case, i receives z. Thus, if i's true preferences are P_i, the worst case from following the action corresponding to P_i is x, while the best outcome from following the action corresponding to P'_i is z. This implies that choosing the action corresponding to P_i is not obviously dominant, and so, i cannot be the first to have a non-singleton action set. By symmetry, the same holds for j and k.[22] Thus, all nodes have only a single action, which implies that G does not OSP-implement $T^{\succ x}$ on $\tilde{\mathcal{P}}^N$. This completes the proof of Theorem 1. ■

[20] More formally, we create a new tree G' which starts with the original tree G and, for each edge e, replaces the function $\alpha(e)$ with $\alpha'(e) = \alpha(e) \cap \tilde{\mathcal{P}}^N$, and then deletes any edge e' for which $\alpha'(e') = \emptyset$ (along with all subsequent nodes and edges, which will never be reached). See Li (2017) and Ashlagi and Gonczarowski (2016) for more detail.

[21] Each of i, j, k must have a node with two actions, or else the social choice rule implemented by G' would not be the TTC social choice rule. Since all $\ell \neq i, j, k$ have only one possible preference ranking, by pruning, each node v such that $\iota(v) = \ell$ has only one possible action. Thus, we can effectively just consider a mechanism defined over $\{i, j, k\}$ and $\{x, y, z\}$.

[22] Note that this problem is symmetric, as each of i, j, k has highest priority at exactly one of x, y, z, and each agent has two possible preference rankings that both rank the object she has highest priority at last, while alternating the preferences between the other two objects.

4.3 Weak acyclicity reinterpreted and dual dictatorships

Looking carefully at the proof of the theorem, we see that the main role of the weak acyclicity condition is to guarantee that there are at most two agents who have top priority for any object: $\mathcal{T}(\succ_X) \leq 2$. In fact, it is possible to interpret this latter statement as the condition driving the characterization of OSP-implementability. The condition needed is slightly stronger than just $\mathcal{T}(\succ_X) \leq 2$; we also need that for any possible *submarket,* $(I', X', \succ'_{X'})$, we continue to have $\mathcal{T}(\succ'_{X'}) \leq 2$ as well. Formally:

Theorem 2. *A priority structure $\succ_X$ is weakly acyclic if and only if $\mathcal{T}(\succ'_{X'}) \leq 2$ for every contraction $\succ'_{X'}$ of $\succ_X$ (including $\succ_X$ itself).*

Proof. The forward direction of this proposition is just a restatement of property (P3) above. For the converse, assume that $\mathcal{T}(\succ'_{X'}) \leq 2$ for all contractions $\succ'_{X'}$ (including the original $\succ_X$ itself), but that $\succ_X$ was not weakly acyclic. Since $\succ_X$ is not weakly acyclic, there exist three agents $i, j, k \in I$ and three objects $x, y, z \in X$ such that $i \succ_x j, k$, $j \succ_y i, k$ and $k \succ_z i, j$. Consider the submarket consisting only of these three agents and three objects, with the contraction $\succ'_{X'}$ defined in the usual way. Then, by construction, we have $Top(\succ'_x) = i$, $Top(\succ'_y) = j$ and $Top(\succ'_z) = k$, which implies $\mathcal{T}(\succ'_{X'}) = 3$, a contradiction. ■

Thus, this is another way to understand why Kesten/Ergin-acyclicity are sufficient for OSP-implementation, but are not necessary. These other acyclicity conditions are both sufficient for $\mathcal{T}(\succ_X) \leq 2$, but they are not necessary. They allow for cycles where the top priority at all objects is distributed amongst two agents (see Example 1), and it turns out that this is precisely what is needed for OSP-implementation.

If $\mathcal{T}(\succ'_{X'}) = 1$ for every contraction of $\succ'_{X'}$, then the TTC mechanism reduces to a simple serial dictatorship. Allowing $\mathcal{T}(\succ'_{X'}) \leq 2$ expands the set of possible mechanisms into what may be called *dual dictatorship*s. In a dual dictatorship, the set of objects X is partitioned into two subsets, Y_0 and

Z_0. Each set is then "endowed" to one agent (it is possible for the same agent to be endowed both sets). Each agent then becomes the "dictator" over her endowment: if either agent owns her favorite object, then she is assigned to that object and is removed from the market. If not, then each agent's favorite object is endowed to the other dictator. They swap these objects, and leave the market. The remaining objects are then again divided into two sets, Y_1 and Z_1, and each set is endowed to one agent, with the restriction that if an agent is ever endowed with an object, she never loses that endowment until she is assigned and leaves the market. The process is then repeated until everyone is assigned.[23] Since we know that TTC is Pareto efficient, and dual dictatorships are a special case, they also will be Pareto efficient. In addition, they will be obviously strategy-proof, which does not hold for TTC more generally.

5 Conclusion

We study whether the (strategy-proof) top trading cycles mechanism can be implemented in a manner such that it is additionally *obviously* strategy-proof, a solution concept introduced recently by Li (2017). We fully characterize the set of OSP-implementable TTC mechanisms using a new acyclicity condition that is weaker than two standard acyclicity conditions common in the literature. Under the acyclicity condition, TTC becomes equivalent to a dual dictatorship, where at any given time, at most two agents "own" all of the seats and can either choose their favorite (if they own it) or swap with the other dictator (so long as the other dictator wants one of her objects). As the acyclicity condition is unlikely to be satisfied in many practical allocation problems, the strategy-proofness of TTC may be difficult for non-experts to

[23] In Example 1, agents i and j are the first dual dictators, with i owning x and y and j owning z. If i takes x, then y is passed to k, and j and k are the dual dictators, while if i takes y, x is passed to j, who is the sole dictator. If i wants z and j wants x or y, then they swap, and k is left with whatever remains.

understand. This may explain why TTC has a reputation of being difficult to understand and is rarely observed in practice, despite the existence of a rich theoretical literature promoting its virtues.

References

Abdulkadiroglu, A., Y.-K. Che, P. A. Pathak, A. E. Roth, and O. Tercieux (2017): "Minimizing Justified Envy in School Choice: The Design of New Orleans' OneApp," Tech. rep., National Bureau of Economic Research.

Abdulkadiroğlu, A. and T. Sönmez (2003): "School choice: A mechanism design approach," *American economic review*, 729–747.

Abdulkadiroğlu, A., Y.-K. Che, and Y. Yasuda (2011): "Resolving Conflicting Preferences in School Choice: The "Boston" Mechanism Reconsidered," *American Economic Review*, 101, 399–410.

Abdulkadiroğlu, A., P. A. Pathak, and A. E. Roth (2005a): "The New York City high school match," *American Economic Review, Papers and Proceedings*, 95, 364–367.

Abdulkadiroğlu, A., P. A. Pathak, A. E. Roth, and T. Sönmez (2005b): "The Boston Public School Match," *American Economic Review, Papers and Proceedings*, 95, 364–367.

Abdulkadiroğlu, A. and T. Sönmez (1999): "House Allocation with Existing Tenants," *Journal of Economic Theory*, 88, 233–260.

——— (2003): "School Choice: A Mechanism Design Approach," *American Economic Review*, 93, 729–747.

Arribillaga, R. P., J. Massó, and A. Neme (2017): "Not All Majority-based Social Choice Functions Are Obviously Strategy-proof," .

Ashlagi, I. and Y. A. Gonczarowski (2016): "No stable matching mechanism is obviously strategy-proof," *arXiv preprint arXiv:1511.00452*.

Bade, S. and Y. A. Gonczarowski (2016): "Gibbard-Satterthwaite

Success Stories and Obvious Strategyproofness," *arXiv preprint arXiv:1610.04873*.

Bogomolnaia, A., R. Deb, and L. Ehlers (2005): "Strategy-proof assignment on the full preference domain," *Journal of Economic Theory*, 123, 161–186.

Chen, Y. and T. Sönmez (2006): "School Choice: An Experimental Study," 127, 202–231.

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

Dur, U. and T. Morrill (2017): "Competitive equilibria in school assignment," *Games and Economic Behavior*.

Dur, U. M. and M. U. Ünver (2015): "Two-sided matching via balanced exchange: Tuition and worker exchanges," .

Ergin, H. (2002): "Efficient Resource Allocation on the Basis of Priorities," *Econometrica*, 70, 2489–2498.

Featherstone, C. and M. Niederle (2014): "Improving on Strategyproof School Choice Mechanisms: An Experimental Investigation," Working paper, Stanford University.

Hakimov, R. and O. Kesten (2014): "The equitable top trading cycles mechanism for school choice," Tech. rep., WZB discussion paper.

Hassidim, A., A. Romm, and R. I. Shorrer (2015): "Strategic" behavior in a strategy-proof environment," Tech. rep., working paper.

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

Leshno, J. and I. Lo (2017): "The Simple Structure of Top Trading Cycles in School Choice," .

Li, S. (2017): "Obviously Strategy-Proof Mechanisms," *American Economic Review*, 107.

Ma, J. (1994): "Strategy-proofness and the strict core in a market with indivisibilities," *International Journal of Game Theory*, 23, 75–83.

Mackenzie, A. (2017): "A Revelation Principle for Obviously Strategy-proof Implementation," *working paper*.

Miralles, A. (2009): "School choice: The case for the Boston mechanism," in *Auctions, Market Mechanisms and Their Applications*, Springer, 58–60.

Morrill, T. (2013): "An alternative characterization of top trading cycles," *Economic Theory*, 54, 181–197.

——— (2015): "Two simple variations of top trading cycles," *Economic Theory*, 60, 123–140.

Pápai, S. (2000): "Strategyproof assignment by hierarchical exchange," *Econometrica*, 68, 1403–1433.

Pathak, P. A. (2016): "What really matters in designing school choice mechanisms," in *Preparation for Advances in Economics and Econometrics, 11th World Congress of the Econometric Society*.

Pathak, P. A. and T. Sönmez (2008): "Leveling the Playing Field: Sincere and Sophisticated Players in the Boston Mechanism," *American Economic Review*, 98, 1636–1652.

Pycia, M. and P. Troyan (2016): "Obvious dominance and random priority," *SSRN 2853563*.

Pycia, M. and M. U. Ünver (2017): "Incentive compatible allocation and exchange of discrete resources," *Theoretical Economics*, 12, 287–329.

Rees-Jones, A. (2017): "Suboptimal behavior in strategy-proof mechanisms: Evidence from the residency match," *Games and Economic Behavior*.

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

——— (1982b): "Incentive Compatibility in a Market with Indivisible Goods," *Economics Letters*, 9, 127–132.

Shapley, L. S. and H. E. Scarf (1974): "On Cores and Indivisibility," *Journal of Mathematical Economics*, 1, 23–28.

Shorrer, R. and S. Sóvágó (2017): "Obvious Mistakes in a Strategically

Simple College Admissions Environment," .

Troyan, P. (2012): "Comparing School Choice Mechanisms by Interim and Ex-Ante Welfare," *Games and Economic Behavior*, 75, 936–947.

Zhang, L. and D. Levin (2017a): "Bounded Rationality and Robust Mechanism Design: An Axiomatic Approach," *American Economic Review Papers and Proceedings*, 107, 235–39.

——— (2017b): "Partition Obvious Preference and Mechanism Design: Theory and Experiment," .